ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΕΦΑΡΜΟΣΜΕΝΕΣ ΜΑΘΗΜΑΤΙΚΕΣ ΕΠΙΣΤΗΜΕΣ

ΚΡΙΤΗΡΙΑ ΠΛΗΡΟΦΟΡΙΑΣ ΓΙΑ ΤΗΝ ΕΠΙΛΟΓΗ ΜΕΤΑΒΛΗΤΩΝ ΣΤΙΣ ΜΗΧΑΝΕΣ ΔΙΑΝΥΣΜΑΤΙΚΗΣ ΥΠΟΣΤΗΡΙΞΗΣ ΚΑΙ ΕΦΑΡΜΟΓΕΣ

(INFORMATION CRITERIA FOR VARIABLE SELECTION IN SUPPORT VECTOR MACHINES AND APPLICATIONS)

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ
ΤΗΣ
ΕΥΓΕΝΙΑΣ Ι.ΣΤΟΥΦΗ

Επιβλέπων: ΧΡΗΣΤΟΣ ΚΟΥΚΟΥΒΙΝΟΣ
Καθηγητής Ε.Μ.Π.

Αθήνα, Ιανουάριος 2015

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΕΦΑΡΜΟΣΜΕΝΕΣ ΜΑΘΗΜΑΤΙΚΕΣ ΕΠΙΣΤΗΜΕΣ

ΚΡΙΤΗΡΙΑ ΠΛΗΡΟΦΟΡΙΑΣ ΓΙΑ ΤΗΝ ΕΠΙΛΟΓΗ ΜΕΤΑΒΛΗΤΩΝ ΣΤΙΣ ΜΗΧΑΝΕΣ ΔΙΑΝΥΣΜΑΤΙΚΗΣ ΥΠΟΣΤΗΡΙΞΗΣ ΚΑΙ ΕΦΑΡΜΟΓΕΣ

(INFORMATION CRITERIA FOR VARIABLE SELECTION IN SUPPORT VECTOR MACHINES AND APPLICATIONS)

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ
ΤΗΣ
ΕΥΓΕΝΙΑΣ Ι.ΣΤΟΥΦΗ

Επιβλέπων: ΧΡΗΣΤΟΣ ΚΟΥΚΟΥΒΙΝΟΣ
Καθηγητής Ε.Μ.Π.

Αθήνα, Ιανουάριος 2015

Πίνακας περιεχομένων

ΠΕΡΙΕΧΟΜΕΝΑ…… 5

ΠΕΡΙΛΗΨΗ……11

ABSTRACT……13

ΕΥΧΑΡΙΣΤΙΕΣ……15

ΚΕΦΑΛΑΙΟ 1: ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ……17
1.1 ΕΙΣΑΓΩΓΙΚΑ ΣΤΟΙΧΕΙΑ ΓΙΑ ΤΟ DATA MINING..... 17
1.2 ΔΙΑΔΙΚΑΣΙΑ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ..... 18
1.3 ΣΥΝΙΣΤΩΣΕΣ ΔΕΔΟΜΕΝΩΝ..... 19
1.3.1 ΟΡΟΛΟΓΙΑ.....19
1.3.2 ΤΥΠΟΙ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ..... 19
1.4 ΜΕΘΟΔΟΙ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ..... 20
1.5 ΤΑ ΣΥΣΤΑΤΙΚΑ ΤΩΝ ΑΛΓΟΡΙΘΜΩΝ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ..... 21
1.6 ΚΑΤΗΓΟΡΙΕΣ ΤΟΥ DATA MINING-ΕΙΔΗ ΜΑΘΗΣΗΣ..... 22
1.6.1 ΚΑΤΗΓΟΡΙΕΣ ΤΟΥ DATA MINING..... 22
1.6.2 ΕΙΔΗ ΜΑΘΗΣΗΣ ΚΑΙ ΜΕΘΟΔΟΙ ΤΟΥ DATA MINING……22
1.7 ΕΦΑΡΜΟΓΕΣ ΤΟΥ DATA MINING..... 23

ΚΕΦΑΛΑΙΟ 2: ΜΗΧΑΝΕΣ ΔΙΑΝΥΣΜΑΤΙΚΗΣ ΥΠΟΣΤΗΡΙΞΗΣ (SVMS)……25
2.1 ΕΙΣΑΓΩΓΗ.....25
2.2 ΒΕΛΤΙΣΤΟ ΔΙΑΧΩΡΙΣΤΙΚΟ ΥΠΕΡΕΠΙΠΕΔΟ..... 25
2.3 Ο ΤΑΞΙΝΟΜΗΤΗΣ ΔΙΑΝΥΣΜΑΤΙΚΗΣ ΥΠΟΣΤΗΡΙΞΗΣ..... 27
2.4 ΥΠΟΛΟΓΙΣΜΟΣ ΤΟΥ ΤΑΞΙΝΟΜΗΤΗ ΔΙΑΝΥΣΜΑΤΙΚΗΣ ΥΠΟΣΤΗΡΙΞΗΣ..... 29
2.5 ΥΠΟΛΟΓΙΣΜΟΣ ΤΟΥ SVM ΓΙΑ ΤΑΞΙΝΟΜΗΣΗ ΚΑΙ ΣΥΝΑΡΤΗΣΕΙΣ ΠΥΡΗΝΑ (KERNEL FUNCTIONS)..... 31
2.6 ΤΕΧΝΑΣΜΑ ΠΥΡΗΝΑ (KERNEL TRICK)..... 33
2.7 ΟΙ SVMS ΩΣ ΜΕΘΟΔΟΣ ΠΟΙΝΙΚΟΠΟΙΗΣΗΣ..... 33

ΚΕΦΑΛΑΙΟ 3: ΜΗΧΑΝΕΣ ΔΙΑΝΥΣΜΑΤΙΚΗΣ ΥΠΟΣΤΗΡΙΞΗΣ ΠΟΛΛΑΠΛΩΝ ΚΛΑΣΕΩΝ (MULTI-CLASS SVMS)……37
3.1 ΕΙΣΑΓΩΓΗ……37
3.2 ONE-AGAINST-ALL MULTI-CLASS SVM……37
3.3 ONE-AGAINST-ONE MULTICLASS SVM……38
3.4 ΥΠΟΛΟΓΙΣΤΙΚΗ ΠΟΛΥΠΛΟΚΟΤΗΤΑ ΓΙΑ MULTI-CLASS SVM……39
3.5 ΣΥΖΗΤΗΣΗ……39
3.6 ERROR-CORRECTING OUTPUT CODES……40

ΚΕΦΑΛΑΙΟ 4: ΚΡΙΤΗΡΙΑ ΠΛΗΡΟΦΟΡΙΑΣ……43
4.1 ΕΙΣΑΓΩΓΗ……43
4.2 ΜΕΡΟΛΗΨΙΑ ΔΙΑΚΥΜΑΝΣΗ ΚΑΙ ΠΟΛΥΠΛΟΚΟΤΗΤΑ ΜΟΝΤΕΛΟΥ……43
4.3 Η ΑΠΟΣΥΝΘΕΣΗ ΜΕΡΟΛΗΨΙΑΣ-ΔΙΑΚΥΜΑΝΣΗΣ……45
4.4 ΒΕΛΤΙΩΣΗ ΤΟΥ ΠΟΣΟΣΤΟΥ ΣΦΑΛΜΑΤΟΣ ΕΚΠΑΙΔΕΥΣΗΣ……46
4.5 ΕΚΤΙΜΗΣΕΙΣ ΤΟΥ ΕΝΤΟΣ- ΔΕΙΓΜΑΤΟΣ ΣΦΑΛΜΑΤΟΣ ΠΡΟΒΛΕΨΗΣ ΚΑΙ ΚΡΙΤΗΡΙΟ ΠΛΗΡΟΦΟΡΙΑΣ ΑΚΑΙΚΕ……47
4.6 ΑΠΟΔΟΤΙΚΟΣ ΑΡΙΘΜΟΣ ΠΑΡΑΜΕΤΡΩΝ……48
4.7 ΜΠΕΫΖΙΑΝΗ ΠΡΟΣΕΓΓΙΣΗ ΚΑΙ ΚΡΙΤΗΡΙΟ ΠΛΗΡΟΦΟΡΙΑΣ BIC……49
4.8 ΔΙΑΣΤΑΥΡΩΜΕΝΗ ΕΠΙΚΥΡΩΣΗ……50
4.8.1 ΟΡΙΣΜΟΣ……50
4.8.2 HOLD-OUT ΕΠΙΚΥΡΩΣΗ……51
4.8.3 K-ΠΛΑΣΙΑ ΔΙΑΣΤΑΥΡΩΜΕΝΗ ΕΠΙΚΥΡΩΣΗ……52
4.8.4 10-ΠΛΑΣΙΑ ΔΙΑΣΤΑΥΡΩΜΕΝΗ ΕΠΙΚΥΡΩΣΗ……52
4.8.5 LEAVE-ONE-OUT ΔΙΑΣΤΑΥΡΩΜΕΝΗ ΕΠΙΚΥΡΩΣΗ (LOOCV)……54
4.9 ΔΙΑΣΤΑΥΡΩΜΕΝΑ ΕΠΙΚΥΡΩΜΕΝΟ ΠΟΣΟΣΤΟ ΣΦΑΛΜΑΤΟΣ (CROSS VALIDATED ERROR RATE)……54
4.9.1 ΕΙΣΑΓΩΓΗ……54
4.9.2 ΑΞΙΟΛΟΓΩΝΤΑΣ ΤΗΝ ΑΠΟΔΟΣΗ ΤΗΣ ΕΠΙΛΟΓΗΣ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ……55
4.9.3 Η ΣΧΕΤΙΚΗ ΣΗΜΑΣΙΑ ΤΩΝ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ ΣΤΟΥΣ SVM ΤΑΞΙΝΟΜΗΤΕΣ……55
4.9.4 ΑΝΑΔΡΟΜΙΚΗ ΤΑΞΙΝΟΜΗΣΗ ΚΑΙ ΕΠΙΛΟΓΗ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ……57
4.10 ΜΕΘΟΔΟΣ GACV (ΓΕΝΙΚΕΥΜΕΝΗ ΠΡΟΣΕΓΓΙΣΗ ΤΗΣ ΔΙΑΣΤΑΥΡΩΜΕΝΗΣ ΕΠΙΚΥΡΩΣΗΣ) 58
4.10.1 ΑΠΟΣΤΑΣΗ KULLBACK-LEIBLER……58
4.10.2 ΣΥΓΚΡΙΤΙΚΗ KULLBACK-LEIBLER ΑΠΟΣΤΑΣΗ (CKL DISTANCE)……58
4.10.3 Η ΓΕΝΙΚΕΥΜΕΝΗ KULLBACK-LEIBLER ΑΠΟΣΤΑΣΗ (GCKL DISTANCE)……59
4.10.4 ΜΙΑ ΥΠΟΛΟΓΙΣΙΜΗ PROXY ΓΙΑ ΤΗ GCKL-ΠΡΟΣΕΓΓΙΣΗ ΤΗΣ ΔΙΑΣΤΑΥΡΩΜΕΝΗΣ ΕΠΙΚΥΡΩΣΗΣ……59
4.10.5 ΤΟ LEAVING-OUT-ONE ΛΗΜΜΑ……60
4.10.6 ΜΙΑ ΠΡΟΣΕΓΓΙΣΗ ΓΙΑ ΤΟ $h_{ii} = \frac{f_{\lambda_i} - f_{\lambda}^{[-i]}}{y_i - \mu_{\lambda}^{[-i]}}$……61
4.10.7 ΤΟ ΤΥΧΑΙΟΠΟΙΗΜΕΝΟ ΙΧΝΟΣ ΕΚΤΙΜΗΣΗΣ ΤΗΣ GACV ΓΙΑ ΤΟ $g(\tau) = [\varepsilon - \tau]_+$……61
4.10.8 ΣΥΖΗΤΗΣΗ ΓΙΑ ΤΗ ranGACV……62
4.11 ΕΝΔΕΙΞΗ ΥΠΟΛΟΓΙΣΜΕΝΗ ΜΕ ΤΗ ΜΕΘΟΔΟ LAPLACE……63
4.11.1 ΕΙΣΑΓΩΓΗ……63
4.11.2 ΤΟ ΠΛΑΙΣΙΟ ΤΗΣ ΕΝΔΕΙΞΗΣ ΚΑΙ Η ΕΦΑΡΜΟΓΗ ΣΤΟ SVM……63
4.11.3 ΠΡΟΣΟΜΟΙΩΣΗ……64

4.12 ξα-ΕΚΤΙΜΗΣΗ..... 67
4.12.1 ΕΙΣΑΓΩΓΗ..... 67
4.12.2 ΤΑΞΙΝΟΜΗΣΗ ΚΕΙΜΕΝΟΥ..... 67
4.12.3 ΕΚΤΙΜΗΤΕΣ ΕΠΑΡΚΟΥΣ ΑΠΟΔΟΣΗΣ ΓΙΑ ΤΑ SVMS.....68

4.12.4 ΠΟΣΟΣΤΟ ΣΦΑΛΜΑΤΟΣ..... 68
4.12.5 ΑΝΑΚΛΗΣΗ, ΑΚΡΙΒΕΙΑ ΚΑΙ F1..... 69
4.12.6 ΠΕΙΡΑΜΑΤΑ..... 69
4.13 ΜΕΘΟΔΟΣ ΕΛΑΧΙΣΤΟΠΟΙΗΣΗΣ ΕΓΓΥΗΜΕΝΟΥ ΚΙΝΔΥΝΟΥ (GUARANTEED RISK MINIMIZATION).....73
4.13.1 ΕΙΣΑΓΩΓΗ..... 73
4.13.2 ΟΡΙΣΜΟΙ..... 73
4.13.3 Η ΜΕΘΟΔΟΣ ΔΙΑΡΘΡΩΤΙΚΗΣ ΕΛΑΧΙΣΤΟΠΟΙΗΣΗΣ ΚΙΝΔΥΝΟΥ (STRUCTURAL RISK MINIMIZATION) ΚΑΙ ΑΝΩ ΦΡΑΓΜΑ ΕΝΟΣ ΠΡΑΓΜΑΤΙΚΟΥ ΚΙΝΔΥΝΟΥ (TRUE RISK)..... 74
4.13.4 ΔΙΑΣΤΑΣΗ VAPNIK- CHERVONENKIS..... 75
4.13.5 ΔΙΑΡΘΡΩΤΙΚΗ ΕΛΑΧΙΣΤΟΠΟΙΗΣΗ ΚΙΝΔΥΝΟΥ (STRUCTURAL RISK MINIMIZATION). 75
4.13.6 ΕΛΑΧΙΣΤΟΠΟΙΗΣΗ ΕΜΠΕΙΡΙΚΟΥ ΚΙΝΔΥΝΟΥ (EMPIRICAL RISK MINIMIZATION)..... 75
4.13.7 ΑΛΓΟΡΙΘΜΟΣ ΕΠΙΛΟΓΗΣ ΜΟΝΤΕΛΟΥ GRM..... 76
4.13.8 ΠΕΙΡΑΜΑΤΑ..... 76
4.14 ΜΗΚΟΣ ΕΛΑΧΙΣΤΗΣ ΠΕΡΙΓΡΑΦΗΣ (MINIMUM DESCRIPTION LENGTH)..... 78
4.15 ΜΕΘΟΔΟΙ BOOTSTRAP..... 79

ΚΕΦΑΛΑΙΟ 5: ΝΕΑ ΚΡΙΤΗΡΙΑ ΠΛΗΡΟΦΟΡΙΑΣ.....81
5.1 ΚΡΙΤΗΡΙΑ ΠΛΗΡΟΦΟΡΙΑΣ ΓΙΑ ΜΗΧΑΝΕΣ ΠΥΡΗΝΑ..... 81
5.1.1 ΕΙΣΑΓΩΓΗ..... 81
5.1.2 Η ΛΟΓΙΣΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ ΠΥΡΗΝΑ (KLR).....82
5.1.3 ΤΟ ΛΟΓΙΣΤΙΚΟ ΜΠΕΥΖΙΑΝΟ ΠΛΑΙΣΙΟ ΓΙΑ ΤΑ SVMS ΚΑΙ ΤΗΝ KLR..... 83
5.1.4 ΤΟ ΜΠΕΥΖΙΑΝΟ ΠΛΑΙΣΙΟ ΤΟΥ SOLLICH ΓΙΑ ΤΑ SVMS..... 83
5.1.5 ΤΟ ΚΡΙΤΗΡΙΟ ΣΥΣΤΗΜΑΤΟΠΟΙΗΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ (RIC)..... 84
5.1.6 ΤΟ ΚΡΙΤΗΡΙΟ ΣΥΣΤΗΜΑΤΟΠΟΙΗΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ ΠΥΡΗΝΑ (KRIC)..... 85
5.1.7 Η ΠΡΟΣΕΓΓΙΣΤΙΚΗ ΜΕΘΟΔΟΣ ΤΟΥ NYSTROM ΓΙΑ ΤΟΝ ΥΠΟΛΟΓΙΣΜΟ ΤΟΥ KRIC..... 88
5.1.8 ΠΕΙΡΑΜΑΤΑ..... 90
5.1.9 ΣΥΜΠΕΡΑΣΜΑΤΑ..... 95
5.2 ΕΝΑ ΚΡΙΤΗΡΙΟ ΠΛΗΡΟΦΟΡΙΑΣ ΓΙΑ ΕΠΙΛΟΓΗ ΜΕΤΑΒΛΗΤΩΝ ΣΤΙΣ ΜΗΧΑΝΕΣ ΔΙΑΝΥΣΜΑΤΙΚΗΣ ΥΠΟΣΤΗΡΙΞΗΣ..... 97
5.2.1 ΕΙΣΑΓΩΓΗ..... 97
5.2.2 ΥΠΑΡΧΟΥΣΕΣ ΤΕΧΝΙΚΕΣ ΕΠΙΛΟΓΗΣ ΜΕΤΑΒΛΗΤΗΣ.....97
5.2.3 ΤΕΧΝΙΚΕΣ ΚΑΤΑΤΑΞΗΣ ΜΕΤΑΒΛΗΤΗΣ.....98
5.2.4 ΤΑ ΝΕΑ ΚΡΙΤΗΡΙΑ ΠΛΗΡΟΦΟΡΙΑΣ.....99
5.2.5 ΜΟΤΙΒΟ ΠΡΟΣΟΜΟΙΩΣΗΣ.....102
5.2.6 ΔΟΚΙΜΕΣ ΣΕ ΠΡΑΓΜΑΤΙΚΑ ΔΕΔΟΜΕΝΑ..... 110

ΚΕΦΑΛΑΙΟ 6: ΚΡΙΤΗΡΙΑ ΠΛΗΡΟΦΟΡΙΑΣ ΓΙΑ ΕΠΙΛΟΓΗ ΜΟΝΤΕΛΟΥ ΣΤΑ MULTICLASS SVMS.....115
6.1 ΕΙΣΑΓΩΓΗ.....115

6.2 ΑΠΑΙΤΟΥΜΕΝΕΣ ΕΝΝΟΙΕΣ....116
6.2.1 ΠΡΟΣΕΓΓΙΣΗ ΜΕΜΟΝΩΜΕΝΗΣ ΜΗΧΑΝΗΣ (SINGLE-MACHINE APPROACH)....116
6.2.2 ΜΕΘΟΔΟΙ ΕΡΕΥΝΑΣ ΒΑΣΙΣΜΕΝΕΣ ΣΤΟ ΠΛΕΓΜΑ (GRID-BASED SEARCH METHODS).116
6.2.3 ΜΕΘΟΔΟΙ ΒΕΛΤΙΣΤΟΠΟΙΗΣΗΣ ΒΑΣΙΣΜΕΝΕΣ ΣΤΗΝ ΚΛΙΣΗ (GRADIENT-BASED OPTIMIZATION TECHNIQUES)....117
6.2.4 BROYDEN-FLETCHER-GOLDFARB-SHANNO (BFGS) QUASI-NEWTON ΜΕΘΟΔΟΣ....117
6.3 ΦΡΑΓΜΑ ΑΚΤΙΝΩΤΟΥ ΠΕΡΙΘΩΡΙΟΥ ΓΙΑ ΔΥΑΔΙΚΑ SVMS....118
6.4 ΚΡΙΤΗΡΙΟ ΕΠΙΛΟΓΗΣ ΜΟΝΤΕΛΟΥ I....119
6.5 ΚΡΙΤΗΡΙΟ ΕΠΙΛΟΓΗΣ ΜΟΝΤΕΛΟΥ II....122
6.6 ΥΠΟΛΟΓΙΣΤΙΚΟ ΘΕΜΑ....125
6.7 ΠΕΙΡΑΜΑΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ....126
6.8 ΙΔΙΟΤΗΤΕΣ ΤΩΝ ΔΥΟ ΚΡΙΤΗΡΙΩΝ ΕΠΙΛΟΓΗΣ ΜΟΝΤΕΛΟΥ....127
6.9 ΠΕΙΡΑΜΑΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ ΓΙΑ ΤΑ ΣΥΝΟΛΑ ΔΕΔΟΜΕΝΩΝ ΑΝΑΦΟΡΑΣ....129
6.10 ΣΥΜΠΕΡΑΣΜΑΤΑ....137

ΚΕΦΑΛΑΙΟ 7: ΑΞΙΟΛΟΓΗΣΗ ΜΟΝΤΕΛΟΥ ΚΑΙ ΕΦΑΡΜΟΓΕΣ....139
7.1 ΕΙΣΑΓΩΓΗ....139
7.1.1 ΣΤΑΤΙΣΤΙΚΟΣ ΕΛΕΓΧΟΣ ΥΠΟΘΕΣΕΩΝ....140
7.1.2 ΚΡΙΤΗΡΙΑ ΑΠΟΔΟΣΗΣ ΤΟΥ ΜΟΝΤΕΛΟΥ-CONFUSION MATRIX....141
7.2 ΕΙΣΑΓΩΓΗ ΣΤΗΝ R ΚΑΙ ΓΕΝΙΚΗ ΜΕΘΟΔΟΛΟΓΙΑ....143
7.2.1 ΤΟ ΠΑΚΕΤΟ R....143
7.2.2 ΒΑΣΙΚΑ ΒΗΜΑΤΑ....144
7.3 ΠΡΩΤΗ ΕΦΑΡΜΟΓΗ....144
7.3.1 ΟΡΙΣΜΟΣ ΤΟΥ CREDIT SCORING ΚΑΙ ΤΟ ΠΡΟΒΛΗΜΑ ΠΟΥ ΚΑΛΕΙΤΑΙ ΝΑ ΛΥΣΕΙ....144
7.3.2 ΠΕΡΙΓΡΑΦΗ ΤΗΣ ΕΦΑΡΜΟΓΗΣ ΚΑΙ ΤΟΥ ΣΥΝΟΛΟΥ ΔΕΔΟΜΕΝΩΝ (GERMAN CREDIT DATA)....145
7.3.3 ΧΕΙΡΙΣΜΟΣ ΤΟΥ ΣΥΝΟΛΟΥ ΔΕΔΟΜΕΝΩΝ (GERMAN CREDIT DATA) ΣΤΗΝ R....147
7.4 ΔΕΥΤΕΡΗ ΕΦΑΡΜΟΓΗ....148
7.4.1 ΤΟ ΣΥΝΟΛΟ ΔΕΔΟΜΕΝΩΝ IRIS DATA SET....148
7.4.2 ΠΕΡΙΓΡΑΦΗ ΤΗΣ ΕΦΑΡΜΟΓΗΣ ΚΑΙ ΤΟΥ ΣΥΝΟΛΟΥ ΔΕΔΟΜΕΝΩΝ (IRIS DATA SET)...148
7.4.3 ΧΕΙΡΙΣΜΟΣ ΤΟΥ ΣΥΝΟΛΟΥ ΔΕΔΟΜΕΝΩΝ (GERMAN CREDIT DATA) ΣΤΗΝ R....149

ΚΕΦΑΛΑΙΟ 8: ΣΥΝΟΨΗ-ΣΥΜΠΕΡΑΣΜΑΤΑ....151

ΒΙΒΛΙΟΓΡΑΦΙΑ....153

ΠΑΡΑΡΤΗΜΑΤΑ....151
ΠΑΡΑΡΤΗΜΑ Α: ΑΠΟΔΕΙΞΗ ΤΗΣ ΣΧΕΣΗΣ (5.1.7.2)....157
ΠΑΡΑΡΤΗΜΑ Β: ΣΕΙΡΙΑΚΗ ΕΛΑΧΙΣΤΗ ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ (SEQUENTIAL MINIMAL OPTIMISATION(SMO))....159
ΠΑΡΑΡΤΗΜΑ Γ:....160
Γ.1 GERMAN CREDIT DATA....160
Γ.1.1 ΠΥΡΗΝΕΣ....160
Γ.1.2 ΣΥΝΟΛΟ ΕΚΠΑΙΔΕΥΣΗΣ....161
Γ.1.3 ΣΥΝΟΛΟ ΔΟΚΙΜΗΣ....164

Γ.2 IRIS DATA SET..... 166
Γ.2.1 ΠΥΡΗΝΕΣ..... 166
Γ.2.2 ΣΥΝΟΛΟ ΕΚΠΑΙΔΕΥΣΗΣ..... 167
Γ.2.3 ΣΥΝΟΛΟ ΔΟΚΙΜΗΣ.....171

Περίληψη

Η πρόοδος στην ψηφιακή απόκτηση δεδομένων και στην τεχνολογία αποθήκευσης έχει σαν αποτέλεσμα την αύξηση των τεραστίων βάσεων δεδομένων. Η ανάλυση αυτών των βάσεων για την εύρεση ανυποψίαστων σχέσεων και τη σύνοψη των δεδομένων με νέους τρόπους που να είναι κατανοητοί όσο και χρήσιμοι, αποτελεί την Εξόρυξη Δεομένων (Data Mining). Κύριοι στόχοι του D.M είναι η πρόβλεψη και η περιγραφή, ένας από τους τρόπους επίτευξής τους είναι η ταξινόμηση, δηλαδή η εκμάθηση μιας λειτουργίας που να ταξινομεί ένα στοιχείο δεδομένων σε μία από τις πολλές προκαθορισμένες κλάσεις. Στην παρούσα εργασία, θα αναλυθούν θεωρητικά οι μηχανές διανυσματικής υποστήριξης δύο και περισσοτέρων κλάσεων, οι οποίες αποτελούν μια νέα μέθοδο ταξινόμησης. Ο πρωταρχικός στόχος, όμως, του D.M. είναι η εξαγωγή νέων πληροφοριών από τα δεδομένα, ο οποίος μπορεί να επιτευχθεί με διάφορες τεχνικές που διακρίνονται στις μεθόδους με επίβλεψη και χωρίς επίβλεψη. Η πρώτη κατηγορία απαρτίζεται από τα στάδια της επιλογής και της εκτίμησης μοντέλου, για αυτά τα δύο στάδια έχει αναπτυχθεί μια μεγάλη ποικιλία μεθόδων και κριτηρίων. Στόχος και κύριο θέμα της παρούσης εργασίας είναι η παρουσίαση αυτών των κριτηρίων και μεθόδων και η εισαγωγή νέων Κριτηρίων πληροφορίας για τα δυαδικά και τα multiclass SVMs και η πειραματική σύγκριση της απόδοσής τους. Στο τέλος της συγκεκριμένης εργασίας, αξιολογείται και η πειραματική εφαρμογή των μεθόδων διανυσματικής υποστήριξης σε προβλήματα ταξινόμησης δύο και τριών κλάσεων.

Πιο συγκεκριμένα:

- Το πρώτο Κεφάλαιο, αποτελεί μια εισαγωγή για τις βασικές έννοιες της Εξόρυξης Δεδομένων όπως είναι η διαδικασία του, οι συνιστώσες των δεδομένων, οι μέθοδοί του, τα συστατικά των αλγορίθμων του, οι κατηγορίες του, τα είδη μάθησής του και οι εφαρμογές του.
- Στο δεύτερο Κεφάλαιο, αναλύεται το θεωρητικό υπόβαθρο των δυαδικών SVMs, και πιο συγκεκριμένα, το βέλτιστο διαχωριστικό υπερεπίπεδο, ο ταξινομητής διανυσματικής υποστήριξης, το δυαδικό SVM, οι πυρήνες, το τέχνασμα πυρήνα, και τέλος το SVM ως μέθοδος ποινικοποίησης.
- Στο τρίτο Κεφάλαιο, το δυικό SVM εφαρμόζεται και σε προβλήματα ταξινόμησης με περισσότερες από δύο κλάσεις μέσω των μεθόδων one-against-all, one-against-one και των error-correcting output codes.
- Στο τέταρτο Κεφάλαιο, παρουσιάζεται μια μεγάλη ποικιλία κριτηρίων και μεθόδων που χρησιμοποιούνται για την επιλογή και την εκτίμηση μοντέλου όπως τα: AIC, BIC, διασταυρωμένη επικύρωση, διασταυρωμένα επικυρωμένο ποσοστό σφάλματος, ένδειξη υπολογισμένη με τη μέθοδο Laplace, ξα-εκτίμηση, GRM, MDL και bootstrap. Επίσης, αναλύεται το θεωρητικό τους υπόβαθρο και παρουσιάζονται τα αντίστοιχα πειραματικά αποτελέσματα για την απόδοσή τους.
- Στο πέμπτο Κεφάλαιο, παρουσιάζονται τα νέα Κριτήρια πληροφορίας για τα δυαδικά SVMs, και συγκεκριμένα τα: RIC, KRIC, SVMICa και SVMICb, το θεωρητικό τους υπόβαθρο, κάποιες προαπαιτούμενες έννοιες, η προσεγγιστική μέθοδος του Nystrom για τον υπολογισμό του KRIC και πειραματική σύγκριση της απόδοσής

τους με αυτή των κριτηρίων του τετάρτου Κεφαλαίου μέσω μοτίβων προσομοίωσης αλλά και δοκιμών σε σύνολα δεδομένων του πραγματικού κόσμου.

- Στο έκτο Κεφάλαιο, παρουσιάζονται δύο νέα Κριτήρια πληροφορίας για τα SVMs περισσοτέρων των δύο κλάσεων, τα οποία χρησιμοποιούν το σφάλμα ακτινωτού περιθωρίου για τα δυαδικά SVMs, κάποιες προαπαιτούμενες έννοιες, μια πειραματική σύγκριση της αποδοτικότητάς τους για την επιλογή μοντέλου στα multiclass SVMs, μια σύγκριση του υπολογιστικού τους κόστους με αυτό άλλων μεθόδων και τέλος κάποια πειραματικά αποτελέσματα για τα σύνολα δεδομένων αναφοράς.
- Στο έβδομο Κεφάλαιο, αναλύεται θεωρητικά ο Στατιστικός Έλεγχος Υποθέσεων και τα κριτήρια απόδοσης του μοντέλου (Confusion Matrix). Επιπλέον, αξιολογείται πειραματικά η εφαρμογή των μεθόδων διανυσματικής υποστήριξης σε προβλήματα δύο και τριών κλάσεων.
- Τέλος, στο όγδοο Κεφάλαιο παρουσιάζεται μια σύνοψη της εργασίας και εξάγονται κάποια γενικά συμπεράσματα.

Abstract

Progress in digital data acquisition and storage technology have resulted in an increase of massive databases. The analysis of these databases to find unsuspecting relationships and to summarize the data in novel ways that are understandable and useful, is Data Mining. The main objectives of the DM is the prediction and description, one of the ways to achieve them is the classification, i.e. learning a function that classifies a data item into one of the several predefined classes. In this thesis, we theoretically analyze the Support vector machines of two or more classes, which form a new classification method. The primary goal, however, of DM is the extraction of new information from the data, which can be achieved by various techniques which are distinguished in the supervised and unsupervised methods. The first category consists of the stages of model selection and assessment, for these two stages has been developed a wide variety of methods and criteria. Objective and main theme of this thesis is the presentation of these criteria and methods and the introduction of new information Criteria for binary and multiclass SVMs and the experimental comparison of their performance. Finally, the experimental application of the methods of support vector classification problems in two and three classes is evaluated.

More specifically:

- The first chapter is an introduction to the basic concepts of Data Mining, for instance: its procedure, the data components, its methods, the components of its algorithms, its categories and types of learning and its applications.
- In the second chapter, the theoretical background of binary SVMs is analyzed, and more specifically, the optimal separating hyperplane, the support vector classifier, the binary SVM, the kernels, the kernel trick, and finally the SVM as a penalization method.
- In the third chapter, the binary SVM is applied to classification problems with more than two classes through the methods of one-against-all, one-against-one and the error-correcting output codes.
- In the fourth chapter, a wide variety of criteria and methods used for the model selection and assessment, is analyzed, such as: AIC, BIC, cross validation, cross-validated error rate, evidence calculated by Laplace's method, ξα-estimation, GRM, MDL and bootstrap. Moreover, their theoretical background is analyzed and the corresponding experimental results for their performance are presented.
- In the fifth chapter the new information criteria for binary SVMs are presented, namely: RIC, KRIC, SVMICa and SVMICb, their theoretical background, some prerequisite concepts, the Nystrom approximation method for the calculation of the KRIC and the experimental comparison of their performance to the corresponding performance of the criteria of the fourth chapter through simulation patterns and tests on real data.

- In the sixth chapter, two new information criteria for SVMs of more than two classes, which use the radius-margin bound for binary SVMs, some prerequisite concepts, an experimental comparison of their efficiency for model selection in multiclass SVMs, a comparison of the computation load to that of other methods and finally some experimental results on the benchmark datasets are presented.
- In the Seventh Chapter, Statistical hypothesis testing and performance criteria of the model (Confusion Matrix) are analyzed. Moreover, the application of support vector methods to problems of two and three classes are experimentally evaluated.
- Finally, in the eighth chapter a summary of the thesis is presented and some general conclusions are exported.

Ευχαριστίες

Αρχικά, θα ήθελα να ευχαριστήσω θερμά τον Καθηγητή του Εθνικού Μετσόβιου Πολυτεχνείου κ. Χ. Κουκουβίνο για τη δυνατότητα ανάληψης της επικείμενης διπλωματικής εργασίας, την καθοδήγηση και την υποστήριξή του κατά τη διάρκεια της εκπόνησής της, καθώς επίσης και για τη συνεχή του προσφορά καθ' όλη τη διάρκεια των σπουδών μου που με οδήγησε στο να αγαπήσω τη Στατιστική.

Ιδιαίτερες ευχαριστίες θα ήθελα να εκφράσω στην υποψήφια διδάκτορα Χριστίνα Παρπούλα για την πολύτιμη βοήθειά της, το συμβουλευτικό της έργο και το συνεχές ενδιαφέρον κατά τη διάρκεια εκπόνησης της διπλωματικής μου εργασίας, καθώς επίσης και στην υποψήφια διδάκτορα Κρυσταλλένια Δρόσου για τις πολύτιμες συμβουλές της στο κομμάτι της Εφαρμογής.

Επιπλέον, θα ήθελα να ευχαριστήσω τη συμφοιτήτριά μου και διπλωματούχο του Εθνικού Μετσόβιου Πολυτεχνείου, Δανάη Γιαννούλη, για την πολύτιμη προσφορά της βοήθειάς της όποτε της εζητήθη.

Τέλος, θα ήθελα να εκφράσω την ευγνωμοσύνη μου και τις ευχαριστίες μου στην οικογένειά μου, στους φίλους μου, καθώς επίσης και στους συμφοιτητές μου για την αμέριστη υπομονή τους καθώς επίσης και τη διαρκή τους συμπαράσταση, βοήθεια και υποστήριξη κατά τη διάρκεια εκπόνησης αυτής της εργασίας, αλλά και καθ' όλη τη διάρκεια των σπουδών μου.

Στουφή Ευγενία

Εθνικό Μετσόβιο Πολυτεχνείο,
Σχολή Εφαρμοσμένων Μαθηματικών
και Φυσικών Επιστημών
Αθήνα, 2015

1

ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ

1.1 ΕΙΣΑΓΩΓΙΚΑ ΣΤΟΙΧΕΙΑ ΓΙΑ ΤΟ DATA MINING

Η πρόοδος στην ψηφιακή απόκτηση δεδομένων και στην τεχνολογία αποθήκευσης έχει ως αποτέλεσμα την αύξηση των τεράστιων βάσεων δεδομένων, από τα τετριμμένα (όπως τα δεδομένα συναλλαγών σούπερ μάρκετ, τα αρχεία χρήσης πιστωτικών καρτών, τα στοιχεία τηλεφωνικών κλήσεων και τα στατιστικά στοιχεία της κυβέρνησης) μέχρι τα πιο εξεζητημένα (όπως οι εικόνες των αστρονομικών σωμάτων, μοριακές βάσεις δεδομένων και τα ιατρικά αρχεία). Είναι επόμενο λοιπόν να έχει αυξηθεί το ενδιαφέρον για τη δυνατότητα χρησιμοποίησης αυτών των βάσεων δεδομένων, καθώς ο ιδιοκτήτης αυτών έχει το πλεονέκτημα εξαγωγής πιθανόν χρήσιμων πληροφοριών από αυτές.

Ο καθορισμός ενός επιστημονικού πεδίου είναι πάντα ένα αμφιλεγόμενο έργο, οι ερευνητές συχνά διαφωνούν για το ακριβές εύρος και τα όρια του τομέα των σπουδών τους .Έχοντας αυτά κατά νου και αποδεχόμενοι ότι οι υπόλοιποι μπορούν να διαφωνούν για τις λεπτομέρειες, θα υιοθετήσουμε ως ορισμό της εξόρυξης δεδομένων τον εξής:

Ορισμός :Η εξόρυξη δεδομένων είναι η ανάλυση των (συχνά μεγάλων) συνόλων παρατηρήσεων δεδομένων για την εύρεση ανυποψίαστων σχέσεων και τη σύνοψη των δεδομένων με νέους τρόπους που να είναι τόσο κατανοητοί όσο και χρήσιμοι στον ιδιοκτήτη των δεδομένων.

Σημείωση :Οι σχέσεις και οι περιλήψεις που προέρχονται μέσα από μία διαδικασία εξόρυξης δεδομένων συχνά αναφέρονται ως μοντέλα ή μοτίβα. Τα παραδείγματα περιλαμβάνουν γραμμικές εξισώσεις ,κανόνες, κλάσεις, γραφήματα, δομές δένδρων και επαναλαμβανόμενα μοντέλα σε χρονοσειρές .

Ο παραπάνω ορισμός αναφέρεται σε ‘’παρατηρησιακά” δεδομένα σε αντιδιαστολή με τα’’ πειραματικά” δεδομένα .Η εξόρυξη δεδομένων συνήθως ασχολείται με τα στοιχεία που έχουν ήδη συλλεχθεί για κάποιο σκοπό, εκτός από την ανάλυση δεδομένων (π.χ. μπορεί να έχουν συλλεχθεί προκειμένου να διατηρηθεί μια up-to-date ημερομηνία καταγραφής των συναλλαγών σε μία τράπεζα).Αυτό σημαίνει ότι οι στόχοι στην εξάσκηση εξόρυξης δεδομένων, δεν παίζουν κανένα ρόλο στη στρατηγική της συλλογής δεδομένων .Αυτός είναι ένας τρόπος με τον οποίο η εξόρυξη δεδομένων διαφέρει κατά πολύ από τη Στατιστική στην οποία τα δεδομένα συχνά συλλέγονται με τη χρήση αποτελεσματικών στρατηγικών για να απαντήσει σε συγκεκριμένα ερωτήματα. Για αυτό το λόγο η εξόρυξη δεδομένων συχνά αναφέρεται ως “δευτερεύουσα “ Ανάλυση Δεδομένων.

Ο ορισμός αναφέρει επίσης, ότι τα σύνολα δεδομένων που εξετάζονται στην Εξόρυξη Δεδομένων είναι συχνά μεγάλα. Αν συμμετείχαν μόνο μικρά σύνολα δεδομένων, θα

συζητάγαμε απλά για κλασική διερευνητική Ανάλυση Δεδομένων, όπως αυτή ασκείται από τους στατιστικούς. Όταν όμως βρισκόμαστε αντιμέτωποι με μεγάλα σύνολα δεδομένων, νέα προβλήματα προκύπτουν .Μερικά από αυτά αφορούν θέματα εκκαθάρισης ,αποθήκευσης ή πρόσβασης στα δεδομένα ,άλλα όμως σχετίζονται με πιο θεμελιώδη θέματα όπως με τον τρόπο καθορισμού της αντιπροσωπευτικότητας των δεδομένων, τον τρόπο ανάλυσης των δεδομένων σε ένα εύλογο χρονικό διάστημα και τον τρόπο απόφασης του κατά πόσο μια φαινομενική σχέση είναι ένα τυχαίο περιστατικό που δεν αντανακλά μια υποκειμενική πραγματικότητα. Συχνά τα διαθέσιμα στοιχεία δεν περιλαμβάνουν ένα μόνο δείγμα από το πλήρη πληθυσμό, ο στόχος μπορεί να είναι η γενίκευση από το δείγμα στον πληθυσμό. Για παράδειγμα, ίσως να επιθυμούσαμε να προβλέψουμε πώς είναι πιθανόν να συμπεριφερθούν μελλοντικοί πελάτες ή να καθορίσουμε πρωτεϊνικές δομές που δεν έχουμε δει ακόμα. Τέτοιες γενικεύσεις είναι πιθανό να μην είναι επιτεύξιμες μέσω τυπικών στατιστικών προσεγγίσεων, διότι τα δεδομένα δεν είναι κλασικά στατιστικά τυχαία δείγματα, αλλά μάλλον “ευκαιριακά” δείγματα ή δείγματα “ευκολίας”. Μερικές φορές μπορεί να θέλουμε να συνοψίσουμε ή να συμπιέσουμε ένα πολύ μεγάλο σύνολο δεδομένων με ένα τέτοιο τρόπο ώστε το αποτέλεσμα να είναι πιο κατανοητό, χωρίς καμία έννοια γενίκευσης .Αυτό το ζήτημα θα προέκυπτε, για παράδειγμα, αν είχαμε πλήρη στοιχεία απογραφής για μια συγκεκριμένη χώρα ή μια βάση δεδομένων εκατομμυρίων μεμονομένων συναλλαγών λιανικής .

Οι σχέσεις και οι δομές που θα βρίσκονται μέσα σε ένα σύνολο δεδομένων πρέπει, φυσικά να είναι καινούργιες. Δεν υπάρχει νόημα να αναμασά κανείς καθιερωμένες ή αναγκαίες σχέσεις εκτός αν στοχεύει στην επιβεβαίωση μιας υπόθεσης στην οποία κάποιος επιδιώκει να καθορίσει εάν ένα καθιερωμένο πρότυπο υπάρχει επίσης και στο νέο σύνολο δεδομένων . Σαφώς η καινοτομία θα πρέπει να μετράται σε σχέση με την εκ των προτέρων γνώση του χρήστη. Δυστυχώς λίγοι αλγόριθμοι εξόρυξης δεδομένων λαμβάνουν υπόψη την προηγούμενη γνώση του χρήστη.

Ενώ η καινοτομία είναι μια σημαντική ιδιότητα των σχέσεων που επιδιώκουμε, δεν είναι επαρκής στο να πληροί τις προϋποθέσεις μιας σχέσεως όπως της αξίζει. Ειδικότερα, πρέπει οι σχέσεις να είναι επίσης κατανοητές, για παράδειγμα, απλές σχέσεις γίνονται πιο εύκολα κατανοητές από τις πιο πολύπλοκες και ίσως να είναι προτιμότερες ή ισάξιες. Η εξόρυξη δεδομένων τοποθετείται συχνά στο ευρύτερο πλαίσιο της ανακάλυψης γνώσης σε βάσεις δεδομένων ή αλλιώς Knowledge Discovery Database (KDD). Αυτός ο όρος προέρχεται από το ερευνητικό πεδίο της τεχνητής νοημοσύνης (AI). Η διαδικασία KDD περιλαμβάνει διάφορα στάδια : την επιλογή των δεδομένων στόχου, την προεπεξεργασία των δεδομένων, την μετατροπή τους (αν είναι απαραίτητη),την εκτέλεση εξόρυξης δεδομένων για την εξαγωγή προτύπων και σχέσεων και στη συνέχεια την ερμηνεία και αξιολόγηση των ανακαλυφθέντων δομών. Και πάλι τα ακριβή όρια της εξόρυξης δεδομένων δεν είναι εύκολο να δηλωθούν. Για παράδειγμα, για πολλούς ανθρώπους η μετατροπή των δεδομένων είναι ένα αναπόσπαστο μέρος της εξόρυξης δεδομένων. Η διαδικασία της αναζήτησης των σχέσεων μέσα σε ένα σύνολο δεδομένων περιλαμβάνει έναν αριθμό βημάτων :

1.Τον προσδιορισμό της φύσης και της δομής της αναπαράστασης που θα χρησιμοποιηθεί .
2.Την απόφαση σχετικά με το πώς θα τροποποιηθούν τα δεδομένα και τη σύγκριση του πόσο καλά ταιριάζουν στις διαφορετικές αναπαραστάσεις τα δεδομένα (επιλέγοντας μια συνάρτηση βαθμολογίας)
3.Την επιλογή μιας αλγοριθμικής διαδικασίας για τη βελτιστοποίηση της συνάρτησης βαθμολογίας.
4.Την απόφαση ποιες αρχές της διαχείρισης των δεδομένων απαιτούνται για την αποτελεσματική εφαρμογή των αλγορίθμων.

1.2 ΔΙΑΔΙΚΑΣΙΑ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

Παρουσιάζουμε επιγραμματικά τα οκτώ κύρια στάδια της Διαδικασίας Εξόρυξης Δεδομένων :

(I)Ορισμός των στόχων και των προς ανάλυση αντικειμένων

(ii)Επιλογή, οργάνωση και επισκόπηση των δεδομένων

(iii)Διερευνητική ανάλυση των δεδομένων

(iv)Καθορισμός των στατιστικών μεθόδων

Υπάρχουν τρεις κύριες κατηγορίες μεθόδων :

(a)Περιγραφικές μέθοδοι

(b)Μέθοδοι πρόβλεψης

(c)Τοπικές μέθοδοι

(v)Ανάλυση των στοιχείων

(vi)Αξιολόγηση των στατιστικών μεθόδων

(vii)Ανάπτυξη του μοντέλου-Σταθεροποίηση της γνώσης-Παρουσίαση της γνώσης

(viii)Εφαρμογή των μεθόδων

1.3 ΣΥΝΙΣΤΩΣΕΣ ΔΕΔΟΜΕΝΩΝ

1.3.1 ΟΡΟΛΟΓΙΑ

Η κατανόηση της μορφής των δεδομένων και των εξαγόμενων, είναι ίσως περισσότερο σημαντική από τα ενδιάμεσα βήματα σε έναν αλγόριθμο εξόρυξης πληροφορίας

Συνιστώσες Δεδομένων:

(i) Αντίληψη (Concept): το αντικείμενο της μάθησης.

Στόχος: Εύρεση κατανοητής και λειτουργικής περιγραφής ενός concept.

(ii) Υπόδειγμα (Instance): το ξεχωριστό και ανεξάρτητο παράδειγμα ενός concept.

Σημείωση :Πιο σύνθετα σχήματα εισόδου είναι πιθανά .

(iii)Χαρακτηριστικό (attribute): η μετρήσιμη συνιστώσα ενός υποδείγματος. Κάθε υπόδειγμα αποτελείται από ένα προκαθορισμένο σύνολο στοιχείων τα οποία καλούνται χαρακτηριστικά (attributes), στην πράξη όμως ο αριθμός τους μπορεί να ποικίλει. Επίσης η ύπαρξη ενός χαρακτηριστικού μπορεί να εξαρτάται από την τιμή ενός άλλου.

1.3.2 ΤΥΠΟΙ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ

Πιθανοί τύποι χαρακτηριστικών ανάλογα με τα επίπεδα μέτρησής τους (levels of measurement) είναι οι εξής:

(A)Ονομαστικά (Nominal)

Οι τιμές τους είναι διακριτά σύμβολα, δηλαδή αποτελούν περιγραφή ή όνομα. Δεν υπάρχει ή συνεπάγεται καμιά συσχέτιση (κατάταξη ,μέτρο απόστασης) μεταξύ των ονομαστικών τιμών, μόνο η σύγκριση ως προς την ισότητα της τιμής του χαρακτηριστικού στα διάφορα υποδείγματα έχει νόημα

(Β)Τακτικά (Ordinal)

Σε αυτή την κατηγορία χαρακτηριστικών ορίζεται διάταξη στις τιμές, αλλά δεν ορίζεται απόσταση μεταξύ τους. Επίσης η πρόσθεση και αφαίρεση δεν έχουν νόημα, ενώ η διάκριση μεταξύ των κατηγοριών των χαρακτηριστικών Α και Β δεν είναι πάντα ευκρινής .

(Γ)Περιοδικά (interval)

Οι τιμές των περιοδικών χαρακτηριστικών, δεν είναι μόνο διατεταγμένες αλλά μετρώνται σε σταθερές και ισαπέχουσες τιμές. Η διαφορά ανάμεσα σε δύο τιμές έχει έννοια ωστόσο, το άθροισμα ή το γινόμενο δεν έχει έννοια καθώς δεν έχει οριστεί το σημείο μηδέν .

(Δ)Αναλογικά (ratio)

Η μέθοδος μέτρησης ορίζει σημείο μηδέν, όμως ο ορισμός του είναι μάλλον σχετικός παρά απόλυτος. Τα αναλογικά χαρακτηριστικά μεταχειρίζονται ως πραγματικοί αριθμοί (όλες οι μαθηματικές πράξεις είναι επιτρεπτές).

1.4 ΜΕΘΟΔΟΙ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

Οι δύο πρωταρχικοί στόχοι της εξόρυξης δεδομένων στην πράξη τείνουν να είναι η πρόβλεψη (prediction) και η περιγραφή (description). Η πρόβλεψη χρησιμοποιεί κάποιες μεταβλητές ή πεδία της βάσης δεδομένων για να προβλέψει άγνωστες ή μελλοντικές μεταβλητές από άλλες μεταβλητές ενδιαφέροντος και η περιγραφή εστιάζει στην εξεύρεση ανθρώπινα ερμηνεύσιμων μοτίβων που να περιγράφουν τα δεδομένα. Παρόλο που τα όρια μεταξύ της πρόβλεψης και της περιγραφής δεν είναι αυστηρά, η διάκριση είναι χρήσιμη για την κατανόηση του συνολικού στόχου ανακάλυψης. Η σημασία της πρόβλεψης και της περιγραφής, για συγκεκριμένες εφαρμογές της εξόρυξης δεδομένων, μπορεί να ποικίλει σημαντικά. Οι στόχοι της πρόβλεψης και της περιγραφής, μπορούν να επιτευχθούν με τη χρήση μιας ποικιλίας μεθόδων εξόρυξης δεδομένων.

Η κατάταξη (classification) είναι η εκμάθηση μιας λειτουργίας που ταξινομεί ένα στοιχείο δεδομένων σε μία από τις πολλές προκαθορισμένες κλάσεις .

Η παλινδρόμηση (regression) είναι η εκμάθηση μιας λειτουργίας που ταξινομεί ένα στοιχείο δεδομένων σε μια πραγματική αξιολογημένη μεταβλητή πρόβλεψης .

Η ομαδοποίηση (clustering) είναι μια κοινή περιγραφική διαδικασία, στην οποία κάποιος προσπαθεί να εντοπίσει πεπερασμένο σύνολο κατηγοριών ή συστάδων. Οι κατηγορίες μπορεί να είναι αμοιβαία αποκλειστικές και εξαντλητικές ή να αποτελούνται από μία πιο πλούσια αντιπροσώπευση, όπως η ιεραρχική ή οι επικαλυπτόμενες κατηγορίες.

Η συνόψιση (summarization) περιλαμβάνει μεθόδους για την εύρεση μιας συμπαγούς περιγραφής για ένα υποσύνολο των δεδομένων. Πιο εξελιγμένες μέθοδοι περιλαμβάνουν την παραγωγή των συνοπτικών κανόνων, πολυμεταβλητές τεχνικές απεικόνισης,και την ανακάλυψη των συσχετισμών μεταξύ των μεταβλητών .

Η μοντελοποίηση εξάρτησης (dependency modeling) συνίσταται στην εύρεση ενός μοντέλου που περιγράφει σημαντικές εξαρτήσεις μεταξύ των μεταβλητών. Υπάρχουν μοντέλα εξάρτησης σε δύο επίπεδα:

(1)Το διαρθρωτικό επίπεδο του μοντέλου καθορίζει συχνά σε γραφική μορφή ποιες μεταβλητές είναι τοπικά εξαρτώμενες .

(2)Το ποσοτικό επίπεδο του μοντέλου, καθορίζει τις αντοχές των εξαρτήσεων χρησιμοποιώντας κάποια αριθμητική κλίμακα.

Η αλλαγή και απόκλιση ανίχνευσης (change and deviation detection) επικεντρώνεται στην ανακάλυψη των πιο σημαντικών αλλαγών στα δεδομένα από τις προηγούμενα μετρημένες κανονιστικές τιμές.

1.5 ΤΑ ΣΥΣΤΑΤΙΚΑ ΤΩΝ ΑΛΓΟΡΙΘΜΩΝ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

Το επόμενο βήμα είναι να κατασκευάσουμε ειδικούς αλγορίθμους για να εφαρμόσουμε τις γενικές μεθόδους που περιγράψαμε. Μπορούν να εντοπιστούν τρία βασικά συστατικά σε κάθε αλγόριθμο εξόρυξης δεδομένων:

(1)Μοντέλο αξιολόγησης

(2) Μοντέλο αναπαράστασης

(3)'Ερευνα

Αυτή η περιοριστική άποψη δεν είναι πλήρης ή πλήρως περιεκτική, μάλλον είναι ένας βολικός τρόπος για να εκφράσει τις βασικές έννοιες των αλγορίθμων εξόρυξης με ένα σχετικά ενιαίο και συνεκτικό τρόπο. Η Αναπαράσταση του Μοντέλου είναι η γλώσσα που χρησιμοποιείται για να περιγράψει ανιχνεύσιμα μοτίβα. Αν η αναπαράσταση είναι πολύ περιορισμένη, τότε καμιά ποσότητα χρόνου εκπαίδευσης ή παραδειγμάτων δεν είναι ικανή να παράγει ένα ακριβές μοντέλο για τα δεδομένα. Είναι σημαντικό ένας αναλυτής δεδομένων να κατανοεί πλήρως τις αναπαραστατικές παραδοχές που μπορεί να είναι εγγενείς σε μια συγκεκριμένη μέθοδο. Είναι εξίσου σημαντικό ένας σχεδιαστής αλγορίθμων να δηλώνει ξεκάθαρα ποιες αναπαραστατικές παραδοχές, γίνονται σε ένα συγκεκριμένο αλγόριθμο. Να σημειωθεί ότι η αυξημένη παραστατική δύναμη για τα μοντέλα αυξάνει τον κίνδυνο της υπερπροσαρμογής των δεδομένων εκπαίδευσης και έχει ως αποτέλεσμα την μειωμένη ακρίβεια πρόβλεψης σε invisible δεδομένα.

Τα Κριτήρια Αξιολόγησης του Μοντέλου είναι ποσοτικές δηλώσεις του πόσο καλά ένα συγκεκριμένο μοτίβο (ένα μοντέλο και οι παράμετροί του) πληροί τους στόχους της διαδικασίας KDD. Για παράδειγμα τα μοντέλα πρόβλεψης συχνά κρίνονται από την εμπειρική ακρίβεια πρόβλεψης σε κάποια σύνολα ελέγχου. Τα περιγραφικά μοντέλα μπορούν να αξιολογηθούν όσον αφορά την προγνωστική ακρίβεια, την καινοτομία, τη χρησιμότητα και την κατανόηση του προσαρμοσμένου μοντέλου .

Οι μέθοδοι αναζήτησης αποτελούνται από δύο συνιστώσες :

(1) Αναζήτηση των παραμέτρων

(2) Αναζήτηση του μοντέλου

Μόλις το μοντέλο αναπαράστασης και τα κριτήρια αξιολόγησης του μοντέλου καθοριστούν, τότε το πρόβλημα εξόρυξης δεδομένων έχει περιοριστεί καθαρά σε ένα έργο βελτιστοποίησης, δηλαδή την εύρεση εκείνων των παραμέτρων και των μοντέλων από την επιλεγμένη οικογένεια που βελτιστοποιούν τα κριτήρια αξιολόγησης .

Στην αναζήτηση των παραμέτρων, ο αλγόριθμος πρέπει να ψάξει για τις παραμέτρους που βελτιστοποιούν τα κριτήρια αξιολόγησης με την προϋπόθεση ότι δίνονται τα παρατηρηθέντα δεδομένα και ένα σταθερό μοντέλο αναπαράστασης. Η αναζήτηση του μοντέλου

εμφανίζεται ως ένας βρόχος πάνω στη μέθοδο αναζήτησης παραμέτρων: το μοντέλο αναπαράστασης αλλάζει έτσι ώστε να θεωρηθεί μια οικογένεια μοντέλων .

1.6 ΚΑΤΗΓΟΡΙΕΣ ΤΟΥ DATA MINING-ΕΙΔΗ ΜΑΘΗΣΗΣ

1.6.1 ΚΑΤΗΓΟΡΙΕΣ ΤΟΥ DATA MINING

(A)Το κατευθυνόμενο (direct) Data Mining που εξηγεί ή ταξινομεί κάποια πεδία στόχους

(B)Το μη κατευθυνόμενο (indirect) Data Mining που βρίσκει ομοιότητες σε ομάδες εγγράφων χωρίς τη χρήση συγκεκριμένου πεδίου στόχου ή προκαθορισμένων κλάσεων

1.6.2 ΕΙΔΗ ΜΑΘΗΣΗΣ ΚΑΙ ΜΕΘΟΔΟΙ ΤΟΥ DATA MINING

Ο κύριος στόχος του Data Mining είναι η εξαγωγή νέων πληροφοριών από τα δεδομένα. Η ανακάλυψη της γνώσης γίνεται με τεχνικές που διακρίνονται σε δύο βασικές κατηγορίες:

(1)Μέθοδοι με επίβλεψη (supervised)

(2)Μέθοδοι χωρίς επίβλεψη (unsupervised)

(A1)Μέθοδοι με επίβλεψη

Αλγόριθμοι εκμάθησης με επίβλεψη είναι εκείνοι που χρησιμοποιούνται στην ταξινόμηση και στην πρόβλεψη. Ουσιαστικά μοντελοποιούν μια μεταβλητή απόκρισης βασιζόμενοι σε μια ή περισσότερες επεξηγηματικές μεταβλητές. Οι μέθοδοι με επίβλεψη χωρίζονται στις εξής υποκατηγορίες :

(I)Άμεσης μάθησης (παραμετρικές προσεγγίσεις)

Αποτελούν παραμετρικές μεθόδους μάθησης όπου η f(x) είναι γραμμική συνάρτηση των χαρακτηριστικών μεταβλητών. Σε αυτή την κατηγορία ανήκουν οι μέθοδοι γραμμικής παλινδρόμησης, συρρίκνωσης, λογιστικής παλινδρόμησης, διακριτής ανάλυσης και γενικευμένης διακριτής ανάλυσης.

(ii)Άμεσης μάθησης (ημί-μη παραμετρικές προσεγγίσεις)

Με τον όρο ημί-μη παραμετρικές προσεγγίσεις εννοούμε ότι το μοντέλο για την εκτίμηση της f(x) πλησιάζει στο να είναι μη παραμετρική, άλλά δεν είναι πλήρως μη παραμετρική. Μια λίστα τέτοιων μεθόδων περιλαμβάνει: τα νευρωνικά δίκτυα, την slice inverse παλινδρόμηση, τα γενικευμένα προσθετικά μοντέλα και τα πολυπαραγοντικά προσαρμοστικά splines παλινδρόμησης.

(iii)Άμεσης μάθησης (μη παραμετρικές προσεγγίσεις)

Σε αυτή την κατηγορία ανήκουν οι εξής μέθοδοι: Kernel, κρησαρίσματος, μέθοδοι που βασίζονται σε δέντρα, πολυπαραγοντικά προσαρμοστικά splines παλινδρόμησης και οι μέθοδοι κοντινότερου γείτονα.

(iv)Έμμεσης μάθησης

Οι μέθοδοι έμμεσης μάθησης εκτιμούν την f(x) με την ελαχιστοποίηση κάποιας λογικής συνάρτησης απώλειας αντί να εκτιμούν την f(x) άμεσα. Σε αυτή την κατηγορία μεθόδων ανήκουν: τα SVMs (που θα αναλυθούν εκτενώς σε επόμενη ενότητα),το ξεχωριστό υπερεπίπεδο, η προσέγγιση συνάρτησης μέσω συστηματοποίησης .

(A2)Επιλογή μοντέλου στη μάθηση με επίβλεψη

Όλες οι μέθοδοι εκμάθησης απαρτίζονται από δύο στάδια :

Στάδιο πρώτο: επιλογή μοντέλου

Επιλογή του καλύτερου μοντέλου με βάση συγκεκριμένα μέτρα απόδοσης.

Στάδιο δεύτερο: εκτίμηση μοντέλου

Εκτίμηση του σφάλματος πρόβλεψης του τελικού μοντέλου

Για την επιλογή και εκτίμηση του μοντέλου έχουν αναπτυχθεί αρκετά κριτήρια και μέθοδοι όπως τα κριτήρια των συντελεστών R^2, Cp Mallow's,τα κριτήρια πληροφορίας AIC και BIC, η μέθοδος MDL και η μέθοδος της διασταυρωμένης επικύρωσης.

(B)Μάθηση χωρίς επίβλεψη (unsupervised methods)

Αλγόριθμοι εκμάθησης χωρίς επίβλεψη είναι εκείνοι που χρησιμοποιούνται όταν δεν υπάρχει μια μεταβλητή απόκρισης να προβλεφθεί ή να ταξινομηθεί. Ουσιαστικά αυτού του τύπου οι μέθοδοι χρησιμοποιούνται όταν δεν υπάρχει κάποιο πεδίο να προβλεφθεί αλλά οι σχέσεις των δεδομένων εξερευνούνται ώστε να ανακαλυφθεί η γενική δομή τους. Μερικές από αυτές τις τεχνικές είναι: τα kohonen networks, two step, k means, η ανάλυση κυρίων συνιστωσών, η ανάλυση λανθάνοντος συστατικού, η πολυδιάστατη κλιμάκωση και η ανάλυση συστάδων.

1.7 ΕΦΑΡΜΟΓΕΣ ΤΟΥ DATA MINING

Το Data Mining (DM) βρίσκει εφαρμογή σε πάρα πολλούς επιστημονικούς κλάδους. Ο τομέας που εξυπηρετείται περισσότερο είναι ο χρηματοπιστωτικός (επιλογή δανειοληπτών) και οι τηλεπικοινωνίες (παρέχει ασφάλεια στον κυβερνοχώρο ανιχνεύοντας εισβολές μέσω της αναγνώρισης ασυνήθιστων λειτουργιών). Το DM έχει συμβάλλει σημαντικά στη διαχείριση του κόστους και στην αύξηση των εσόδων. Επίσης άλλοι τομείς που εξυπηρετούνται από το D.M είναι η Βιοϊατρική (το DM στην πρόβλεψη της δραστικότητας των φαρμάκων), η χημεία (μέσω της πρόβλεψης δομών των χημικών ενώσεων), η βιολογία (εντοπισμός γονιδίων σε γονιδίωμα). Στον τομέα της Ιατρικής το D.M κάνει εφικτή την παροχή επιστημονικών αποφάσεων για τη διάγνωση και θεραπεία μιας ασθένειας με τη σωστή διαχείριση των πληροφοριών από τη βάση δεδομένων του νοσοκομείου για την πρόβλεψη της αποτελεσματικότητας των χειρουργικών επεμβάσεων και των εξετάσεων αναπτύσσοντας έτσι την τηλεϊατρική. Ένας άλλος τομέας που εξυπηρετείται επίσης από το DM είναι αυτός της βιομηχανίας παραγωγής ενέργειας συγκεκριμένα βοηθά την έγκυρη πρόβλεψη της μελλοντικής ζήτησης ενέργειας, το διαχωρισμό του αργού πετρελαίου από το φυσικό αέριο και την κατασκευή πυρηνικών σφαιριδίων καύσιμου. Δύο άλλοι τομείς που επωφελούνται επίσης από το DM είναι αυτοί της οικολογίας (εντοπισμός πετρελαιοκηλίδων, έγκυρη προειδοποίηση οικολογικών καταστροφών και παράνομη απόρριψη αποβλήτων), της αστρονομίας (έλεγχος καταλογογράφησης ουράνιων αντικειμένων δυσδιάκριτων με οπτική

επιθεώρηση) και της βιομηχανίας (διάγνωση και πρόβλεψη βλαβών σε ηλεκτρομηχανολογικά συστήματα όπως κινητήρες και γεννήτριες)

Στη συνέχεια παραθέτουμε μερικές εφαρμογές του D.M σε συνδυασμό με παραδείγματα:

(1) Ανάλυση εταιριών και διαχείριση ρίσκου

I) Προβλέψεις

ii) Διαχείριση πελατολογίου

iii) Βελτιωμένη χρηματοδότηση

Παράδειγμα 1: Κατασκευή δένδρων αποφάσεων από ιστορικά στοιχεία τραπεζικών δανείων για την παραγωγή αλγορίθμων ώστε να αποφασίζεται αν πρέπει ή όχι να δοθεί ένα δάνειο σε υποψήφιο πελάτη .

Παράδειγμα 2: Ιατρικά εργαστήρια θέλουν να συσχετίσουν ασθένειες με χαρακτηριστικά των ασθενών όπως ο τόπος διαμονής, οι διατροφικές συνήθειες, παλαιότερες ασθένειες και άλλα ώστε να καταφέρουν να βγάλουν κάποια ιατρικά συμπεράσματα και καινούρια γνώση με τη βοήθεια των συγκεκριμένων χαρακτηριστικών.

(2) Ανάλυση αγοράς και διαχείρισης

(I)Target marketing

(ii) Customer relation manager

(iii) Market basket analysis (super market)

(iv) Cross selling

Παράδειγμα 1: Τράπεζες

Έλεγχος ποιότητας και ανάλυση ανταγωνιστικότητας

Παράδειγμα 2: Η περίπτωση “Diapers and Beer”.

Η παρατήρηση ότι οι πελάτες που αγοράζουν πάνες αγοράζουν και μπύρα, επιτρέπει στα καταστήματα να τοποθετούν αυτά τα είδη σχετικά κοντά γνωρίζοντας ότι οι πελάτες θα κάνουν τη διαδρομή μεταξύ των ραφιών με τις πάνες και αυτών με τις μπύρες. Τοποθετώντας ανάμεσά τους πατατάκια αυξάνουν τις πωλήσεις και στα τρία είδη.

Παράδειγμα 3: Εταιρία πωλήσεων ηλεκτρονικών συσκευών θέλει να μελετήσει τις αγοραστικές συνήθειες των πελατών της ώστε να προγραμματίσει ανάλογα την διαφημιστική καμπάνια .

(3) Εντοπισμός απάτης και διαχείρισης ρίσκου

Άλλες εφαρμογές που χρησιμοποιούν DM :

(I) Εξόρυξη κειμένου (news group, email, documents και Web Analysis)

(ii) Ευφυείς απαντήσεις σε ερωτήματα

Παράδειγμα 1: Άτομα που σκηνοθετούν ατυχήματα για να εισπράξουν από τις ασφαλιστικές εταιρίες ή κάποιοι που κάνουν ξέπλυμα “μαύρου χρήματος “ εντοπίζοντας ύποπτες μεταφορές χρημάτων ή κάποιοι που κλέβουν τους παρόχους τηλεπικοινωνιών και κάνουν τηλεφωνήματα που έχουν κάποια επαναλαμβανόμενα σχέδια είτε προς μία κλειστή ομάδα ατόμων (κινητά) είτε κάποια συγκεκριμένη ώρα της μέρας.

Παράδειγμα 2: Εντοπισμός ακατάλληλων ιατρικών μεθόδων και θεραπειών.

2

ΜΗΧΑΝΕΣ ΔΙΑΝΥΣΜΑΤΙΚΗΣ ΥΠΟΣΤΗΡΙΞΗΣ (SVMS)

2.1ΕΙΣΑΓΩΓΗ

Οι μηχανές διανυσματικής υποστήριξης (SVMs) είναι μια νέα μέθοδος ταξινόμησης που αναπτύχθηκε από τον Vapnik και την ομάδα του στα AT&T Bell Labs. Τα SVMs εφαρμόστηκαν σε προβλήματα ταξινόμησης ως εναλλακτικές μέθοδοι για τα πολυστρωματικά δίκτυα. Η ταξινόμηση πετυχαίνεται από μια γραμμική ή μια μη γραμμική διαχωριστική επιφάνεια στο χώρο εισόδου του συνόλου δεδομένων. Ο στόχος των SVMs είναι η ελαχιστοποίηση της προσδοκίας της εξόδου του σφάλματος δοκιμής. Τα SVMs καταγράφουν ένα δεδομένο σύνολο δυαδικών επιγεγραμμένων δεδομένων εκπαίδευσης σε ένα χώρο χαρακτηριστικών υψηλής διάστασης και διαχωρίζουν τις δύο κλάσεις των δεδομένων με ένα υπερεπίπεδο μεγίστου περιθωρίου. Για να καταλάβουμε την SVM προσέγγιση, πρέπει να ξέρουμε δύο έννοιες: τη δυαδικότητα και τους πυρήνες.

2.2 ΒΕΛΤΙΣΤΟ ΔΙΑΧΩΡΙΣΤΙΚΟ ΥΠΕΡΕΠΙΠΕΔΟ

Το βέλτιστο διαχωριστικό υπερεπίπεδο διαχωρίζει τις δύο κλάσεις και μεγιστοποιεί την απόσταση μέχρι τα πιο κοντινά σημεία από κάθε κλάση. Αυτό, με τη μεγιστοποίηση του περιθωρίου μεταξύ των δύο κλάσεων των δεδομένων εκπαίδευσης, οδηγεί σε καλύτερη απόδοση ταξινόμησης στα δεδομένα δοκιμής. Θεωρούμε το πρόβλημα βελτιστοποίησης:

$$\max_{\beta,\beta_0,\|\beta\|=1} M \qquad (2.2.1)$$

υποκείμενο σε $y_i\left(x_i^T\beta+\beta_0\right)\geq M, i=1,\dots,N$ (2.2.2)

Το σύνολο των συνθηκών εξασφαλίζει ότι όλα τα σημεία απέχουν τουλάχιστον μια ενδεικτική απόσταση M από το όριο απόφασης που ορίζεται από τα β $και$ β_0 και ψάχνουμε το μεγαλύτερο τέτοιο M και τις σχετικές παραμέτρους. Μπορούμε να ξεφορτωθούμε τον περιορισμό $\|\beta\|=1$ αντικαθιστώντας τις συνθήκες με:

$$\frac{1}{\|\beta\|}y_i\left(x_i^T\beta+\beta_0\right)\geq M \qquad (2.2.3)$$

Ή ισοδύναμα:

$$y_i\left(x_i^T\beta+\beta_0\right)\geq M\|\beta\| \qquad (2.2.4)$$

Από τη στιγμή που για κάθε β $και$ β_0 που ικανοποιούν αυτές τις ανισότητες, κάθε θετικά κλιμακωτό πολλαπλάσιο τις ικανοποιεί επίσης, μπορούμε αυθαίρετα να θέσουμε $\|\beta\|=\frac{1}{M}$. Ως εκ τούτου η (2.2.1) είναι ισοδύναμη με:

$$\min\frac{1}{2}\|\beta\|^2$$
$$y_i\left(x_i^T\beta+\beta_0\right)\geq M, i=1,\dots,N \qquad (2.2.5)$$

Οι περιορισμοί ορίζουν ένα περιθώριο γύρω από το γραμμικό όριο απόφασης πάχους $\frac{1}{\|\beta\|}$. Ως εκ τούτου, επιλέγουμε τα β $και$ β_0 για να μεγιστοποιήσουμε το πάχος του. Αυτό είναι ένα κυρτό πρόβλημα βελτιστοποίησης (τετραγωνικό κριτήριο με γραμμικούς ανισοτικούς περιορισμούς). Η συνάρτηση Langrange (πρωταρχική) που ελαχιστοποιείται υπό τους περιορισμούς β $και$ β_0 είναι:

$$L_p=\frac{1}{2}\|\beta\|^2-\sum_i^N a_i\left[y_i\left(x_i^T\beta+\beta_0\right)-1\right] \qquad (2.2.6)$$

Θέτοντας τις παραγώγους ίσες με μηδέν, παίρνουμε:

$$\beta=\sum_{i=1}^N a_i y_i x_i \qquad (2.2.7)$$

$$0=\sum_{i=1}^N a_i y_i \qquad (2.2.8)$$

και αντικαθιστώντας αυτές στην (2.2.6) παίρνουμε το λεγόμενο δυικό Wolfe:

$$L_D=\sum_{i=1}^N a_i-\frac{1}{2}\sum_{i=1}^N\sum_{k=1}^N a_i a_k y_i y_k x_i^T x_k \qquad (2.2.9)$$

Υπόκειται σε $a_i\geq 0$.

Αυτή η λύση λαμβάνεται από τη μεγιστοποίηση της L_D στο θετικό orthant, ένα απλούστερο κυρτό πρόβλημα βελτιστοποίησης, για το οποίο, τυπικό λογισμικό μπορεί να χρησιμοποιηθεί. Επιπροσθέτως η λύση πρέπει να ικανοποιεί τις συνθήκες Karush-Kuhn-Tucker, οι οποίες περιλαμβάνουν τις σχέσεις (2.2.7), (2.2.8), (2.2.9) και την

$$a_i\left[y_i\left(x_i^T\beta + \beta_0\right) - 1\right] = 0 \qquad (2.2.10)$$

Από αυτά μπορούμε να δούμε ότι:

- Εάν $a_i > 0$, τότε $y_i\left(x_i^T\beta + \beta_0\right) = 1$, ή διαφορετικά, τα x_i είναι στο όριο του περιθωρίου;
- Εάν $y_i\left(x_i^T\beta + \beta_0\right) > 1$, τα x_i δεν είναι στο όριο του περιθωρίου και τα $a_i = 0$.

Από τη σχέση (2.2.7) βλέπουμε ότι το διάνυσμα λύση β ορίζεται όσον αφορά ένα γραμμικό συνδυασμό των διανυσμάτων υποστήριξης x_i-αυτά τα σημεία που ορίζονται να είναι στο όριο του περιθωρίου μέσω των $a_i > 0$. Ομοίως το β_0 λαμβάνεται λύνοντας την (2.2.10) για κάθε ένα από τα σημεία υποστήριξης.

Το βέλτιστο διαχωριστικό υπερεπίπεδο παράγει μια συνάρτηση $\hat{f}(x) = x^T\hat{\beta} + \hat{\beta}_0$ για την ταξινόμηση των νέων παρατηρήσεων:

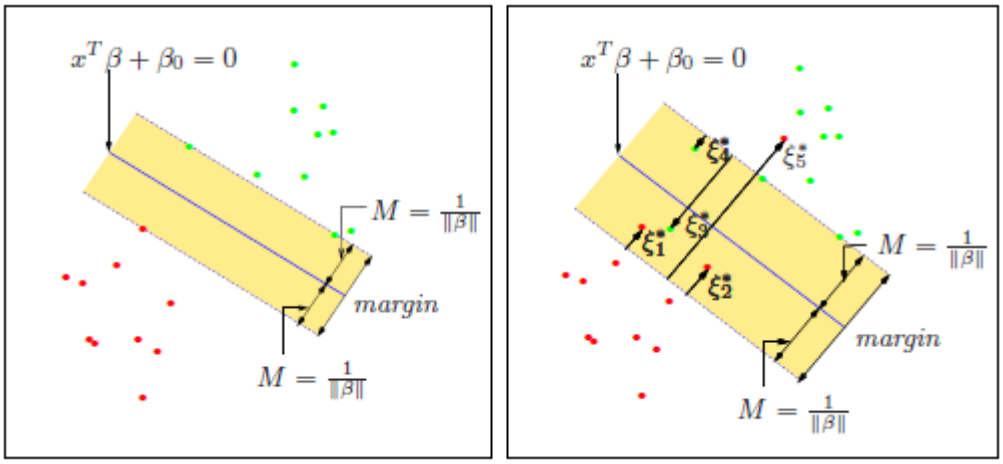
$$\hat{G}(x) = sign\hat{f}(x) \qquad (2.2.11)$$

Παρόλο που καμιά από τις παρατηρήσεις εκπαίδευσης δεν πέφτουν στο περιθώριο (από κατασκευής), αυτή δε θα είναι απαραιτήτως η περίπτωση για τις παρατηρήσεις δοκιμής. Η διαίσθηση είναι ότι ένα μεγάλο περιθώριο στα δεδομένα εκπαίδευσης θα οδηγήσει σε καλό διαχωρισμό στα δεδομένα δοκιμής.

Η περιγραφή της λύσης όσον αφορά τα σημεία υποστήριξης φαίνεται να υποδεικνύει ότι το βέλτιστο υπερεπίπεδο επικεντρώνεται περισσότερο στα σημεία που μετράνε και είναι περισσότερο εύρωστα στα ατελή μοντέλα. Να σημειωθεί, εντούτοις, ότι η αναγνώριση αυτών των σημείων υποστήριξης απαιτεί τη χρήση όλων των δεδομένων.

Όταν τα δεδομένα δεν είναι διαχωρίσιμα, δε θα υπάρχει εφικτή λύση για το πρόβλημα και ένας εναλλακτικός σχηματισμός χρειάζεται. Και πάλι μπορεί κανείς να μεγεθύνει το χώρο με τη χρήση μετασχηματισμών βάσης, αλλά αυτό μπορεί να οδηγήσει σε τεχνητό διαχωρισμό μέσω υπερπροσαρμογής.

2.3 Ο ΤΑΞΙΝΟΜΗΤΗΣ ΔΙΑΝΥΣΜΑΤΙΚΗΣ ΥΠΟΣΤΗΡΙΞΗΣ



ΣΧΗΜΑ 2.3.1: Ταξινομητές διανυσματικής υποστήριξης. Το αριστερό ταμπλό δείχνει τη διαχωρίσιμη περίπτωση. Το όριο απόφασης είναι η συμπαγής γραμμή, ενώ οι

διακεκομμένες γραμμές φράσσουν στο σκιασμένο μέγιστο περιθώριο εύρους 2M=$2/\|\boldsymbol{\beta}\|$. Το δεξί ταμπλό δείχνει τη μη διαχωρίσιμη περίπτωση (επικάλυψη). Τα σημεία με την ετικέτα ξ_j^*, βρίσκονται στη λάθος πλευρά του περιθωρίου τους κατά ένα ποσό $\xi_j^* = M\xi_j$; τα σημεία στη σωστή πλευρά έχουν $\xi_j^* = 0$. Το περιθώριο μεγιστοποιείται, υποκείμενο σε ένα συνολικό προϋπολογισμό $\sum \xi_i < σταθεράς$.
Ως εκ τούτου το $\sum \xi_j^*$ είναι η συνολική απόσταση των σημείων στη λάθος πλευρά του περιθωρίου τους

Τα δεδομένα εκπαίδευσής μας αποτελούνται από N ζεύγη $(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$, με $x_i \in \mathbb{R}^p$ και $y_i \in \{-1,1\}$. Ορίζουμε ένα υπερεπίπεδο από:

$$\{x: f(x) = x^T\beta + \beta_0 = 0\}, \tag{2.3.1}$$

όπου β είναι ένα μοναδιαίο διάνυσμα: $\|\beta\| = 1$. Ένας κανόνας ταξινόμησης που επάγεται από το $f(x)$ είναι:

$$G(x) = sign[x^T\beta + \beta_0] \tag{2.3.2}$$

Το $f(x)$ στη (2.3.1) δίνει την σημειωμένη απόσταση από ένα σημείο x μέχρι το υπερεπίπεδο $f(x) = x^T\beta + \beta_0 = 0$. Από τη στιγμή που οι κλάσεις είναι διαχωρίσιμες, μπορούμε να βρούμε μια συνάρτηση $f(x) = x^T\beta + \beta_0$ με $y_i f(x_i) > 0 \ \ \forall i$. Ως εκ τούτου μπορούμε να βρούμε το υπερεπίπεδο που δημιουργεί το μέγιστο περιθώριο μεταξύ των σημείων εκπαίδευσης για τις κλάσεις 1 και -1 (ΣΧΗΜΑ 2.2.1). Το πρόβλημα βελτιστοποίησης

$$\max_{\beta,\beta_0,\|\beta\|=1} M \tag{2.3.3}$$

υποκείμενο σε $y_i(x_i^T\beta + \beta_0) \geq M, i = 1, \dots, N,$
περιγράφει αυτή την έννοια. Η ρίγα στο σχήμα είναι M μονάδες μακριά από το υπερεπίπεδο σε κάθε πλευρά και ως εκ τούτου εύρους 2M μονάδων. Καλείται περιθώριο.

Δείξαμε ότι αυτό το πρόβλημα μπορεί πιο βολικά να αναδιατυπωθεί ως:

$$\max_{\beta,\beta_0} \|\beta\| \tag{2.3.4}$$

υποκείμενο σε $y_i(x_i^T\beta + \beta_0) \geq M, i = 1, \dots, N,$
όπου, ο πρότυπος περιορισμός αναφέρεται στο β. Να σημειωθεί ότι M=$1/\|\beta\|$. Η έκφραση (2.3.4) είναι ο συνήθης τρόπος γραφής του κριτηρίου διανύσματος υποστήριξης για διαχωρίσιμα δεδομένα. Αυτό είναι ένα κυρτό πρόβλημα βελτιστοποίησης (τετραγωνικό κριτήριο, γραμμικοί ανισοτικοί περιορισμοί).

Ας υποθέσουμε τώρα ότι οι κλάσεις επικαλύπτονται στο χώρο χαρακτηριστικών. Ένας τρόπος για να αντιμετωπίσουμε την επικάλυψη είναι να μεγιστοποιήσουμε το M, αλλά να επιτρέψουμε για κάποια σημεία να είναι στη λάθος πλευρά του περιθωρίου. Ορίζουμε τις χαλαρές (slack) μεταβλητές $\xi = (\xi_1, \xi_2, \dots, \xi_N)$. Υπάρχουν δύο φυσικοί τρόποι για να τροποποιήσουμε τους περιορισμούς στη (2.3.3) οι οποίοι περιγράφονται παρακάτω.

$$y_i(x_i^T\beta + \beta_0) \geq M - \xi_i \tag{2.3.5}$$

ή

$$y_i(x_i^T\beta + \beta_0) \geq M(1 - \xi_i) \tag{2.3.6}$$

$\forall i, \xi_i \geq 0, \sum_{i=1}^{N} \xi_i < \sigma\tau\alpha\theta\varepsilon\rho\acute{\alpha}\varsigma. O\iota$ δύο επιλογές οδηγούν σε διαφορετικές λύσεις. Η πρώτη επιλογή μοιάζει πιο φυσική, καθώς μετρά την επικάλυψη στην πραγματική απόσταση από το περιθώριο. Η δεύτερη επιλογή μετρά την επικάλυψη στη σχετική απόσταση, η οποία αλλάζει με το εύρος του περιθωρίου M. Εντούτοις, η πρώτη επιλογή έχει ως αποτέλεσμα ένα μη-κυρτό πρόβλημα βελτιστοποίησης, ενώ το δεύτερο είναι κυρτό. Έτσι η (2.3.6) οδηγεί στον ''πρότυπο'' ταξινομητή διανυσματικής υποστήριξης, τον οποίο θα χρησιμοποιούμε από εδώ και στο εξής.

Εδώ βασική ιδέα είναι η ιδέα του σχηματισμού. Η αξία ξ_i στον περιορισμό $y_i(x_i^T\beta + \beta_0) \geq M(1 - \xi_i)$ είναι το αναλογικό ποσό κατά το οποίο η πρόβλεψη $f(x_i) = x_i^T\beta + \beta_0$ βρίσκεται στη λάθος πλευρά του περιθωρίου της. Ως εκ τούτου φράσσοντας το άθροισμα $\sum \xi_i$, φράσσουμε το ολικό αναλογικό ποσό κατά το οποίο οι προβλέψεις πέφτουν στη λάθος πλευρά του περιθωρίου τους. Οι εσφαλμένες ταξινομήσεις συμβαίνουν όταν $\xi_i > 1$, έτσι φράσσοντας το $\sum \xi_i$ σε μια τιμή K σημαίνει ότι φράσσουμε τον ολικό αριθμό των εσφαλμένων ταξινομήσεων εκπαίδευσης στο K.

Έτσι, μπορούμε να ρίξουμε τον πρότυπο περιορισμό στο β, ορίζοντας M=$1/\|\beta\|$ και γράφουμε τη (2.3.4) στην ισοδύναμη μορφή:

$$\min \quad \|\beta\| \text{ υποκείμενο σε} \begin{cases} y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i \\ \xi_i \geq 0, \sum \xi_i < \sigma\tau\alpha\theta\varepsilon\rho\acute{\alpha}\varsigma \end{cases} \qquad (2.3.7)$$

Αυτός είναι ο συνήθης τρόπος που ορίζεται ο ταξινομητής διανυσματικής υποστήριξης για τη μη διαχωρίσιμη περίπτωση. Εντούτοις προκαλεί σύγχυση η παρουσία της σταθερής κλίμακας ''1'' στον περιορισμό $y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i$, και προτιμούμε να αρχίζουμε με την (2.3.6). Το δεξί ταμπλό στο ΣΧΗΜΑ 2.3.1 απεικονίζει αυτή την περίπτωση επικάλυψης. Από τη φύση του κριτηρίου (2.3.7), βλέπουμε ότι τα σημεία εντός του ορίου της τάξης δεν παίζουν μεγάλο ρόλο στο σχηματισμό του ορίου.

2.4 ΥΠΟΛΟΓΙΣΜΟΣ ΤΟΥ ΤΑΞΙΝΟΜΗΤΗ ΔΙΑΝΥΣΜΑΤΙΚΗΣ ΥΠΟΣΤΗΡΙΞΗΣ

Το πρόβλημα (2.3.7) είναι τετραγωνικό με γραμμικούς ανισοτικούς περιορισμούς, ως εκ τούτου είναι ένα κυρτό πρόβλημα βελτιστοποίησης. Περιγράφουμε μια λύση τετραγωνικού προγραμματισμού με τη χρήση πολλαπλασιαστών Lagrange. Υπολογιστικά είναι βολικό να εκφράσουμε ξανά τη (2.3.7) στην παρακάτω ισοδύναμη μορφή:

$$\min_{\beta,\beta_0} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N} \xi_i \qquad (2.4.1)$$

Υπόκειται σε $\xi_i \geq 0, y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i, \forall i$

όπου η παράμετρος ''κόστους'' C αντικαθιστά τη σταθερά στη (2.3.7). Η διαχωρίσιμη περίπτωση αντιστοιχεί στο C=∞.

Η συνάρτηση Lagrange (πρωταρχική) είναι η

$$L_p = \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} a_i\left[y_i(x_i^T\beta + \beta_0) - (1 - \xi_i)\right] - \sum_{i=1}^{N} \mu_i \xi_i \qquad (2.4.2)$$

την οποία ελαχιστοποιούμε ως προς τα β, β_0 και ξ_i. Θέτοντας τις αντίστοιχες παραγώγους ίσες με το μηδέν, παίρνουμε:

$$\beta=\sum_{i=1}^{N} a_i\, y_i x_i, \tag{2.4.3}$$
$$0=\sum_{i=1}^{N} a_i\, y_i \tag{2.4.4}$$
$$a = C - \mu_i \tag{2.4.5}$$

καθώς επίσης και τους περιορισμούς θετικότητας $a_i, \mu_i, \xi_i \geq 0\ \forall i$. Αντικαθιστώντας τις (2.4.3)-(2.4.5) στην (2.4.2), λαμβάνουμε την ακόλουθη Λαγκρανζιανή (Wolfe) δυική αντικειμενική συνάρτηση:

$$L_D = \sum_{i=1}^{N} a_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{i'=1}^{N} a_i a_i' y_i y_i' x_i^T x_i', \tag{2.4.6}$$

η οποία δίνει ένα χαμηλότερο φράγμα της αντικειμενικής συνάρτησης (2.4.1) για κάθε εφικτό σημείο. Μεγιστοποιούμε την L_D που υπόκειται στους $0 < a_i < C$ και $\sum_{i=1}^{N} a_i\, y_i = 0$. Επιπροσθέτως των (2.4.3)-(2.4.5), οι συνθήκες Karush-Kuhn-Tucker περιλαμβάνουν τους περιορισμούς:

$$a_i\left[y_i\left(x_i^T\beta + \beta_0\right) - (1 - \xi_i)\right] = 0 \tag{2.4.7}$$
$$\mu_i \xi_i = 0 \tag{2.4.8}$$
$$y_i\left(x_i^T\beta + \beta_0\right) - (1 - \xi_i) \geq 0 \tag{2.4.9}$$

για $i = 1, \dots, N$. Μαζί αυτές οι εξισώσεις (2.4.7)-(2.4.9) χαρακτηρίζουν μοναδικά το πρωταρχικό και δυικό πρόβλημα.

Από την (2.4.3) βλέπουμε ότι η λύση για το β έχει τη μορφή:

$$\hat{\beta} = \sum_{i=1}^{N} \hat{a}_i y_i\, x_i \tag{2.4.10}$$

Με μη μηδενικούς συντελεστές $\hat{a}_i$, μόνο για εκείνες τις παρατηρήσεις i για τις οποίες οι περιορισμοί στην (2.4.9) πληρούνται ακριβώς (λόγω της (2.4.7)). Αυτές οι παρατηρήσεις καλούνται διανύσματα υποστήριξης (support vectors-SVs), καθώς το $\hat{\beta}$ αναπαρίσταται όσον αφορά αυτές και μόνο. Μεταξύ αυτών των σημείων υποστήριξης, κάποια βρίσκονται στην άκρη του περιθωρίου ($\hat{\xi}_i = 0$) και ως εκ τούτου από τις (2.4.8) και (2.4.5) θα χαρακτηρισθούν από $0 < \hat{a}_i < C$ (μη φραγμένα SVs). Τα υπόλοιπα ($\hat{\xi}_i > 0$) έχουν $\hat{a}_i = C$ (φραγμένα SVs). Από την (2.4.7) μπορούμε να δούμε ότι οποιοδήποτε από αυτά τα σημεία περιθωρίου ($0 < \hat{a}_i, \hat{\xi}_i = 0$) μπορούν να χρησιμοποιηθούν στην επίλυση για β_0 και τυπικά χρησιμοποιούμε ένα μέσο όλων των λύσεων για αριθμητική ευστάθεια.

Η μεγιστοποίηση του δυικού (2.4.6) είναι ένα απλούστερο πρόβλημα κυρτού τετραγωνικού προγραμματισμού από το πρωταρχικό (2.4.2) και μπορεί να επιλυθεί με τυπικές τεχνικές.

Δεδομένων των λύσεων $\hat{\beta}_0$ και $\hat{\beta}$, η συνάρτηση απόφασης μπορεί να γραφεί ως:

$$\begin{aligned}\hat{G}(x) &= sign\left[\hat{f}(x)\right] \\ &= sign\left[x^T\hat{\beta} + \hat{\beta}_0\right].\end{aligned} \tag{2.4.11}$$

Η παράμετρος συντονισμού της διαδικασίας είναι η παράμετρος ‘’κόστους’’ C.

2.5 ΥΠΟΛΟΓΙΣΜΟΣ ΤΟΥ SVM ΓΙΑ ΤΑΞΙΝΟΜΗΣΗ ΚΑΙ ΣΥΝΑΡΤΗΣΕΙΣ ΠΥΡΗΝΑ (KERNEL FUNCTIONS)

Μπορούμε να αναπαραστήσουμε το πρόβλημα βελτιστοποίησης (2.4.2) και τη λύση του με ένα ειδικό τρόπο που περιλαμβάνει μόνο τα χαρακτηριστικά εισόδου μέσω των εσωτερικών προϊόντων. Το κάνουμε απευθείας για τα μετασχηματισμένα διανύσματα χαρακτηριστικών $h(x_i)$. Έπειτα βλέπουμε ότι για συγκεκριμένες επιλογές του h, αυτά τα ενδότερα προϊόντα μπορούν να υπολογιστούν με ελάχιστο κόστος.

Η δυική συνάρτηση Lagrange (2.4.6) έχει τη μορφή:

$$L_D = \sum_{i=1}^{N} a_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{i'=1}^{N} a_i a_{i'} y_i\, y_{i'} \langle h(x_i), h(x_{i'})\rangle \qquad (2.5.1)$$

Από τη (2.4.3) βλέπουμε ότι η λύση της συνάρτησης $f(x)$ μπορεί να γραφτεί:

$$\begin{aligned} f(x) &= h(x)^T\beta + \beta_0 \\ &= \sum_{i=1}^{N} a_i y_i \langle h(x), h(x_i)\rangle + \beta_0 \end{aligned} \qquad (2.5.2)$$

Όπως και προηγουμένως, δεδομένων των a_i, το β_0 μπορεί να καθοριστεί λύνοντας την $y_i f(x_i) = 1$ στην (2.5.2) για κάθε x_i για το οποίο $0 < \alpha_\iota < C$.

Έτσι και η (2.5.1) και η (2.5.2) περιλαμβάνουν το h(x) μόνο μέσω ενδότερων προϊόντων. Στην πραγματικότητα, δεν προσδιορίζουμε το μετασχηματισμό h(x) καθόλου, αλλά απαιτούμε μόνο γνώση της συνάρτησης πυρήνα (kernel function)

$$K(x, x') = \langle h(x), h(x')\rangle, \forall\, x, x' \qquad (2.5.3)$$

η οποία υπολογίζει τα ενδότερα προϊόντα στο μετασχηματισμένο χώρο. Η K πρέπει να είναι μια συμμετρική (ημί-) ορισμένη συνάρτηση.

<u>Θεώρημα του Mercer</u>: Έστω X ένα συμπαγές υποσύνολο του $\mathbb{R}^n$. Ας υποθέσουμε ότι K είναι μια συνεχής συμμετρική συνάρτηση τέτοια ώστε ο ακέραιος τελεστής $T_K{:}L_2(X) \to L_2(X)$,

$$(T_K f)(\cdot) \coloneqq \int_X K(\cdot, x) f(x)\, dx \qquad (2.5.4)$$

να είναι θετικός, αυτό είναι

$$\int_{X\times X} K(x, x') f(x) f(x')\, dx\, dx' \geq 0 \qquad (2.5.5)$$

για όλες τις $f \in L^2(X)$. Έπειτα μπορούμε να επεκτείνουμε το $K(x, x')$ σε μια ομοιομόρφως συγκλίνουσα σειρά (στο $X \times X$) όσον αφορά τις ιδιοσυναρτήσεις του T_K , $\varphi_k \in L_2(X)$, κανονικοποιημένο με τέτοιο τρόπο που $\|\varphi_k\|_{L_2} = 1$, και θετικά συσχετισμένες ιδιοσυναρτήσεις $\lambda_k \geq 0$,

$$K(x, x') = \sum_{k=1}^{\infty} \lambda_k \varphi_k(x) \varphi_k(x') \qquad (2.5.6)$$

Ορισμός Kernel Matrix: Ο πίνακας πυρήνα είναι η κεντρική δομή στις μηχανές πυρήνα και αναμειγνύει πληροφορίες για τα δεδομένα και τους πυρήνες:

K(1,1)	K(1,2)	K(1,3)	...	K(1,m)
K(2,1)	K(2,2)	K(2,3)		K(2,m)
...	...	...	...	...
K(m,1)	K(m,2)	K(m,3)	...	K(m,m)

Εισάγουμε κάποια παραδείγματα συναρτήσεων πυρήνα.

1. Παράδειγμα: Γκαουσιανός πυρήνας (Ακτινικής Βάσης Συνάρτηση Πυρήνα ή σφαιρικός GRBF πυρήνας)

$$K(x,x') = exp\left(\frac{-\|x-x'\|}{2\sigma^2}\right) \qquad (2.5.7)$$

όπου σ>0, είναι η παράμετρος πυρήνα γνωστή και ως Γκαουσιανό εύρος.

2. Παράδειγμα: Ελλειψοειδής GRBF πυρήνας

$$K(x,x') = exp\left(-\sum_{i=1}^{N}\frac{\|x_i-x_i'\|}{2\sigma_i{}^2}\right) \qquad (2.5.8)$$

όπου το σ_i, είναι το Γκαουσιανό εύρος για την i-οστή διάσταση

3. Παράδειγμα: ο πολυωνυμικός πυρήνας

$$K(x,x') = (1 + c\langle x,x'\rangle)^d \qquad (2.5.9)$$

όπου $d \in N$ και $c > 0$.

4. Παράδειγμα: ο πυρήνας νευρωνικού δικτύου (σιγμοειδής πυρήνας)

$$K(x,x') = tanh(b\langle x,x'\rangle - c) \qquad (2.5.10)$$

όπου b>0 και $c \in \mathbb{R}$.

Οι παράμετροι στους πυρήνες, όπως είναι οι *σ,c,d* και *b*, παραπάνω, ονομάζονται παράμετροι πυρήνα.

Ο Γκαουσιανός πυρήνας και ο πολυωνυμικός πυρήνας είναι πυρήνες Mercer. Ο πυρήνας νευρωνικού δικτύου δεν είναι ένας πυρήνας Mercer για κάθε b και c.

Από τη (2.4.2) βλέπουμε ότι η λύση μπορεί να γραφεί ως

$$\hat{f}(x) = \sum_{i=1}^{N}\hat{a}_i y_i K(x,x') + \hat{\beta}_0 \qquad (2.5.11)$$

Ο ρόλος της παραμέτρου C είναι πιο ξεκάθαρος σε ένα διευρυμένο χώρο χαρακτηριστικών, καθώς ο τέλειος διαχωρισμός είναι συχνά εφικτός εκεί. Μια μεγάλη τιμή της C θα αποθαρρύνει κάθε θετικό ξ_i και θα οδηγήσει σε ένα κυματοειδές όριο υπερπροσαρμογής στον αρχικό χώρο χαρακτηριστικών; μια μικρή

τιμή της C θα ενθαρρύνει μια μικρή τιμή του $\|\beta\|$, η οποία με τη σειρά της προκαλεί την $f(x)$ και ως εκ τούτου το όριο να είναι ομαλότερο.

Χρησιμοποιώντας τις συναρτήσεις πυρήνα, το δυικό πρόβλημα των SVMs γίνεται ως ακολούθως.

<u>Το δυικό πρόβλημα των SVMs με συνάρτηση πυρήνα στο χώρο χαρακτηριστικών</u>

$$\sum_{i=1}^{N} a_i - \frac{1}{2}\sum_{i,i'=1}^{N} a_i a_i' y_i y_i' K(x_i, x_i') \qquad (2.5.12)$$

$$\upsilon\pi\acute{o}\ \tau o\upsilon\varsigma\ \pi\varepsilon\rho\iota o\rho\iota\sigma\mu o\acute{\upsilon}\varsigma\ C \geq a_i \geq 0\ \gamma\iota\alpha\ i = 1,..,N \qquad (2.5.13)$$

$$\sum_{i=1}^{N} a_i y_i = 0$$

2.6 ΤΕΧΝΑΣΜΑ ΠΥΡΗΝΑ (KERNEL TRICK)

Αντικαθιστώντας τη (2.4.10) στη (2.4.11) παίρνουμε τη συνάρτηση απόφασης :

$$\hat{G}(x) = sign\sum_{i=1}^{N}\left(\hat{a_i} y_i x_i x^T + \hat{\beta}_0\right) \qquad (2.6.1)$$

και ξαναγράφουμε την (2.6.1) όσον αφορά τις συναρτήσεις πυρήνα παίρνοντας την ακόλουθη εξίσωση:

$$\hat{G}(x) = sign\sum_{i=1}^{N}\left(\hat{a_i} y_i K(x, x') + \hat{\beta}_0\right) \qquad (2.6.2)$$

Αυτό είναι γνωστό ως το τέχνασμα πυρήνα (kernel trick). Δηλαδή, με τη χρήση μιας κατάλληλης συνάρτησης πυρήνα, μπορούμε να επωφεληθούμε από τις χαρτογραφήσεις (mappings) στο χώρο χαρακτηριστικών χωρίς να πρέπει να υπολογίσουμε τις αντίστοιχες χαρτογραφήσεις, καθώς οι υπολογισμοί στο χώρο χαρακτηριστικών πάντα απλοποιούνται σε υπολογισμούς στο χώρο εισόδου. Με τη συνετή επιλογή των συναρτήσεων πυρήνα μπορούμε να ελέγξουμε την πολυπλοκότητα του μοντέλου. Το τέχνασμα βασίζεται στην εύρεση του κατάλληλου πυρήνα προκειμένου να κατασκευαστεί ένα μοντέλο για ένα συγκεκριμένο σύνολο δεδομένων.

Ο λόγος που το τέχνασμα είναι εύχρηστο , είναι ότι υπάρχουν πολλά προβλήματα ταξινόμησης που δεν είναι γραμμικά διαχωρίσιμα στο χώρο των εισόδων x, οι οποίες μπορεί να βρίσκονται σε ένα χώρο χαρακτηριστικών υψηλότερης διάστασης.

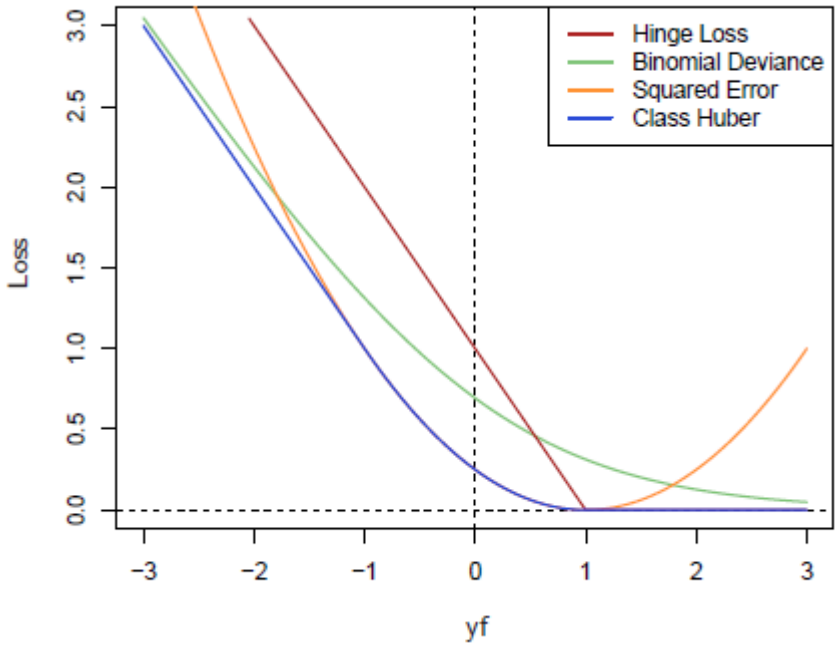
2.7 Οι SVMs ΩΣ ΜΕΘΟΔΟΣ ΠΟΙΝΙΚΟΠΟΙΗΣΗΣ

Με $f(x) = x^T\beta + \beta_0$, θεωρούμε το πρόβλημα βελτιστοποίησης:

$$\min_{\beta_0,\beta}\sum_{i=1}^{N}[1 - y_i f(x_i)]_+ + \frac{\lambda}{2}\|\beta\|^2 \qquad (2.7.1)$$

όπου ο υποδείκτης ''+'' υποδεικνύει θετικό μέρος. Αυτό έχει τη μορφή $\alpha\pi\acute{\omega}\lambda\epsilon\iota\alpha + \pi o\iota\nu\acute{\eta}$, που είναι οικείο παράδειγμα στην εκτίμηση συναρτήσεων. Είναι εύκολο να δείξουμε ότι η λύση της (2.7.1), με $\lambda = 1/C$, είναι η ίδια όπως αυτή για τη (2.4.1).

Η επιλογή της hinge συνάρτησης απώλειας $L(y, f) = [1 - yf]_+$ δείχνει να είναι η πιο λογική επιλογή για την ταξινόμηση δύο τάξεων, όταν συγκρίνεται με άλλες πιο παραδοσιακές συναρτήσεις απώλειας. Το ΣΧΗΜΑ (2.7.1) τη συγκρίνει με την απώλεια λογαριθμικής πιθανότητας για λογιστική παλινδρόμηση καθώς επίσης και με το τετραγωνικό σφάλμα απώλειας και με μια παραλλαγή αυτής. Η (αρνητική) λογαριθμική πιθανότητα ή η διωνυμική διακύμανση έχουν παρόμοιες ουρές με την SVM απώλεια, δίνοντας μηδενική ποινή σε σημεία εντός του περιθωρίου τους, και μια γραμμική ποινή σε σημεία στη λάθος πλευρά και μακριά. Το τετραγωνικό σφάλμα από την άλλη πλευρά δίνει τετραγωνική ποινή, και σημεία εντός του περιθωρίου τους που έχουν επίσης ισχυρή επίδραση στο μοντέλο. Το τετράγωνο της hinge απώλειας $L(y, f) = [1 - yf]_+^2$ είναι σαν την τετραγωνική, εκτός του ότι είναι μηδέν για σημεία εντός του περιθωρίου τους. Ακόμα αυξάνεται τετραγωνικά στην αριστερή ουρά και θα είναι λιγότερο εύρωστη από την ''εξαρτημένη'' ή τη διακύμανση στις εσφαλμένα ταξινομημένες παρατηρήσεις. Πρόσφατα προτάθηκε μια Huberized εκδοχή της τετραγωνικής hinge συνάρτησης απώλειας, η οποία μετατρέπεται ομαλά σε μια γραμμική απώλεια στο yf=-1.



ΣΧΗΜΑ 2.7.1: Η συνάρτηση απώλειας διανυσματικής υποστήριξης (hinge loss), συγκρίνεται με την αρνητική απώλεια λογαριθμικής πιθανότητας (binomial deviance) για τη λογιστική παλινδρόμηση, την απώλεια τετραγωνικού σφάλματος και μια ''Huberized'' εκδοχή της τετραγωνικά hinge απώλειας. Όλες αναπαριστώνται ως μια συνάρτηση του yf αντί του f, εξαιτίας της συμμετρίας μεταξύ της περίπτωσης y=+1 και της y=-1. Η διακύμανση και η Huberized έχουν τις ίδιες asymptotes όπως η SVM απώλεια, αλλά είναι στρογγυλοποιημένες στο εσωτερικό. Όλες είναι κλιμακωτές στο να έχουν περιοριστική κλίση αριστερής ουράς -1.

Μπορούμε να χαρακτηρίσουμε αυτές τις συναρτήσεις απώλειας όσον αφορά το τι εκτιμούν στο πληθυσμιακό επίπεδο. Θεωρούμε την ελαχιστοποίηση του $EL(Y, f(X))$. Ο πίνακας του ΣΧΗΜΑΤΟΣ 2.7.2 συνοψίζει τα αποτελέσματα. Ενώ η hinge απώλεια εκτιμά τον ίδιο τον ταξινομητή G(x), όλες οι άλλες εκτιμούν ένα μετασχηματισμό της κλάσης των εκ

των υστέρων πιθανοτήτων. Η Huberized τετραγωνική εξαρτημένη απώλεια μοιράζεται ελκυστικές ιδιότητες της λογιστικής παλινδρόμησης (ομαλή συνάρτηση απώλειας, εκτιμά πιθανότητες), καθώς επίσης και η SVM hinge απώλεια (σημεία υποστήριξης).

Loss Function	$L[y, f(x)]$	Minimizing Function
Binomial Deviance	$\log[1 + e^{-yf(x)}]$	$f(x) = \log \frac{\Pr(Y = +1\|x)}{\Pr(Y = -1\|x)}$
SVM Hinge Loss	$[1 - yf(x)]_+$	$f(x) = \text{sign}[\Pr(Y = +1\|x) - \frac{1}{2}]$
Squared Error	$[y - f(x)]^2 = [1 - yf(x)]^2$	$f(x) = 2\Pr(Y = +1\|x) - 1$
"Huberised" Square Hinge Loss	$-4yf(x), \quad yf(x) < -1$ $[1 - yf(x)]_+^2 \quad$ otherwise	$f(x) = 2\Pr(Y = +1\|x) - 1$

ΣΧΗΜΑ 2.7.2: Οι πληθυσμιακοί ελαχιστοποιητές για τις διαφορετικές συναρτήσεις απώλειας στο ΣΧΗΜΑ 2.7.1. Η λογιστική παλινδρόμηση χρησιμοποιεί τη διωνυμική λογαριθμική πιθανότητα ή τη διακύμανση. Η SVM hinge απώλεια εκτιμά την επιλογή των εκ των υστέρων πιθανοτήτων κλάσης, ενώ οι άλλες εκτιμούν ένα γραμμικό μετασχηματισμό αυτών των πιθανοτήτων.

Η διατύπωση της (2.7.1) καθιστά το SVM ως ένα πρόβλημα εκτίμησης συστηματοποιημένης συνάρτησης, όπου οι συντελεστές συσχέτισης της γραμμικής επέκτασης $f(x) = \beta_0 + h(x)^T\beta$, συρρικνώνονται μέχρι το μηδέν (εξαιρουμένης της σταθεράς). Αν η h(x) αναπαριστά μια ιεραρχική βάση έχοντας κάποια οργανωμένη δομή (όπως οργανωμένη στην τραχύτητα (roughness)), τότε η ομοιόμορφη συρρίκνωση έχει περισσότερο νόημα αν τα τραχύτερα στοιχεία h_j στο διάνυσμα h έχουν μικρότερη νόρμα.

Όλες οι συναρτήσεις απώλειας του ΣΧΗΜΑΤΟΣ 2.7.2, εκτός του τετραγωνικού σφάλματος είναι οι λεγόμενες ''συναρτήσεις απώλειας μεγιστοποίησης περιθωρίου''. Αυτό σημαίνει ότι εάν τα δεδομένα είναι διαχωρίσιμα τότε το όριο του $\hat{\beta}_\lambda$ στη (2.7.1) καθώς λ→0 ορίζει το βέλτιστο διαχωριστικό υπερεπίπεδο.

3

ΜΗΧΑΝΕΣ ΔΙΑΝΥΣΜΑΤΙΚΗΣ ΥΠΟΣΤΗΡΙΞΗΣ ΠΟΛΛΑΠΛΩΝ ΚΛΑΣΕΩΝ (MULTI-CLASS SVMS)

3.1 ΕΙΣΑΓΩΓΗ

Μέχρι τώρα έχουμε υποθέσει ότι τα δεδομένα είναι μόνο σε δύο κλάσεις. Πολλές πρακτικές εφαρμογές περιλαμβάνουν περισσότερες κλάσεις. Για παράδειγμα, η χειρόγραφη ψηφιακή αναγνώριση θεωρεί σύνολα δέκα κλάσεων, π.χ. ψηφία 0 έως 9. Υπάρχουν πολλοί τρόποι να επεκτείνουμε τα SVMs για τέτοιες περιπτώσεις. Στη συνέχεια, θα εξεταστούν δύο απλές μέθοδοι.

3.2 ONE-AGAINST-ALL MULTI-CLASS SVM

Αυτή η μέθοδος κατασκευάζει δυαδικά SVM μοντέλα, έτσι ώστε κάθε ένα από αυτά να εκπαιδεύεται με μια κλάση ως θετική και τις υπόλοιπες ως αρνητικές. Απεικονίζουμε αυτή τη μέθοδο με την απλή περίπτωση των τεσσάρων κλάσεων.

Τα 4 SVMs δύο κλάσεων είναι:

$y_i = 1$	$y_i = -1$	**Συναρτήσεις απόφασης**
Κλάση 1	Κλάσεις 2,3,4	$f^1(x) = (w^1)^T x + b^1$
Κλάση 2	Κλάσεις 1,3,4	$f^2(x) = (w^2)^T x + b^2$
Κλάση 3	Κλάσεις 1,2,4	$f^3(x) = (w^3)^T x + b^3$
Κλάση 4	Κλάσεις 1,2,3	$f^4(x) = (w^4)^T x + b^4$

ΣΧΗΜΑ 3.2.1

Για κάθε τεστ δεδομένων x, αν είναι στην i-οστή κλάση, θα μπορούσαμε να περιμένουμε ότι:

$$f^i(x) \geq 1 \; \kappa\alpha\iota \; f^j(x) \leq -1, \varepsilon\acute{\alpha}\nu \; j \neq i. \qquad (3.2.1)$$

Η "προσδοκία" αυτή προκύπτει άμεσα από τη ρύθμιση της εκπαίδευσης των 4 προβλημάτων δύο κλάσεων και κάτω από την υπόθεση ότι τα δεδομένα διαχωρίζονται σωστά. Ως εκ τούτου, η $f^i(x)$ έχει τις μεγαλύτερες τιμές μεταξύ των $f^1(x), \dots, f^4(x)$ και έτσι ο κανόνας απόφασης είναι ο παρακάτω:

Προβλεπόμενη κλάση= $arg\ \max_{i=1,..,4} f'(x)$ (3.2.3)

3.3 ONE-AGAINST-ONE MULTICLASS SVM

Αυτή η μέθοδος κατασκευάζει επίσης διάφορες SVMs δύο κλάσεων αλλά κάθε μία προκύπτει από την εκπαίδευση των δεδομένων από δύο μόνο διαφορετικές κλάσεις. Για το ίδιο πρόβλημα των τεσσάρων κλάσεων κατασκευάζονται έξι προβλήματα των δύο κλάσεων:

$\boldsymbol{y_i = 1}$	$\boldsymbol{y_i = -1}$	Συναρτήσεις απόφασης
Κλάση 1	Κλάση 2	$f^{12}(x) = (w^{12})^T x + b^{12}$
Κλάση 1	Κλάση 3	$f^{13}(x) = (w^{13})^T x + b^{13}$
Κλάση 1	Κλάση 4	$f^{14}(x) = (w^{14})^T x + b^{14}$
Κλάση 2	Κλάση 3	$f^{23}(x) = (w^{23})^T x + b^{23}$
Κλάση 2	Κλάση 4	$f^{24}(x) = (w^{24})^T x + b^{24}$
Κλάση 3	Κλάση 4	$f^{34}(x) = (w^{34})^T x + b^{34}$

ΣΧΗΜΑ 3.3.1

Για κάθε τεστ δεδομένων x, το θέτουμε σε 6 συναρτήσεις. Αν το πρόβλημα των κλάσεων i και j υποδεικνύει ότι τα δεδομένα x θα έπρεπε να είναι στο i, η κλάση i παίρνει μία ψήφο για παράδειγμα, υποθέτουμε ότι:

Κλάσεις	Νικητής
1 2	1
1 3	1
1 4	1
2 3	2
2 4	4
3 4	3

ΣΧΗΜΑ 3.3.2

Τότε, έχουμε:

Κλάση	1	2	3	4
# ψήφων	3	1	1	1

ΣΧΗΜΑ 3.3.3

Έτσι, το x προβλέπεται να είναι στην πρώτη κλάση. Για ένα σύνολο δεδομένων με k διαφορετικές κλάσεις, αυτή η μέθοδος κατασκευάζει $k(k-1)/2$ SVMs δύο κλάσεων.

Κάποιες φορές, αν και όχι τόσο συχνά, περισσότερες από μία κλάσεις κατέχουν τον υψηλότερο αριθμό ψήφων, αλλά υπάρχουν περαιτέρω στρατηγικές για να διαχειριστούμε αυτή την κατάσταση.

3.4 ΥΠΟΛΟΓΙΣΤΙΚΗ ΠΟΛΥΠΛΟΚΟΤΗΤΑ ΓΙΑ MULTI-CLASS SVM

Καθώς η μέθοδος one-against-one εκπαιδεύει $k(k-1)/2$ ταξινομητές, όπου k ο αριθμός των κλάσεων, μπορούμε να υποθέσουμε ότι αυτή η μέθοδος παίρνει περισσότερο χρόνο εκπαίδευσης. Στην πραγματικότητα, αυτό δεν ισχύει. Ας υποθέσουμε ότι ο χρόνος για την εκπαίδευση ενός προβλήματος δύο κλάσεων είναι $(\cdot)^d$, όπου $(\cdot)$ ο αριθμός των μεταβλητών. Έπειτα οι χρόνοι εκπαίδευσης των δύο προσεγγίσεων πολλαπλών κλάσεων συγκρίνονται στον πίνακα (3.4.1). Αν $d \geq 2$ η one-against-one προσέγγιση έχει συντομότερο χρόνο εκπαίδευσης.

Μέθοδος	One-against-all	One-against-one
Μέσο μέγεθος ανά SVM δύο κλάσεων	l	$2l/k$
# SVMs δύο κλάσεων	k	$k(k-1)/2$
Χρόνος εκπαίδευσης	$kO(l^d)$	$\frac{k(k-1)}{2} O\left(\left(\frac{2l}{k}\right)^d\right)$

ΣΧΗΜΑ 3.4.1

Ως εκ τούτου, εάν το ποσοστό των δεδομένων εκπαίδευσης που είναι διανύσματα υποστήριξης, στα k και στα $k(k-1)/2$ SVMs δύο κλάσεων είναι περίπου το ίδιο και το k δεν είναι υπερβολικά μεγάλο, τότε οι δύο προσεγγίσεις πολλαπλών κλάσεων έχουν παρόμοιο υπολογιστικό χρόνο.

3.5 ΣΥΖΗΤΗΣΗ

Εδώ εξετάσαμε δύο μεθόδους που μας επιτρέπουν να εφαρμόσουμε το εγγενώς δυικό SVM σε προβλήματα ταξινόμησης με περισσότερες από δύο κλάσεις. Δύο άλλες μέθοδοι αναφέρονται συχνά στη βιβλιογραφία: η error-correcting-output-codes ταξινόμηση και η multiobjective μηχανή διανυσματικής υποστήριξης. Η πρώτη επεκτείνει την one-against-all μέθοδο ταξινόμησης επιτρέποντας έναν καλύτερο έλεγχο του πώς κατασκευάζονται επιφάνειες απόφασης και έπειτα χρησιμοποιούνται στην ταξινόμηση άγνωστων σημείων. Η δεύτερη επεκτείνει τη θεωρία των SVMs απευθείας από δυαδικά μοντέλα σε μοντέλα πολλαπλών κλάσεων, έχοντας ως αποτέλεσμα ένα multiobjective πρόβλημα βελτιστοποίησης σαν έναν αλγόριθμο εκπαίδευσης. Και οι δύο προσεγγίσεις έχουν ωραίες θεωρητικές ιδιότητες, αλλά δε χρησιμοποιούνται συχνά στην πράξη, εξαιτίας της αυξημένης υπολογιστικής τους πολυπλοκότητας.

3.6 ERROR-CORRECTING OUTPUT CODES

Οι ECOC προτάθηκαν από τους Dietterich & Bakiri το 1995. Η βασική ιδέα είναι ότι σε κάθε κλάση εκχωρείται μία μοναδική δυαδική ακολουθία μήκους n, εκπαιδεύονται n ταξινομητές ένας για κάθε bit και στη συνέχεια τρέχουν n ταξινομητές στο x, για να πάρουν μια ακολουθία S των n bit και επιλέγεται η κλάση με το πιο κοντινό S.

Class	Code Word					
	vl	hl	dl	cc	ol	or
0	0	0	0	1	0	0
1	1	0	0	0	0	0
2	0	1	1	0	1	0
3	0	0	0	0	1	0
4	1	1	0	0	0	0
5	1	1	0	0	1	0
6	0	0	1	1	0	1
7	0	0	1	0	0	0
8	0	0	0	1	0	0
9	0	0	1	1	0	0

ΣΧΗΜΑ 3.6.1: Παράδειγμα

Η σημασία κάθε στήλης φαίνεται στο ακόλουθο σχή**μα:**

Column position	Abbreviation	Meaning
1	vl	contains vertical line
2	hl	contains horizontal line
3	dl	contains diagonal line
4	cc	contains closed curve
5	ol	contains curve open to left
6	or	contains curve open to right

ΣΧΗΜΑ 3.6.2:

Class	Code Word														
	f_0	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}	f_{12}	f_{13}	f_{14}
0	1	1	0	0	0	0	1	0	1	0	0	1	1	0	1
1	0	0	1	1	1	1	0	1	0	1	1	0	0	1	0
2	1	0	0	1	0	0	0	1	1	1	1	0	1	0	1
3	0	0	1	1	0	1	1	1	0	0	0	0	1	0	1
4	1	1	1	0	1	0	1	1	0	0	1	0	0	0	1
5	0	1	0	0	1	1	0	1	1	1	0	0	0	0	1
6	1	0	1	1	1	0	0	0	0	1	0	1	0	0	1
7	0	0	0	1	1	1	1	0	1	0	1	1	0	0	1
8	1	1	0	1	0	1	1	0	0	1	0	0	0	1	1
9	0	1	1	1	0	0	0	0	1	0	1	0	0	1	1

ΣΧΗΜΑ 3.6.3: Παράδειγμα ενός κώδικα 15-bit για ένα πρόβλημα 10 κλάσεων

<u>Ορισμός απόσταση Hamming:</u> Η απόσταση Hamming μεταξύ δύο σειρών ίσου μήκους είναι ο αριθμός των θέσεων για τις οποίες τα αντίστοιχα σύμβολα διαφέρουν.

Για την επιλογή ενός καλού error-correcting κώδικα, επιλέγουμε αυτόν με την ελάχιστη απόσταση Hamming, μεταξύ οποιουδήποτε ζεύγους κωδικών λέξεων. Αν η απόσταση Hamming είναι d τότε ο κώδικας μπορεί να διορθώσει τουλάχιστον $(d-1)/2$ σφάλματα ενός bit.

Ένας καλός ECOC έχει τις εξής δύο ιδιότητες:

1. Διαχωρισμός γραμμών: Κάθε κωδική λέξη, πρέπει να είναι καλά διαχωρισμένη στην απόσταση Hamming από κάθε άλλη από τις κωδικές λέξεις.
2. Διαχωρισμός στηλών: Κάθε συνάρτηση bit-θέσης f_i πρέπει να είναι ασυσχέτιστη με κάθε άλλη f_j.

Class	Code Word							
	f_0	f_1	f_2	f_3	f_4	f_5	f_6	f_7
c_0	0	0	0	0	1	1	1	1
c_1	0	0	1	1	0	0	1	1
c_2	0	1	0	1	0	1	0	1

ΣΧΗΜΑ 3.6.4: Όλες οι πιθανές στήλες για ένα πρόβλημα 3 κλάσεων.

Εάν υπάρχουν k κλάσεις θα υπάρχουν το πολύ $2^{k-1}-1$, στήλες που μπορούν να χρησιμοποιηθούν μετά την αφαίρεση των συμπληρωμάτων και των στηλών που περιέχουν μόνο στήλες ή μόνο μηδενικά.

4

ΚΡΙΤΗΡΙΑ ΠΛΗΡΟΦΟΡΙΑΣ

4.1 ΕΙΣΑΓΩΓΗ

Η απόδοση γενίκευσης της μαθησιακής μεθόδου σχετίζεται με την ικανότητα πρόβλεψης των ανεξάρτητων δοκιμών δεδομένων. Η αξιολόγηση αυτής της απόδοσης είναι εξαιρετικά σημαντική στην πράξη καθώς καθοδηγεί την επιλογή της μαθησιακής μεθόδου ή του μοντέλου και μας δίνει ένα μέτρο της ποιότητας του επιλεχθέντος μοντέλου

4.2 ΜΕΡΟΛΗΨΙΑ ΔΙΑΚΥΜΑΝΣΗ ΚΑΙ ΠΟΛΥΠΛΟΚΟΤΗΤΑ ΜΟΝΤΕΛΟΥ

Έχουμε μία συνάρτηση στόχου Y, ένα διάνυσμα εισόδου X και ένα μοντέλο πρόβλεψης $\hat{f}(X)$ που έχει εκτιμηθεί από το σύνολο εκπαίδευσης T. Η συνάρτηση απώλειας για τη μέτρηση των λαθών μεταξύ των Y και $\hat{f}(X)$ δηλώνεται από το $L\left(Y, \hat{f}(X)\right)$. Τυπικές επιλογές είναι:

$$L\left(Y, \hat{f}(X)\right) = \begin{cases} \left(Y - \hat{f}(X)\right)^2, \textit{τυπικό σφάλμα} \\ \left|Y - \hat{f}(X)\right|, \ \textit{απόλυτο σφάλμα} \end{cases} \tag{4.2.1}$$

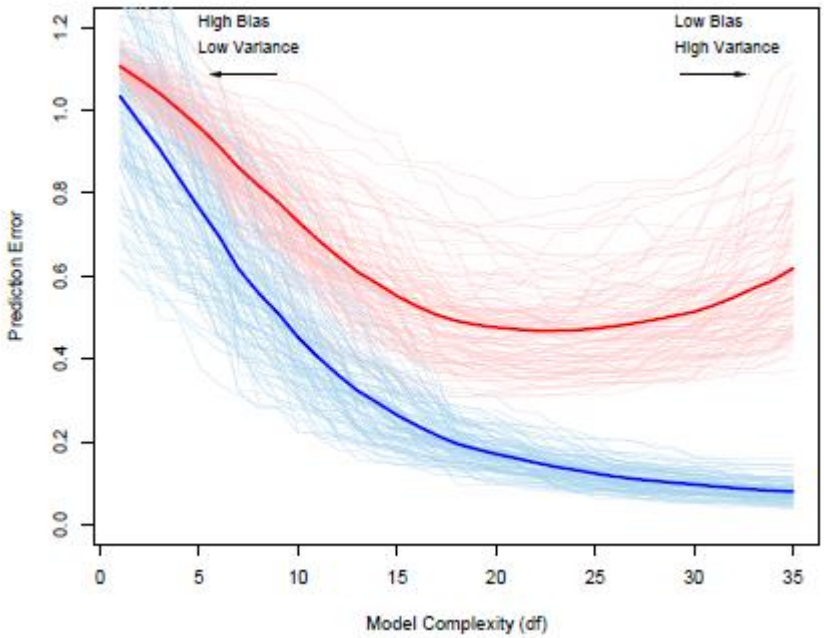
Το σφάλμα δοκιμής, επίσης αναφέρεται ως σφάλμα γενίκευσης και αποτελεί το σφάλμα πρόβλεψης σε ένα ανεξάρτητο δείγμα δοκιμής

$$Err_T = \mathrm{E}\left[L\left(Y - \hat{f}(X)\right)|T\right], \tag{4.2.2}$$

όπου τα X και Y επιλέγονται τυχαία από την από κοινού κατανομή (πληθυσμού). Εδώ το T είναι σταθερό και το σφάλμα δοκιμής αναφέρεται στο σφάλμα για αυτό το συγκεκριμένο σύνολο εκπαίδευσης. Μια σχετική ποσότητα είναι το αναμενόμενο σφάλμα πρόβλεψης ή δοκιμής:

$$Err = \mathrm{E}\left[L\left(Y - \hat{f}(X)\right)\right] = E[\,Err_T] \qquad (4.2.3)$$

Δυστυχώς το σφάλμα εκπαίδευσης δεν αποτελεί καλή εκτίμηση του σφάλματος δοκιμής, καθώς αυτό μειώνεται σταθερά με την πολυπλοκότητα του μοντέλου και πέφτει στο μηδέν με την αύξηση της πολυπλοκότητας του μοντέλου. Ωστόσο ένα μοντέλο με μηδενικό σφάλμα εκπαίδευσης είναι υπερπροσαρμοσμένο στα δεδομένα εκπαίδευσης και τυπικά θα γενικεύσουμε ανεπαρκώς.



ΣΧΗΜΑ 4.2.1 : Το σχήμα δείχνει το σφάλμα πρόβλεψης Err_T για 100 προσομοιωμένα σύνολα εκπαίδευσης, το κάθε ένα μεγέθους 50, καθώς η πολυπλοκότητα του μοντέλου αυξάνεται. Οι συμπαγείς καμπύλες δείχνουν το αναμενόμενο σφάλμα δοκιμής Err και το αναμενόμενο σφάλμα εκπαίδευσης E($\overline{err}$).

Το σφάλμα εκπαίδευσης είναι η μέση απώλεια δείγματος εκπαίδευσης:

$$\overline{err} = \frac{1}{N}\sum_{i=1}^{N} L\left(y_i, \hat{f}(x_i)\right) \qquad (4.2.4)$$

Καθώς το μοντέλο γίνεται όλο και πιο πολύπλοκο, χρησιμοποιεί τα δεδομένα εκπαίδευσης περισσότερο και είναι ικανό να προσαρμοστεί σε πιο υποβόσκουσες δομές. Έτσι υπάρχει μείωση στη μεροληψία αλλά αύξηση στη διακύμανση. Το σφάλμα εκπαίδευσης δεν αποτελεί καλή εκτίμηση του σφάλματος δοκιμής καθώς μειώνεται συστηματικά με την πολυπλοκότητα του μοντέλου.

Τυπικά το μοντέλο μας θα έχει μία παράμετρο ή παραμέτρους συντονισμού και έτσι μπορούμε να γράψουμε τις προβλέψεις μας ως $\hat{f}_a(x)$. Η παράμετρος συντονισμού προκαλεί την ποικιλομορφία της πολυπλοκότητας του μοντέλου και έτσι αναζητούμε την τιμή του α που ελαχιστοποιεί το σφάλμα. Στην πραγματικότητα έχουμε δύο ξεχωριστούς στόχους:

1) Επιλογή μοντέλου : Εκτιμούμε την απόδοση διαφορετικών μοντέλων έτσι ώστε να επιλεχθεί το καλύτερο

2) Αξιολόγηση του μοντέλου: Έχοντας επιλέξει το τελικό μοντέλο, εκτιμούμε το σφάλμα πρόβλεψης-γενίκευσης στα νέα δεδομένα.

Εάν βρισκόμαστε σε κατάσταση πλούσια σε δεδομένα, η καλύτερη προσέγγιση και για τα δύο προβλήματα είναι να χωρίσουμε τυχαία το σύνολο δεδομένων σε τρία μέρη: ένα σύνολο εκπαίδευσης, ένα επικύρωσης και ένα ελέγχου τα οποία χρησιμοποιούνται αντιστοίχως για την προσαρμογή του μοντέλου, την εκτίμηση του σφάλματος πρόβλεψης για την επιλογή μοντέλου και την αξιολόγηση του σφάλματος γενίκευσης του τελικά επιλεγμένου μοντέλου. Είναι δύσκολο να δοθεί ένας γενικός κανόνας στο πώς να επιλέγουμε τον αριθμό των παρατηρήσεων σε κάθε ένα από τα τρία μέρη καθώς αυτό εξαρτάται από την αναλογία του σήματος προς το θόρυβο στα δεδομένα και το μέγεθος του δείγματος εκπαίδευσης. Μια τυπική διαίρεση θα μπορούσε να είναι 50% για εκπαίδευση και από 25% για επικύρωση και εκπαίδευση.

4.3 Η ΑΠΟΣΥΝΘΕΣΗ ΜΕΡΟΛΗΨΙΑΣ-ΔΙΑΚΥΜΑΝΣΗΣ

Ας υποθέσουμε ότι $Y = f(x) + \varepsilon$, όπου Ε(ε) και $Var(\varepsilon) = \sigma_\varepsilon^2$, μπορούμε να παράγουμε μια έκφραση για το αναμενόμενο σφάλμα πρόβλεψης μιας προσαρμογής παλινδρόμησης $\hat{f}(X)$, σε ένα σημείο εισόδου X=x_0, με τη χρήση της απώλειας τετραγωνικού σφάλματος:

$$Err(x_0) = E\left[\left(Y - \hat{f}(x_0)\right)^2 | X = x_0\right]$$
$$= \sigma_\varepsilon^2 + \left[E\hat{f}(x_0) - f(x_0)\right]^2 + E\left[\hat{f}(x_0) - E\hat{f}(x_0)\right]^2$$
$$= \sigma_\varepsilon^2 + Bias^2\left(\hat{f}(x_0)\right) + Var\left(\hat{f}(x_0)\right)$$
$$= Irreducible\ Error + Bias^2 + Variance$$

(4.3.1)

Ο πρώτος όρος είναι η διακύμανση της μεταβλητής στόχου από τον πραγματικό μέσο $f(x_0)$ και δεν μπορεί να παραληφθεί οσοδήποτε καλά και να εκτιμάται το $f(x_0)$ εκτός και αν $\sigma_\varepsilon^2 = 0$.Ο δεύτερος όρος είναι το τετράγωνο της μεροληψίας και εκφράζει το ποσό κατά το οποίο ο μέσος της εκτίμησής μας διαφέρει από τον πραγματικό μέσο και ο τελευταίος όρος είναι η διακύμανση και εκφράζει την αναμενόμενη τετραγωνική απόκλιση του $\hat{f}(x_0)$ από το μέσο του. Τυπικά όσο πιο πολύπλοκο είναι το μοντέλο $\hat{f}$, τόσο πιο χαμηλή είναι η (τετραγωνική) μεροληψία και τόσο υψηλότερη η διακύμανση.

Στην προσαρμογή παλινδρόμησης k-πλησιέστερων-γειτόνων, αυτές οι εκφράσεις έχουν την εξής απλή μορφή :

$$Err(x_0) = E\left[\left(Y - \hat{f}(x_0)\right)^2 | X = x_0\right]$$

$$= \sigma_\varepsilon^2 + \left[f(x_0) - \frac{1}{k}\sum_{l=1}^{k} f\left(x_{(l)}\right)\right]^2 + \frac{\sigma_\varepsilon^2}{k} \qquad (4.3.2)$$

Υποθέτουμε για ευκολία ότι οι είσοδοι εκπαίδευσης x_i είναι σταθεροί και η τυχαιότητα προκύπτει από τα y_i. Ο αριθμός των k γειτόνων είναι αντιστρόφως ανάλογος της πολυπλοκότητας του μοντέλου. Για μικρά k η εκτίμηση $\hat{f}_k(x)$ μπορεί να προσαρμοστεί

καλύτερα στο $f(x)$. Αυξάνοντας το k, η μεροληψία τυπικά αυξάνεται ενώ μειώνεται η διακύμανση.

Για ένα μοντέλο γραμμικής παλινδρόμησης $\hat{f}_p(x) = x^T\hat{\beta}$, όπου το διάνυσμα της παραμέτρου β με p συνιστώσες προσαρμόζεται με τη βοήθεια των ελαχίστων τετραγώνων, έχουμε :

$Err(x_0) = E\left[\left(Y - \hat{f}(x_0)\right)^2 |X = x_0\right]$. Εδώ το $h(x_0) = X(X^TX)^{-1}x_0$, είναι το N-διάστατο διάνυσμα των γραμμικών βαρών που παράγει την προσαρμογή $\hat{f}_p(x_0) = x_0^T\, X(X^TX)^{-1}X^Ty$ και ως εκ τούτου $Var\left[\hat{f}_p(x_0)\right] = \|h(x_0)\|^2\sigma_\varepsilon^2$. Ενώ η διακύμανση αλλάζει με το x_0, ο μέσος της είναι $(p/N)\sigma_\varepsilon^2$. Και ως εκ τούτου έχουμε ότι η ακόλουθη παράσταση

$$\frac{1}{N}\sum_{i=1}^{N} Err(x_i) = \sigma_\varepsilon^2 + \frac{1}{N}\sum_{i=1}^{N}\left[f(x_i) - E\hat{f}(x_i)\right]^2 + \left(\frac{p}{N}\right)\sigma_\varepsilon^2 \qquad (4.3.3)$$

αποτελεί το σφάλμα εντός δείγματος. Εδώ το μοντέλο πολυπλοκότητας είναι ευθέως ανάλογο με τον αριθμό των παραμέτρων p.

4.4 ΒΕΛΤΙΩΣΗ ΤΟΥ ΠΟΣΟΣΤΟΥ ΣΦΑΛΜΑΤΟΣ ΕΚΠΑΙΔΕΥΣΗΣ

Τυπικά το σφάλμα εκπαίδευσης έχει την ακόλουθη μορφή:

$$\overline{err} = \frac{1}{N}\sum_{i=1}^{N} L\left(y_i, \hat{f}(x_i)\right) \qquad (4.4.1)$$

Είναι μικρότερο από το πραγματικό σφάλμα Err_T , διότι τα ίδια δεδομένα χρησιμοποιούνται για να προσαρμόζονται στη μέθοδο και να αξιολογούν το σφάλμα της. Μια μέθοδος προσαρμογής τυπικά προσαρμόζεται στα δεδομένα εκπαίδευσης και ως εκ τούτου το σφάλμα εκπαίδευσης $\overline{err}$ θα είναι μία υπερβολικά αισιόδοξη εκτίμηση του σφάλματος γενίκευσης Err_T. Μέρος της διαφοράς οφείλεται στο πού προκύπτουν τα σημεία αξιολόγησης. Η ποσότητα Err_T μπορεί να θεωρηθεί ως σφάλμα εκτός δείγματος καθώς τα διανύσματα των χαρακτηριστικών ελέγχου δε χρειάζεται να συμπίπτουν με τα διανύσματα των χαρακτηριστικών εκπαίδευσης. Τη φύση της βελτίωσης στο $\overline{err}$ είναι πιο εύκολο να την αντιληφθούμε όταν επικεντρωνόμαστε στο σφάλμα εντός δείγματος:

$$Err_{in} = \frac{1}{N}\sum_{i=1}^{N} E_{Y^0}\left[L\left(Y_i^0, \hat{f}(x_i)\right)|T\right] \qquad (4.4.2)$$

Ο συμβολισμός Y^0 υποδεικνύει ότι παρατηρούμε N νέες τιμές απόκρισης σε κάθε ένα από τα σημεία εκπαίδευσης x_i $i = 1, \dots, N$.Ορίζουμε τη βελτίωση ως τη διαφορά μεταξύ του Err_{in} και του σφάλματος εκπαίδευσης $\overline{err}$

$$op \equiv Err_{in} - \overline{err} \qquad (4.4.3)$$

Η τιμή αυτή είναι συνήθως θετική, καθώς το $\overline{err}$ είναι μεροληπτικό προς τα κάτω, ως μία εκτίμηση του σφάλματος πρόβλεψης.

Για το τετραγωνικό σφάλμα 0-1 και άλλες συναρτήσεις απώλειας γενικώς μπορεί να δειχθεί ότι :

$$\omega = \frac{2}{N}\sum_{i=1}^{N} Cov(\hat{y}_i, y_i), \qquad (4.4.4)$$

όπου το Cov υποδεικνύει τη συνδιασπορά. Έτσι το ποσό κατά το οποίο το $\overline{err}$ υποτιμά το πραγματικό σφάλμα, εξαρτάται από το πόσο ισχυρά οι y_i επιδρούν στην πρόβλεψή τους. Όσο πιο ισχυρά προσαρμόζονται τα δεδομένα, τόσο μεγαλύτερη γίνεται η $Cov(\hat{y}_i, y_i)$, επομένως αυξάνεται η βελτίωση. Έχουμε την εξής σημαντική σχέση:

$$E_y(Err_{in}) = E_y(\overline{err}\,) + \frac{2}{N}\sum_{i=1}^{N} Cov(\hat{y}_i, y_i) \qquad (4.4.5)$$

Αυτή η έκφραση απλοποιείται αν τα $\hat{y}_i$, προέρχονται από γραμμική προσαρμογή με d εισόδους ή βασικές συναρτήσεις. Για παράδειγμα :

$$\sum_{i=1}^{N} Cov(\hat{y}_i, y_i) = d\sigma_\varepsilon^2 \qquad (4.4.6)$$

Για το προσθετικό μοντέλο σφάλματος, $Y = f(x) + \varepsilon$ και έτσι έχουμε ότι

$$E_y(Err_{in}) = E_y(\overline{err}\,) + 2\frac{d}{N}\sigma_\varepsilon^2 \qquad (4.4.7)$$

Η ΄΄αισιοδοξία΄΄ αυξάνει γραμμικά με τον αριθμό d των εισόδων ή των βασικών συναρτήσεων που χρησιμοποιούνται, αλλά μειώνεται καθώς το μέγεθος δείγματος εκπαίδευσης αυξάνεται. Ένας προφανής τρόπος για να εκτιμήσουμε το σφάλμα πρόβλεψης είναι να εκτιμήσουμε τη βελτίωση και έπειτα να την προσθέσουμε στο σφάλμα εκπαίδευσης $\overline{err}$.

Το σφάλμα εντός του δείγματος δεν είναι συνήθως άμεσου ενδιαφέροντος καθώς οι μελλοντικές τιμές των χαρακτηριστικών δεν είναι πιθανό να ταυτιστούν με τις τιμές του συνόλου εκπαίδευσης. Αλλά για σύγκριση μεταξύ των μοντέλων, το σφάλμα εντός του δείγματος είναι βολικό να οδηγεί σε αποδοτικά μοντέλα επιλογής. Ο λόγος είναι ότι το σχετικό (αντί του απολύτου) μέγεθος είναι αυτό που έχει σημασία.

4.5 ΕΚΤΙΜΗΣΕΙΣ ΤΟΥ ΕΝΤΟΣ- ΔΕΙΓΜΑΤΟΣ ΣΦΑΛΜΑΤΟΣ ΠΡΟΒΛΕΨΗΣ ΚΑΙ ΚΡΙΤΗΡΙΟ ΠΛΗΡΟΦΟΡΙΑΣ AKAIKE

Η γενική μορφή των εκτιμήσεων εντός- δείγματος είναι:

$$Err_{in} = \overline{err} + \hat{\omega} \qquad (4.5.1)$$

Το κριτήριο πληροφορίας Akaike είναι μία γενικώς εφαρμόσιμη εκτίμηση του Err_{in} , όταν χρησιμοποιείται η συνάρτηση απώλειας της λογαριθμικής πιθανοφάνειας (ή λογαριθμοπιθανοφάνειας). Στηρίζεται σε μια σχέση παρόμοια με την (4.4.7) , που ισχύει ασυμπτωτικά καθώς το $N \to \infty$:

$$-2E\left[\log Pr_{\hat{\theta}}(Y)\right] \approx -\frac{2}{N}E[\log lik] + 2\frac{d}{N} \qquad (4.5.2)$$

Εδώ η $Pr_{\hat{\theta}}(Y)$ είναι μια οικογένεια πυκνοτήτων για την Y (που περιλαμβάνει την πραγματική πυκνότητα), $\hat{\theta}$ είναι η εκτίμηση μεγίστης πιθανοφάνειας του θ και $\log lik$ είναι η μεγιστοποιημένη λογαριθμική πιθανοφάνεια:

$$\log lik = \sum_{i=1}^{N} \log Pr_{\hat{\theta}}(y_i) \qquad (4.5.3)$$

Για να χρησιμοποιήσουμε το AIC για επιλογή μοντέλου, απλά επιλέγουμε το μοντέλο που δίνει το μικρότερο AIC από το σύνολο των μοντέλων που θεωρούνται. Για μη γραμμικά και άλλα σύνθετα μοντέλα πρέπει να αντικαταστήσουμε το d με κάποιο μέτρο του μοντέλου πολυπλοκότητας.

Δοθέντος ενός συνόλου μοντέλων $\xi_\alpha(x)$ που κατατάσσονται από μία παράμετρο ρύθμισης α, δηλώνεται από το $\overline{err}(\alpha)$ και έστω d(α) το σφάλμα εκπαίδευσης και ο αριθμός των παραμέτρων για κάθε μοντέλο. Τότε για αυτό το σύνολο μοντέλων ορίζουμε:

$$AIC(a) = \overline{err}(a) + 2\frac{d(a)}{N}\hat{\sigma}_\varepsilon^2 \qquad (4.5.4)$$

Η συνάρτηση $AIC(a)$ παρέχει μια εκτίμηση της καμπύλης του σφάλματος δοκιμής και βρίσκουμε την παράμετρο $\hat{\alpha}$ που την ελαχιστοποιεί. Το τελικά επιλεγμένο μοντέλο μας είναι το $\xi_{\hat{\alpha}}(x)$. Να σημειωθεί ότι αν οι βασικές συναρτήσεις επιλέγονται προσαρμοστικά η (4.4.6) δεν ισχύει πλέον. Για παράδειγμα αν έχουμε ένα συνολικό αριθμό p εισόδων και επιλέγουμε το καλύτερα προσαρμοσμένο μοντέλο με d<p εισόδους, η βελτίωση θα υπερβεί το $(2d/N)\hat{\sigma}_\varepsilon^2$ και έτσι ο αποδοτικός αριθμός παραμέτρων προσαρμογής είναι μεγαλύτερος του d.

Χρησιμοποιώντας το AIC επιλέγουμε τον αριθμό των βασικών συναρτήσεων περίπου να ελαχιστοποιεί το Err(M) και για την εντροπία και για την 0-1 απώλεια. Ο απλός τύπος δίνεται από την ακόλουθη σχέση:

$$(2/N)\sum_{i=1}^{N} Cov(\hat{y}_i, y_i) = (2d/N)\sigma_\varepsilon^2$$

Ισχύει ακριβώς για γραμμικά μοντέλα με προσθετικά σφάλματα και απώλεια τετραγωνικού σφάλματος και προσεγγιστικά για γραμμικά μοντέλα και λογαριθμικές πιθανοφάνειες. Στην πραγματικότητα, ο τύπος δεν ισχύει γενικά για την 0-1 απώλεια.

4.6 ΑΠΟΔΟΤΙΚΟΣ ΑΡΙΘΜΟΣ ΠΑΡΑΜΕΤΡΩΝ

Η έννοια του΄΄ αριθμού των παραμέτρων΄΄ μπορεί να γενικευτεί, ειδικά σε μοντέλα στα οποία η συστηματικοποίηση χρησιμοποιείται στην προσαρμογή. Ας υποθέσουμε ότι τοποθετούμε τα αποτελέσματα $y_1, \dots, y_N$ σε ένα διάνυσμα y και ομοίως για τις προβλέψεις $\hat{y}$. Στη συνέχεια μια γραμμικά προσαρμόσιμη μέθοδος είναι εκείνη για την οποία μπορούμε να γράψουμε:

$$\hat{y} = Sy \qquad (4.6.1)$$

όπου S είναι ένας $N \times N$ πίνακας που εξαρτάται από το διάνυσμα εισόδου x_i αλλά όχι από το y_i. Οι γραμμικά προσαρμόσιμες μέθοδοι περιλαμβάνουν τη γραμμική παλινδρόμηση στα πραγματικά χαρακτηριστικά ή σε ένα προερχόμενο από τη βάση σύνολο και μεθόδους εξομάλυνσης που χρησιμοποιούν τετραγωνική συρρίκνωση. Έπειτα ο αποδοτικός αριθμός παραμέτρων ορίζεται ως:

$$df(S) = trace(S) \tag{4.6.2}$$

Το άθροισμα των διαγώνιων στοιχείων του S είναι επίσης γνωστό και ως αποδοτικός αριθμός βαθμών ελευθερίας.

4.7 ΜΠΕΫΖΙΑΝΗ ΠΡΟΣΕΓΓΙΣΗ ΚΑΙ ΚΡΙΤΗΡΙΟ ΠΛΗΡΟΦΟΡΙΑΣ BIC

Το Μπεϋζιανό κριτήριο πληροφορίας (BIC), όπως το AIC, είναι εφαρμόσιμο σε ρυθμίσεις όπου η προσαρμογή διεξάγεται με μεγιστοποίηση της λογαριθμικής πιθανοφάνειας. Η γενική μορφή του BIC είναι:

$$BIC = -2\log lik + (\log N)d \tag{4.7.1}$$

Η στατιστική BIC ($1/2$ $\varphi o\rho\acute{\varepsilon}\varsigma$) είναι επίσης γνωστή ως το κριτήριο Schwarz.
Σύμφωνα με το Γκαουσιανό μοντέλο, υποθέτοντας ότι η διακύμανση σ_ε^2 είναι γνωστή, η ποσότητα $-2\log lik + (\log N)d$ ισούται (μέχρι μία σταθερά) με την $\sum_i \left(y_i - \hat{f}(x_i)\right)^2 / \sigma_\varepsilon^2$, η οποία είναι $N\overline{err}/\sigma_\varepsilon^2$ για την απώλεια του τετραγωνικού σφάλματος. Ως εκ τούτου μπορούμε να γράψουμε:

$$BIC = \frac{N}{\sigma_\varepsilon^2}\left[\overline{err} + (\log N)\frac{d}{N}\sigma_\varepsilon^2\right] \tag{4.7.2}$$

Επομένως το BIC είναι αναλογικά προς το AIC, με τον παράγοντα 2 να αντικαθίσταται από το $\log N$. Υποθέτοντας ότι N>$e^2 \approx 7{,}4$, το BIC τείνει να ποινικοποιεί τα σύνθετα μοντέλα πιο έντονα, δίνοντας προτεραιότητα σε απλούστερα μοντέλα, στην επιλογή. Όπως και με το AIC, η σ_ε^2 τυπικά εκτιμάται από το μέσο τετραγωνικό σφάλμα ενός χαμηλά μεροληπτικού μοντέλου. Να σημειωθεί ωστόσο ότι το μέτρο σφάλματος της εσφαλμένης ταξινόμησης δε τίθεται στα πλαίσια του BIC, καθώς δεν ανταποκρίνεται στη λογαριθμική πιθανοφάνεια των δεδομένων κάτω από οποιοδήποτε μοντέλο πιθανότητας. Παρά την ομοιότητα με το AIC, το BIC κινητοποιείται με ένα εντελώς διαφορετικό τρόπο. Προέρχεται από τη Μπεϋζιανή προσέγγιση στην επιλογή μοντέλου, η οποία περιγράφεται παρακάτω.

Ας υποθέσουμε ότι έχουμε ένα σύνολο υποψηφίων μοντέλων M_m ,$m = 1, \dots, M$ και αντίστοιχες παραμέτρους του μοντέλου θ_m. Ας υποθέσουμε επίσης ότι έχουμε μια εκ των προτέρων κατανομή $Pr(\theta_m|M_m)$ για τις παραμέτρους κάθε μοντέλου M_m, η εκ των υστέρων πιθανότητα του δοθέντος μοντέλου είναι:

$$\begin{aligned} Pr(M_m|Z) &\propto Pr(M_m)Pr(Z|M_m) \\ &\propto Pr(M_m)\int Pr\left(Z\middle|\theta_\mu, M_m\right)Pr(\theta_m|M_m)d\theta_m \end{aligned}$$

(4.7.3)

Όπου το Z αναπαριστά τα δεδομένα εκπαίδευσης $\{x_i, y_i\}_1^N$. Για να συγκρίνουμε τα δύο μοντέλα M_m $\kappa\alpha\iota$ M_l, σχηματίζουμε τα εκ των υστέρων odds:

$$\frac{Pr(M_m|Z)}{Pr(M_l|Z)} = \frac{Pr(M_m)}{Pr(M_l)}\frac{Pr(Z|M_m)}{Pr(Z|M_l)} \quad (4.7.4)$$

Εάν τα odds είναι μεγαλύτερα του ένα, επιλέγουμε το μοντέλο m, αλλιώς επιλέγουμε το l. Η ποσότητα που βρίσκεται στα δεξιά

$$BF(Z) = \frac{Pr(Z|M_m)}{Pr(Z|M_l)} \quad (4.7.5)$$

ονομάζεται συντελεστής Bayes, η συμβολή των δεδομένων προς τα εκ των υστέρων odds.

Χρειαζόμαστε κάποιον τρόπο προσέγγισης της $Pr(Z|M_m)$. Η λεγόμενη προσέγγιση Laplace στο ολοκλήρωμα, ακολουθούμενη από κάποιες άλλες απλοποιήσεις δίνει:

$$\log Pr(Z|M_m) = Pr\left(Z\middle|\hat{\theta}_m, M_m\right) - \frac{d_m}{2}\log N + O(1) \quad (4.7.6)$$

Εδώ το $\hat{\theta}_m$ είναι ένας εκτιμητής μεγίστης πιθανοφάνειας και d_m είναι ο αριθμός των ελευθέρων παραμέτρων στο μοντέλο M_m. Αν ορίσουμε τη συνάρτηση απώλειας να είναι:

$$-2\log Pr\left(\hat{\theta}_m, M_m\right)$$

Αυτό είναι ισοδύναμο με το κριτήριο BIC της εξίσωσης (4.7.1). Ως εκ τούτου, επιλέγουμε το μοντέλο με το ελάχιστο BIC, αυτό είναι ισοδύναμο με το να επιλέξουμε το μοντέλο με την μεγαλύτερη (κατά προσέγγιση) εκ των υστέρων πιθανότητα. Αν υπολογίσουμε το κριτήριο BIC για ένα σύνολο M μοντέλων, που δίνει $BIC_m, m = 1, \dots, M$ τότε μπορούμε να εκτιμήσουμε την εκ των υστέρων πιθανότητα κάθε μοντέλου M_m ως:

$$\frac{e^{-\frac{1}{2}BIC_m}}{\sum_{l=1}^{M} e^{-\frac{1}{2}BIC_m}} \quad (4.7.7)$$

Έτσι μπορούμε να εκτιμήσουμε όχι μόνο το καλύτερο μοντέλο αλλά επίσης να αξιολογήσουμε τα σχετικά πλεονεκτήματα των μοντέλων που θεωρούνται. Για τους σκοπούς της επιλογής μοντέλων δεν υπάρχει σαφής επιλογή μεταξύ των AIC και BIC. Το BIC είναι ασυμπτωτικά συνεπές για την επιλογή μοντέλου. Αυτό σημαίνει ότι δεδομένης μιας οικογένειας μοντέλων, που περιλαμβάνει το πραγματικό μοντέλο, η πιθανότητα ότι το BIC θα επιλέξει το σωστό μοντέλο τείνει στο ένα καθώς το μέγεθος δείγματος $N \to \infty$. Από την άλλη πλευρά, για πεπερασμένα δείγματα, το BIC συχνά επιλέγει μοντέλα τα οποία είναι υπερβολικά απλά λόγω της βαριάς ποινής για την πολυπλοκότητα.

4.8 ΔΙΑΣΤΑΥΡΩΜΕΝΗ ΕΠΙΚΥΡΩΣΗ

4.8.1 ΟΡΙΣΜΟΣ

Η διασταυρωμένη επικύρωση (cross- validation) είναι μία στατιστική μέθοδος αξιολόγησης και σύγκρισης αλγορίθμων μάθησης με διαίρεση των δεδομένων σε δύο τμήματα: το ένα χρησιμοποιείται για τη μάθηση και το άλλο για την επικύρωση του μοντέλου. Στην τυπική διασταυρωμένη επικύρωση, τα σύνολα εκπαίδευσης και επικύρωσης πρέπει να διασταυρωθούν σε διαδοχικούς γύρους έτσι ώστε κάθε δεδομένο να έχει μία ευκαιρία να επικυρωθεί. Η βασική μορφή διασταυρωμένης επικύρωσης είναι η k-

πλάσια διασταυρωμένη επικύρωση (k-fold cross- validation). Άλλες μορφές της διασταυρωμένης επικύρωσης είναι ειδικές περιπτώσεις της k-πλάσιας διασταυρωμένης επικύρωσης ή περιλαμβάνουν επαναλαμβανόμενους γύρους της k-πλάσιας διασταυρωμένης επικύρωσης .

Στην k-πλάσια διασταυρωμένη επικύρωση τα δεδομένα πρώτα κατανέμονται σε k τμήματα ή πτυχές ίσου ή σχεδόν ίσου μεγέθους. Στη συνέχεια k επαναλήψεις εκπαίδευσης και επικύρωσης εκτελούνται έτσι ώστε σε κάθε επανάληψη μια διαφορετική πτυχή των δεδομένων να επαρκεί για επικύρωση ενώ οι εναπομείνασες k-1 να χρησιμοποιούνται για μάθηση.

Η διασταυρωμένη επικύρωση χρησιμοποιείται για την αξιολόγηση ή τη σύγκριση των αλγορίθμων εκμάθησης ως εξής: σε κάθε επανάληψη ένας ή και περισσότεροι αλγόριθμοι εκμάθησης χρησιμοποιούν k-1 πτυχές των δεδομένων για να μάθουν ένα ή και περισσότερα μοντέλα και στη συνέχεια τα εξακριβωμένα μοντέλα καλούνται να κάνουν προβλέψεις σχετικά με τα δεδομένα στην πτυχή επικύρωσης. Η απόδοση κάθε αλγορίθμου μάθησης για κάθε πτυχή μπορεί να παρακολουθηθεί με τη χρήση κάποιων προκαθορισμένων μετρικών ακρίβειας. Μέχρι την ολοκλήρωση, k δείγματα της μετρικής απόδοσης θα είναι διαθέσιμα για κάθε αλγόριθμο. Διαφορετικές μεθοδολογίες, όπως αυτές της μέσης τιμής μπορούν να χρησιμοποιηθούν για να ληφθεί ένα συνολικό μέτρο από αυτά τα δείγματα, ή αυτά τα δείγματα μπορούν να χρησιμοποιηθούν σε έναν έλεγχο στατιστικής υπόθεσης για να δειχθεί ότι ένας αλγόριθμος είναι ανώτερος ενός άλλου.

Υπάρχουν δύο πιθανοί στόχοι στη διασταυρωμένη επικύρωση:

(i) Η εκτίμηση της απόδοσης των μοντέλων μάθησης από τα διαθέσιμα δεδομένα με τη χρήση ενός αλγορίθμου. Με άλλα λόγια η μέτρηση της (δυνατότητας) γενίκευσης ενός αλγορίθμου.

(ii) Η σύγκριση της επίδοσης ενός ή και περισσότερων διαφορετικών αλγορίθμων και η εύρεση του καλύτερου αλγορίθμου για τα διαθέσιμα δεδομένα, ή εναλλακτικά η σύγκριση της απόδοσης δύο ή και περισσότερων παραλλαγών ενός παραμετρικοποιημένου μοντέλου.

Οι δύο παραπάνω στόχοι είναι ιδιαιτέρως συσχετισμένοι, καθώς ο δεύτερος στόχος επιτυγχάνεται αυτόματα, εάν κάποιος γνωρίζει τις ακριβείς εκτιμήσεις απόδοσης. Δεδομένων ενός δείγματος N περιπτώσεων και ενός αλγορίθμου εκμάθησης A, η μέση ακρίβεια διασταυρωμένης επικύρωσης του A, σε αυτές τις N περιπτώσεις, μπορεί να ληφθεί σαν μια εκτίμηση για την ακρίβεια του A σε αφανή δεδομένα, όταν ο A εκπαιδεύεται σε όλες τις N περιπτώσεις. Εναλλακτικά, εάν ο τελικός στόχος είναι η σύγκριση δύο αλγορίθμων μάθησης, τα δείγματα απόδοσης που λαμβάνονται μέσω της διασταυρωμένης επικύρωσης μπορούν να χρησιμοποιηθούν για τη διεξαγωγή ελέγχων στατιστικών υποθέσεων, συγκρίνοντας ένα ζεύγος αλγορίθμων μάθησης.

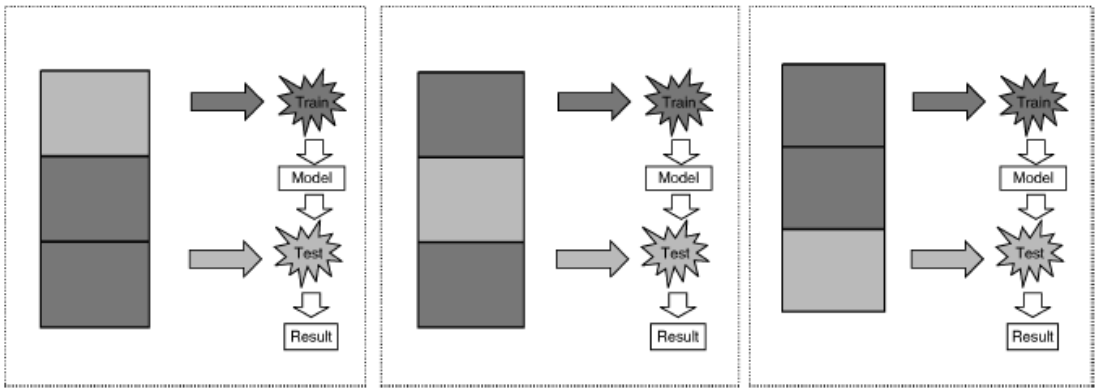
4.8.2 HOLD-OUT ΕΠΙΚΥΡΩΣΗ

Για να αποφύγουμε την υπερπροσαρμογή, ένα ανεξάρτητο σύνολο δοκιμής προτιμάται. Μια φυσική προσέγγιση είναι να χωρίσουμε τα δεδομένα σε δύο μη επικαλυπτόμενα μέρη: το ένα για εκπαίδευση και το άλλο για έλεγχο. Τα δεδομένα ελέγχου παρακρατούνται και δεν εξετάζονται κατά τη διάρκεια της εκπαίδευσης. Η hold-out επικύρωση αποφεύγει την υπερκάλυψη μεταξύ των δεδομένων εκπαίδευσης και ελέγχου, αποδίδοντας μια πιο ακριβή εκτίμηση για την απόδοση γενίκευσης του αλγορίθμου. Το μειονέκτημα είναι ότι αυτή η διαδικασία δε χρησιμοποιεί όλα τα διαθέσιμα δεδομένα και τα αποτελέσματα εξαρτώνται σε μεγάλο βαθμό από την επιλογή για το διαχωρισμό εκπαίδευσης-ελέγχου. Οι περιπτώσεις που επιλέγονται για ένταξη στο σύνολο δοκιμής μπορεί να είναι υπερβολικά εύκολο ή υπερβολικά δύσκολο να ταξινομηθούν και αυτό μπορεί να παραποιήσει τα

αποτελέσματα. Επιπλέον, τα δεδομένα στο σύνολο δοκιμής μπορεί να είναι πολύτιμα για εκπαίδευση και αν παρακρατείται απόδοση πρόβλεψης μπορεί να οδηγήσει σε παραποιημένα αποτελέσματα. Αυτά τα προβλήματα μπορούν μερικώς να αντιμετωπιστούν με την επανάληψη της hold-out επικύρωσης πολλαπλές φορές και υπολογίζοντας το μέσο των αποτελεσμάτων ενώ άλλα δε συμπεριλαμβάνονται καθόλου ή αντιστρόφως κάποια δεδομένα μπορούν πάντα να πέσουν στο σύνολο ελέγχου και ποτέ δεν έχουν την ευκαιρία να συμβάλλουν στη φάση εκμάθησης. Για την αντιμετώπιση αυτών των προκλήσεων και τη χρήση των διαθέσιμων δεδομένων στο μέγιστο, η k-πλάσια διασταυρωμένη επικύρωση χρησιμοποιείται.

4.8.3 Κ-ΠΛΑΣΙΑ ΔΙΑΣΤΑΥΡΩΜΕΝΗ ΕΠΙΚΥΡΩΣΗ

Στην k-πλάσια διασταυρωμένη επικύρωση (k-fold cross- validation) τα δεδομένα πρώτα κατανέμονται σε k πτυχές ίσου (ή σχεδόν ίσου) μεγέθους. Στη συνέχεια διεξάγονται k επαναλήψεις εκπαίδευσης και επικύρωσης, έτσι ώστε σε κάθε επανάληψη μια διαφορετική πτυχή των δεδομένων να επαρκεί για επικύρωση ενώ οι εναπομείνασες k-1 πτυχές χρησιμοποιούνται για μάθηση. Τα δεδομένα συνήθως στρωματοποιούνται πριν χωριστούν σε k πτυχές. Η διαστρωμάτωση είναι η διαδικασία της ανακατανομής των δεδομένων έτσι ώστε να εξασφαλιστεί ότι κάθε πτυχή είναι ένας καλός αντιπρόσωπος του συνόλου. Για παράδειγμα σε ένα δυαδικό πρόβλημα ταξινόμησης, όπου κάθε κλάση αποτελείται από το 50% των δεδομένων είναι καλύτερο να κατανεμηθούν τα δεδομένα έτσι ώστε, σε κάθε πτυχή, κάθε κλάση να αποτελείται από περίπου τις μισές περιπτώσεις.



ΣΧΗΜΑ 4.8.1 : Διαδικασία της τρπλάσιας διασταυρωμένης επικύρωσης

4.8.4 10-ΠΛΑΣΙΑ ΔΙΑΣΤΑΥΡΩΜΕΝΗ ΕΠΙΚΥΡΩΣΗ

Ένας ιδανικός ή στατιστικά άριστος στατιστικός σχεδιασμός πρέπει να παρέχει έναν επαρκή αριθμό ανεξάρτητων μετρήσεων της απόδοσης του αλγορίθμου. Για να γίνουν ανεξάρτητες μετρήσεις της απόδοσης του αλγορίθμου πρέπει να διασφαλιστεί ότι οι παράγοντες που επηρεάζουν τη μέτρηση είναι ανεξάρτητοι από το ένα τρέξιμο στο επόμενο. Αυτοί οι παράγοντες είναι: (i) τα δεδομένα εκπαίδευσης από τα οποία μαθαίνει ο αλγόριθμος, (ii) τα δεδομένα ελέγχου που χρησιμοποιούνται για τη μέτρηση της απόδοσης του αλγορίθμου. Εάν κάποια δεδομένα χρησιμοποιούνται για δοκιμή σε περισσότερους από ένα γύρους τα αποτελέσματα που λαμβάνονται θα είναι εξαρτημένα και μια στατιστική σύγκριση δε θα είναι έγκυρη.

Όχι μόνο τα σύνολα δεδομένων θα πρέπει να ελέγχονται ανεξάρτητα στα διάφορα τρεξίματα, αλλά δε θα πρέπει να υπάρχει καμιά επικάλυψη μεταξύ των δεδομένων που χρησιμοποιούνται για μάθηση και των δεδομένων που χρησιμοποιούνται για επικύρωση στο ίδιο τρέξιμο. Τυπικά ένας αλγόριθμος μάθησης μπορεί να κάνει πιο ακριβείς προβλέψεις στα δεδομένα που έχουν ιδωθεί κατά τη διάρκεια της φάσης μάθησης από ότι σε αυτά που δεν έχουν. Για αυτό το λόγο, μια επικάλυψη μεταξύ του συνόλου εκπαίδευσης και επικύρωσης μπορεί να οδηγήσει σε μια υπερεκτίμηση της μετρικής απόδοσης και απαγορεύεται. Για να ικανοποιηθεί η άλλη απαίτηση, συγκεκριμένα ένα επαρκώς μεγάλο δείγμα, οι περισσότεροι στατιστικοί κάνουν έκκληση για 30+ δείγματα.

Για ένα πραγματικά καλό στατιστικό σχεδιασμό, τα διαθέσιμα δεδομένα θα πρέπει να διαιρεθούν σε $30 \times 2 = 60$ διχοτομήσεις για να εκτελεστούν 30 πραγματικά ανεξάρτητα τρεξίματα εκπαίδευσης-δοκιμής. Ωστόσο, αυτό δεν είναι πρακτικό, διότι η απόδοση των αλγορίθμων μάθησης και κατάταξή τους δεν είναι δεν είναι αναλλοίωτες σε σχέση με τον αριθμό των δειγμάτων που είναι διαθέσιμα για μάθηση. Με άλλα λόγια, μια εκτίμηση ακριβείας σε μια τέτοια περίπτωση, θα μπορούσε να αντιστοιχεί στην ακρίβεια του αλγορίθμου μάθησης όταν μαθαίνει από μόλις το $1/60$ των διαθέσιμων δεδομένων με την προϋπόθεση ότι τα σύνολα εκπαίδευσης και επικύρωσης είναι του ιδίου μεγέθους. Εντούτοις, η ακρίβεια του αλγορίθμου μάθησης στα αφανή δεδομένα, όταν ο αλγόριθμος εκπαιδεύεται σε όλα τα επί του παρόντος διαθέσιμα δεδομένα, είναι πολύ μεγαλύτερη, δεδομένου ότι οι αλγόριθμοι μάθησης εν γένει βελτιώνονται όσον αφορά την ακρίβεια καθώς περισσότερα δεδομένα καθίστανται διαθέσιμα για μάθηση. Ομοίως συγκρίνοντας δύο αλγορίθμους Α και Β, ακόμα και αν ο Α έχει αποδειχτεί ότι είναι ο καλύτερος αλγόριθμος όταν χρησιμοποιείται το $1/60$ των διαθέσιμων δεδομένων δεν υπάρχει εγγύηση ότι θα είναι επίσης ο καλύτερος αλγόριθμος σε σχέση με όταν χρησιμοποιούνται όλα τα διαθέσιμα δεδομένα για μάθηση. Πολλοί υψηλής απόδοσης αλγόριθμοι μάθησης χρησιμοποιούν σύνθετα μοντέλα με πολλές παραμέτρους και απλά δεν αποδίδουν καλά με ένα πολύ μικρό πλήθος δεδομένων. Αλλά μπορεί να είναι εξαιρετικοί όταν επαρκή δεδομένα είναι διαθέσιμα για να μάθουν από αυτά.

Υπενθυμίζεται ότι δύο παράγοντες επηρεάζουν το μέτρο απόδοσης: το σύνολο εκπαίδευσης και το σύνολο ελέγχου. Το σύνολο εκπαίδευσης επηρεάζει τη μέτρηση έμμεσα μέσω του αλγορίθμου μάθησης, ενώ η σύνθεση του συνόλου δοκιμής έχει άμεσο αντίκτυπο στο μέτρο απόδοσης . Ένας λογικός πειραματικός συμβιβασμός μπορεί να είναι να επιτρέπονται τα επικαλυπτόμενα σύνολα εκπαίδευσης διατηρώντας τα σύνολα δοκιμής ανεξάρτητα. Η k-πλάσια διασταυρωμένη επικύρωση κάνει ακριβώς αυτό.

Τώρα το ζητούμενο γίνεται η επιλογή μιας κατάλληλης τιμής για το k. Ένα μεγάλο k είναι φαινομενικά επιθυμητό καθώς με ένα μεγάλο k (i) υπάρχουν περιπτώσεις εκτίμησης απόδοσης και (ii) το μέγεθος του συνόλου εκπαίδευσης είναι πιο κοντά στο μέγεθος του πλήρους συνόλου των δεδομένων. Έτσι αυξάνοντας την πιθανότητα ότι οποιοδήποτε συμπέρασμα που βγαίνει σχετικά με τους αλγορίθμους μάθησης στα πλαίσια της δοκιμής, θα γενικεύεται στην περίπτωση που όλα τα δεδομένα χρησιμοποιούνται για την εκπαίδευση του μοντέλου μάθησης. Καθώς το k αυξάνεται, ωστόσο, η επικάλυψη μεταξύ των συνόλων εκπαίδευσης επίσης αυξάνεται. Για παράδειγμα, με 5-πλάσια διασταυρωμένη επικύρωση, κάθε σύνολο εκπαίδευσης μοιράζεται μόνο τα $3/4$ των περιπτώσεών του με κάθε ένα από τα άλλα 4 σύνολα εκπαίδευσης, ενώ με τη 10-πλάσια διασταυρωμένη επικύρωση, κάθε σύνολο εκπαίδευσης μοιράζεται τα $8/9$ των περιπτώσεών του με κάθε ένα από τα υπόλοιπα 9 σύνολα εκπαίδευσης. Επιπλέον, η αύξηση του k συρρικνώνει το μέγεθος του συνόλου δοκιμής, οδηγώντας σε λιγότερο ακριβείς και λιγότερο λεπτομερείς μετρήσεις της μετρικής απόδοσης. Για παράδειγμα, με ένα σύνολο δοκιμής μεγέθους 10 περιπτώσεων, μπορεί κανείς να μετρήσει την ακρίβεια με προσέγγιση 10%, ενώ με 20 περιπτώσεις η ακρίβεια μπορεί να μετρηθεί με προσέγγιση 5%. Αυτοί οι ανταγωνιστικοί παράγοντες ελήφθησαν όλοι υπόψη και η γενική συναίνεση φαίνεται να είναι ότι ο k=10

είναι ένας καλός συμβιβασμός. Αυτή η τιμή του k είναι ιδιαιτέρως ελκυστική διότι κάνει προβλέψεις με τη χρήση του 90% των δεδομένων, καθιστώντας πιο πιθανή τη γενίκευση στο σύνολο των δεδομένων.

4.8.5 LEAVE-ONE-OUT ΔΙΑΣΤΑΥΡΩΜΕΝΗ ΕΠΙΚΥΡΩΣΗ (LOOCV)

Η leave-one-out διασταυρωμένη επικύρωση είναι μια ειδική περίπτωση της k-πλάσιας διασταυρωμένης επικύρωσης, όπου το k είναι ίσο με τον αριθμό των περιπτώσεων στα δεδομένα. Με άλλα λόγια, σε κάθε επανάληψη σχεδόν όλα τα δεδομένα εκτός μόνο μίας παρατήρησης χρησιμοποιούνται για εκπαίδευση και το μοντέλο ελέγχεται σε αυτή τη μεμονωμένη παρατήρηση. Μια εκτίμηση της ακρίβειας που λαμβάνεται με τη χρήση της LOOCV είναι γνωστό ότι είναι σχεδόν αμερόληπτη, αλλά έχει υψηλή διακύμανση, που οδηγεί σε αναξιόπιστες εκτιμήσεις. Εξακολουθεί να χρησιμοποιείται ευρέως όταν τα διαθέσιμα δεδομένα είναι πολύ σπάνια, ειδικά στη βιοπληροφορική, όπου μόνο δεκάδες δειγμάτων δεδομένων είναι διαθέσιμα.

4.9 ΔΙΑΣΤΑΥΡΩΜΕΝΑ ΕΠΙΚΥΡΩΜΕΝΟ ΠΟΣΟΣΤΟ ΣΦΑΛΜΑΤΟΣ (CROSS VALIDATED ERROR RATE)

4.9.1 ΕΙΣΑΓΩΓΗ

Στη μέθοδο της διασταυρωμένης επικύρωσης (βλ. ενότητα 4.9) γίνεται διαίρεση των δεδομένων σε k περίπου ίσα μέρη:

1	2	3	4	5
Validation	Train	Train	Train	Train

ΣΧΗΜΑ 4.9.1

Για κάθε k=1,..,K, ταιριάζει το μοντέλο με την παράμετρο λ στα άλλα k-1 μέρη, δίνοντας $\hat{\beta}^{-k}(\lambda)$ και υπολογίζει το σφάλμα του στην πρόβλεψη του k-οστού μέρους:

$$E_k(\lambda) = \sum_{i \in k-\text{οστό μέρος}} \left(y_i - x_i \hat{\beta}^{-k}(\lambda)\right)^2 \qquad (4.9.1.1)$$

Αυτό δίνει το σφάλμα διασταυρωμένης επικύρωσης:

$$CV(\lambda) = \frac{1}{K} \sum_{k=1}^{K} E_k(\lambda) \qquad (4.9.1.2)$$

Αυτό επαναλαμβάνεται για πολλές τιμές της λ και επιλέγεται αυτή η τιμή για την οποία ελαχιστοποιείται το $CV(\lambda)$ (τυπικά χρησιμοποιούμε k=5 ή k=10).

4.9.2 ΑΞΙΟΛΟΓΩΝΤΑΣ ΤΗΝ ΑΠΟΔΟΣΗ ΤΗΣ ΕΠΙΛΟΓΗΣ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ

Η διασταυρωμένη επικύρωση συχνά χρησιμοποιείται για την αξιολόγηση της ακρίβειας των ταξινομητών. Θα πρέπει να σημειωθεί ότι τα αποτελέσματα της επιλογής χαρακτηριστικών μπορούν να ποικίλουν ακόμα και με μια διαφορά μιας και μόνο περίπτωσης στο σύνολο εκπαίδευσης όταν το μέγεθος δείγματος είναι μικρό. Σε κάποιες περιπτώσεις, τα στάδια της επιλογής χαρακτηριστικών είναι εκτός των διαδικασιών της διασταυρωμένης επικύρωσης, δηλαδή η επιλογή των χαρακτηριστικών γίνεται με όλα τα δείγματα και η διασταυρωμένη επικύρωση, γίνεται μόνο για τη διαδικασία της ταξινόμησης. Αυτό το είδος της διασταυρωμένης επικύρωσης καλείται CV1 και μπορεί να είναι σοβαρά μεροληπτικό ως προς την αξιολόγηση υπέρ των μελετημένων μεθόδων λόγω της διαρροής πληροφορίας στο στάδιο της επιλογής χαρακτηριστικών. Για παράδειγμα, σε μια πιλοτική μελέτη, δημιουργήθηκαν 100 δείγματα με 1000 χαρακτηριστικά για κάθε δείγμα, όλα προερχόμενα αμετάβλητα από τη Γκαουσιανή κατανομή N(0,1). Τυχαία αναθέτουμε τα δείγματα σε δύο ψευδοκλάσεις. Από τη στιγμή που το σύνολο δεδομένων είναι εντελώς μη πληροφοριακό, το αξιόπιστο σφάλμα διασταυρωμένης επικύρωσης θα πρέπει να είναι περίπου 50% ανεξαρτήτως της μεθόδου που χρησιμοποιείται. Αλλά από το σχέδιο CV1 θα μπορούσαμε να επιτύχουμε ένα διασταυρωμένα επικυρωμένο σφάλμα τόσο χαμηλό περίπου ίσο με 0,025, το οποίο δείχνει ότι η μεροληψία η οποία προκαλείται από το ακατάλληλο σχέδιο διασταυρωμένης επικύρωσης, μπορεί να είναι εκπληκτικά μεγάλη. Μια πιο κατάλληλη προσέγγιση είναι να συμπεριλάβουμε τη διαδικασία επιλογής χαρακτηριστικών στη διασταυρωμένη επικύρωση, δηλαδή να αφήσουμε τα δείγματα δοκιμής εκτός του συνόλου εκπαίδευσης πριν να γίνει οποιαδήποτε επιλογή χαρακτηριστικών. Με αυτό τον τρόπο, επικυρώνεται όχι μόνο ο αλγόριθμος ταξινόμησης, αλλά επίσης και η μέθοδος επιλογής χαρακτηριστικών. Αυτό το σχέδιο καλείται CV2.

4.9.3 Η ΣΧΕΤΙΚΗ ΣΗΜΑΣΙΑ ΤΩΝ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ ΣΤΟΥΣ SVM ΤΑΞΙΝΟΜΗΤΕΣ

Η συνάρτηση ενός γραμμικού SVM είναι η εξής;

$$\begin{aligned} g(x) &= sgnf(x) = sgn\{(wx) + b\} \\ &= sgn\{\textstyle\sum_{i=1}^{n} a_i\, \gamma_i (x_i x) + b\} \end{aligned} \qquad (4.9.3.1)$$

όπου i=1,…,n , $\gamma_i \in \{+1, -1\}$. Η αντίστοιχη ετικέτα κλάσης , $w = \sum_{i=1}^{n} a_i\, \gamma_i x_i$, είναι το διάνυσμα βαρών των χαρακτηριστικών και b ένας w βαθμωτός συμψηφισμός. Τα α_i και τα b εκτιμώνται από το σύνολο εκπαίδευσης. Μόνο τα δείγματα που βρίσκονται πιο κοντά στο διαχωριστικό όριο και καλούνται SVs, έχουν μη μηδενικά α_i και ως εκ τούτου η συνάρτηση f(x) είναι ένας γραμμικός συνδυασμός μόνο των SVs . Η αναλογία των SVs του συνόλου εκπαίδευσης αντανακλά ένα άνω φράγμα του αναμενόμενου σφάλματος γενίκευσης.

Στη συνέχεια βρίσκουμε ένα σύνολο χαρακτηριστικών που πετυχαίνει το μέγιστο διαχωρισμό μεταξύ των δύο κλάσεων των δειγμάτων. Λαμβάνοντας υπόψη και το πιθανό ασύμμετρο μέγεθος δείγματος, ορίζεται το ακόλουθο μέτρο :

$$S = \frac{1}{n_1}\sum_{x^+ \in\, \kappa\lambda\acute{\alpha}\sigma\eta\, 1} f(x^+) - \frac{1}{n_2}\sum_{x^- \in\, \kappa\lambda\acute{\alpha}\sigma\eta 2} f(x^-) \qquad (4.9.3.2)$$

όπου τα n_1 και n_2 είναι οι αριθμοί των δειγμάτων στις κλάσεις 1 και 2. Όσο μεγαλύτερο είναι το S, τόσο μεγαλύτερος είναι ο διαχωρισμός μεταξύ των δύο κλάσεων. Όσο μεγαλύτερο είναι το S, τόσο καλύτερα διαχωρισμένες είναι οι δύο κλάσεις θεωρώντας την

(4.9.3.1) και δηλώνοντας τους μέσους του χαρακτηριστικού j στις δύο κλάσεις ως m_j^+ και m_j^- , παίρνουμε:

$$S = \sum_{j=1}^{d} w_j m_j^+ - \sum_{j=1}^{d} w_j m_j^- = \sum_{j=1}^{d} w_j \left(m_j^+ - m_j^-\right) \qquad (4.9.3.3)$$

όπου d ο συνολικός αριθμός των χαρακτηριστικών και w_j το j-οστό στοιχείο του διανύσματος βαρών w. Το S είναι το άθροισμα των όρων που ορίζονται στα μεμονωμένα χαρακτηριστικά, έτσι ορίζουμε τη συμβολή του χαρακτηριστικού j στο S ως

$$s_j = w_j \left(m_j^+ - m_j^-\right) \qquad (4.9.3.4)$$

Το s_j ονομάζεται παράγοντας συμβολής του χαρακτηριστικού j. Ένας εναλλακτικός τρόπος για τη μέτρηση της σπουδαιότητας των χαρακτηριστικών είναι η χρήση του τετραγώνου του βάρους w_j, όπως προκύπτει στον SVM-RFE αλγόριθμο. Αυτό μπορεί να θεωρηθεί ως χρήση των σταθμισμένων αθροισμάτων των SVMs σε κάθε κλάση ως αντιπροσώπων, δηλαδή:

$$r^+ = \sum_{x_i^+ : SVs\ \sigma\tau\eta\nu\ \kappa\lambda\acute{\alpha}\sigma\eta\ 1} \alpha_\iota x_i^+ \quad , \quad r^- = \sum_{x_i^- : SVs\ \sigma\tau\eta\nu\ \kappa\lambda\acute{\alpha}\sigma\eta\ 2} \alpha_\iota x_i^- \qquad (4.9.3.5)$$

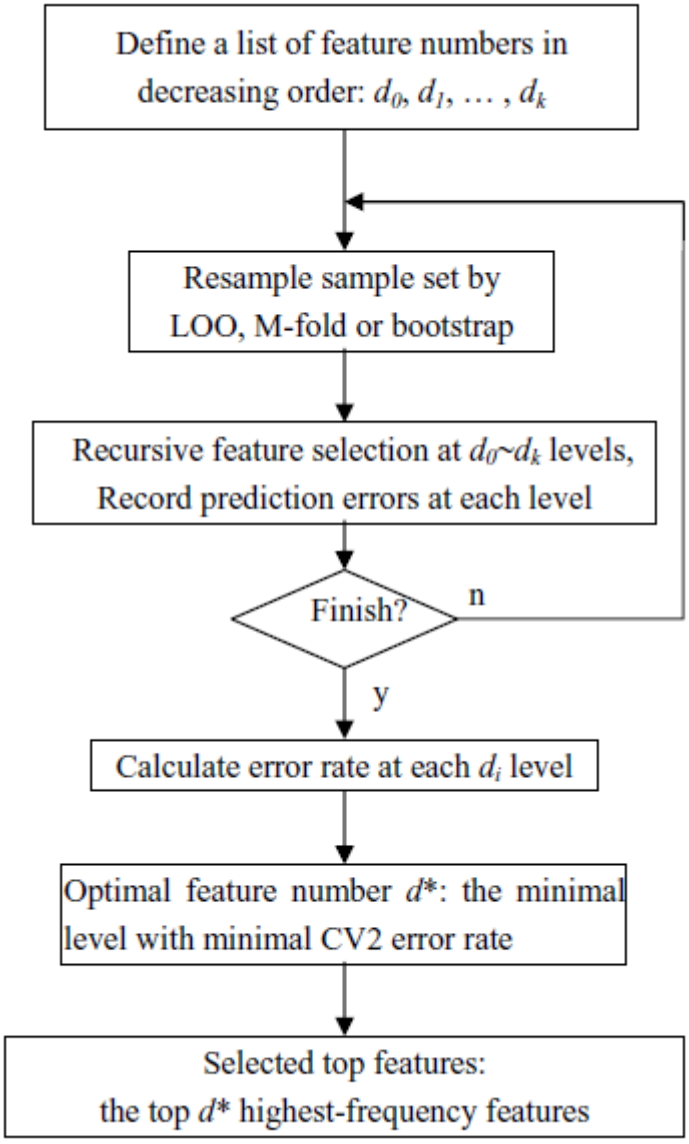
Ο διαχωρισμός μεταξύ των δύο αντιπροσώπων μπορεί να γραφεί ως:

$$\begin{aligned} S_{RFE} &= f(r^+) - f(r^-) \\ &= w(r^+ - r^-) \\ &= w \sum_{x_i : SVs\ \kappa\alpha\iota\ \sigma\tau\iota\varsigma\ \delta\acute{\upsilon}o\ \kappa\lambda\acute{\alpha}\sigma\varepsilon\iota\varsigma} \alpha_i\, \gamma_i x_i \\ &= ww \\ &= \sum_{j=1}^{d} w_j^2 \end{aligned} \qquad (4.9.3.6)$$

το οποίο μπορεί να αποσυντεθεί στο άθροισμα των εισφορών κάθε χαρακτηριστικού:

$$S_j^{RFE} = w_j^2 \qquad (4.9.3.7)$$

Παρατηρούμε ότι στην περίπτωση όπου οι κατανομές των δύο κλάσεων είναι και οι δύο υψηλής διάστασης Γκαουσιανές με πανομοιότυπους πίνακες συνδιακύμανσης Σ, το βέλτιστο w είναι η γραμμική διακρίνουσα συνάρτηση της μορφής $w = (m^+ - m^-)^T \Sigma^{-1}$. Η χρήση των μέσων των κλάσεων ως αντιπροσώπων και των δύο κλάσεων κάνει τη μέθοδο λιγότερο ευαίσθητη στο θόρυβο και στα πιθανά outliers.



ΣΧΗΜΑ 4.9.3 : Διάγραμμα ροής του αλγορίθμου R-SVM

4.9.4 ΑΝΑΔΡΟΜΙΚΗ ΤΑΞΙΝΟΜΗΣΗ ΚΑΙ ΕΠΙΛΟΓΗ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ

Για την επιλογή ενός υποσυνόλου χαρακτηριστικών, τα οποία συμβάλλουν περισσότερο στην ταξινόμηση, ταξινομούμε όλα τα χαρακτηριστικά σύμφωνα με τα s_j (4.9.3.4) και επιλέγουμε τα κορυφαία της λίστας. Χρησιμοποιούμε αυτή τη στρατηγική αναδρομικά στις ακόλουθες διαδικασίες:

ΒΗΜΑ 0 : Όρισε μια φθίνουσα σειρά αριθμών χαρακτηριστικών $d_0 > d_1 > d_2 > \cdots > d_k$ να εκλέγονται στη σειρά των βημάτων επιλογής. Θέσε $i = 0$ $\kappa\alpha\iota$ $d_0 = d$ (δηλαδή αρχίζουμε με όλα τα χαρακτηριστικά).

ΒΗΜΑ 1 : Στο βήμα i, κατασκεύασε τη συνάρτηση απόφασης SVM με τα τρέχοντα d_i χαρακτηριστικά.

ΒΗΜΑ 2 : Ταξινόμησε τα χαρακτηριστικά σύμφωνα με τους παράγοντες συμβολής s_j , στο εκπαιδευμένο SVM και επέλεξε τα κορυφαία d_{i+1} χαρακτηριστικά.
ΒΗΜΑ 3: Θέσε i=i+1. Επανέλαβε από το βήμα 1 μέχρι i=k

Ακολουθούμε το σχήμα CV2 για να εκτιμήσουμε το ποσοστό σφάλματος σε κάθε επίπεδο. Στα πειράματα διασταυρωμένης επικύρωσης, διαφορετικά υποσύνολα εκπαίδευσης δημιουργούν διαφορετικές λίστες χαρακτηριστικών (παρόλο που τα περισσότερα από αυτά επικαλύπτονται στα συνήθη πειράματα). Μια μέθοδος επιλογής με βάση τη συχνότητα υιοθετείται για να επιλεχθούν οι λίστες χαρακτηριστικών. Μετά τα βήματα της αναδρομικής επιλογής χαρακτηριστικών για κάθε υποσύνολο, μετράμε για κάθε ένα από τα d_i επίπεδα, τη συχνότητα των χαρακτηριστικών που επιλέγονται μεταξύ όλων των γύρων των πειραμάτων της διασταυρωμένης επικύρωσης. Τα d_i κορυφαία, πιο συχνά επιλεγμένα χαρακτηριστικά αναφέρονται ως τα τελικά d_i χαρακτηριστικά (ονομάζονται τα κορυφαία χαρακτηριστικά).

Στις περισσότερες περιπτώσεις, τα CV2 σφάλματα συνήθως ακολουθούν μια καμπύλη σχήματος U κατά μήκος των σταδίων επιλογής χαρακτηριστικών (αριθμός χαρακτηριστικών). Η εύρεση του ελαχίστου αριθμού χαρακτηριστικών που μπορούν να δώσουν το ελάχιστο CV2 ποσοστό σφάλματος είναι συχνά επιθυμητή για πραγματικές εφαρμογές.

4.10 ΜΕΘΟΔΟΣ GACV (ΓΕΝΙΚΕΥΜΕΝΗ ΠΡΟΣΕΓΓΙΣΗ ΤΗΣ ΔΙΑΣΤΑΥΡΩΜΕΝΗΣ ΕΠΙΚΥΡΩΣΗΣ)

Ο στόχος της μεθόδου GACV είναι η ελαχιστοποίηση της συγκριτικής Kullback-Leibler απόστασης (GCKL) της f_λ στην πραγματική f, σαν συνάρτηση του λ.

4.10.1 ΑΠΟΣΤΑΣΗ KULLBACK-LEIBLER

Η Kullback-Leibler απόσταση είναι ένα μέτρο πληροφορίας, που περιλαμβάνει δύο κατανομές πιθανότητας συσχετισμένες με το ίδιο πείραμα. Η K-L απόκλιση επίσης γνωστή ως σχετική εντροπία, είναι ένα μέτρο του κατά πόσο διαφορετικές είναι δύο κατανομές πιθανότητας. Η K-L απόκλιση των πιθανοτήτων κατανομών P και Q σε ένα πεπερασμένο σύνολο X ορίζεται ως:

$$D(P||Q) = \sum_{x \in X} P(x) \log \frac{P(x)}{Q(x)} \qquad (4.10.1.1)$$

Η K-L απόκλιση μεταξύ των P και Q μπορεί επίσης να θεωρηθεί ως ο μέσος αριθμός των bits που σπαταλώνται με την κωδικοποίηση περιστατικών από την κατανομή P με ένα κώδικα βασισμένο σε μια κατάλληλη κατανομή Q. Αυτή η K-L απόκλιση είναι ένα μη συμμετρικό θεωρητικό μέτρο πληροφορίας της απόστασης P από την Q. Όσο μικρότερη είναι η σχετική εντροπία τόσο πιο όμοια είναι η κατανομή των δύο μεταβλητών.

4.10.2 ΣΥΓΚΡΙΤΙΚΗ KULLBACK-LEIBLER ΑΠΟΣΤΑΣΗ (CKL DISTANCE)

Για τις μονοπαραγοντικές αποκρίσεις Bernoulli, η ποινικοποιημένη μέθοδος πιθανότητας εκτιμά τη $\log it$ συνάρτηση $f(x) = \log[p(x)/(1-p(x))]$, ελαχιστοποιώντας :

$$-\frac{1}{n}\sum_{i=1}^{n}[y_i f(x_i) - b(f(x_i))] + \frac{\lambda}{2}J(f) \qquad (4.10.2.1)$$

Η διαφορά μεταξύ του πραγματικού f και του εκτιμημένου f_λ, μπορεί να μετρηθεί με τη CKL απόσταση:

$$CKL(f, f_\lambda) = -\frac{1}{n}\sum_{i=1}^{n}[\mu_i f_\lambda(x_i) - b(f_\lambda(x_i))] \qquad (4.10.2.2)$$

όπου $\mu_i = b^{\cdot}(f_i) = E_f(Y_i)$. Η CKL μπορεί να θεωρηθεί ως η αναμενόμενη αρνητική συνάρτηση λογαριθμικής πιθανότητας για την f_λ. Εντούτοις το μ_i, είναι μια άγνωστη ποσότητα. Για να ψάξουμε για το καλύτερο λ, για να ελαχιστοποιήσουμε την CKL απόσταση, μια μετρήσιμη proxy πρέπει να εξαχθεί.

4.10.3 Η ΓΕΝΙΚΕΥΜΕΝΗ KULLBACK-LEIBLER ΑΠΟΣΤΑΣΗ (GCKL DISTANCE)

Έστω ότι τα μη παρατηρούμενα y_i θα παραχθούν σύμφωνα με μοντέλο πιθανότητας με $p(t) = p_{true}(t)$, να είναι η πιθανότητα που ένα άγνωστο στοιχείο με χαρακτηριστικό διάνυσμα t , να ανήκει στην κατηγορία A. Έστω επίσης y_i η μη παρατηρούμενη αξία του y που συνδέεται με το t_j. Δεδομένης της f_λ, ορίζεται η γενικευμένη συγκριτική Kullback-Leibler απόσταση (GCKL), σε σχέση με τη g ως:

$$GCKL(f_{true}, f_\lambda) = GCKL(\lambda) = E_{true}\frac{1}{n}\sum_{j=1}^{n} g\left(y_j, f_{\lambda_j}\right) \qquad (4.3.10.1)$$

Δεδομένου ότι για $y_{i=}\pm 1$ και 0<ε<1 , $V_\varepsilon(y_i, f_i) = \max\ \{0, |1 - y_i f_i| - \varepsilon\}$, η $g(t) = [|1-\tau| - \varepsilon]_+$, αντιστοιχεί στο σύνηθες SVM. Να σημειωθεί ότι αυτή η $g(t)$ δεν είναι μονότονη στο τ , αλλά είναι 0 για $\tau \in [1-\varepsilon, 1+\varepsilon]$ και αυξάνει γραμμικά εκτός του διαστήματος καθώς βγαίνει εκτός του διαστήματος προς κάθε κατεύθυνση. Η σχέση της GCKL απόστασης, με το ποσοστό της εσφαλμένης ταξινόμησης, δεν είναι τόσο άμεση αλλά εξακολουθεί να είναι χρήσιμη.

4.10.4 ΜΙΑ ΥΠΟΛΟΓΙΣΙΜΗ PROXY ΓΙΑ ΤΗ GCKL-ΠΡΟΣΕΓΓΙΣΗ ΤΗΣ ΔΙΑΣΤΑΥΡΩΜΕΝΗΣ ΕΠΙΚΥΡΩΣΗΣ

Στη συνέχεια βρίσκεται μια μετρήσιμη proxy για τη GCKL(λ) ,όπου μια proxy για τη GCKL(λ) είναι μια μετρήσιμη συνάρτηση, της οποίας ο ελαχιστοποιητής I_λ είναι μια καλή προσέγγιση του ελαχιστοποιητή της GCKL(λ), στην περίπτωση που ο I_λ είναι αυστηρά κυρτός. Συγκεκριμένα περιγράφεται μια διαδικασία leaving-out-one διασταυρωμένης επικύρωσης για τη GCKL και μια σειρά προσεγγίσεων έτσι ώστε να ληφθεί μια προσέγγιση της proxy για τη GCKL. Έπειτα γίνεται μια διαδικασία τυχαιοποίησης και πιο συγκεκριμένα μια εκτίμηση Monte Carlo.

Έστω $f_\lambda^{[-i]}$ η λύση του εξής προβλήματος μεταβολής: εύρεση $f \in \{span\varphi_\nu\} \oplus H_k$ για την ελαχιστοποίηση της:

$$I_\lambda[f,Y] = \frac{1}{n}\sum_{\substack{j=1\\ j\neq i}} g\left(y_j, f_j\right) + J_\lambda(f) \qquad (4.10.4.1)$$

όπου $J_\lambda(f)$ είναι μια τετραγωνική ποινή της f εξαρτώμενης του λ.
Εν συνεχεία ορίζεται η leaving-out-one συνάρτηση $V_0(\lambda)$ ως εξής:

$$V_0(\lambda) = \frac{1}{n}\sum_{i=1}^{n} g\left(y_i f_\lambda^{[-i]}\right), \qquad (4.10.4.2)$$

και θεωρείται ως μια εκτίμηση για τη GCKL(λ). Έστω επίσης

$$V_0(\lambda) = OBS(\lambda) + D(\lambda) \qquad (4.10.4.3)$$

όπου $OBS(\lambda)$ είναι το παρατηρηθέν ταίρι του f_λ για τα δεδομένα

$$OBS(\lambda)=\frac{1}{n}\sum_{i=1}^{n} g\left(y_i\, f_{\lambda_\iota}\right) \qquad (4.10.4.4)$$

και η

$$D(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left[g\left(y_i f_\lambda^{[-i]}\right) - g\left(y_i\, f_{\lambda_\iota}\right)\right] \approx -\frac{1}{n}\sum_{i=1}^{n} g_i' h_{ii}\frac{y_i' - \mu_{\lambda_i}}{1-\mu_i' h_{ii}} \qquad (4.10.4.5)$$

όπου

$$g_i' = \frac{\partial g}{\partial f_{\lambda_i}} \qquad (4.10.4.6)$$

$$\mu_i' = \frac{\partial \mu}{\partial f_{\lambda_i}} \qquad (4.10.4.7)$$

$$h_{ii} = \frac{f_{\lambda_\iota} - f_\lambda^{[-i]}}{y_i - \mu_\lambda^{[-i]}} \qquad (4.10.4.8)$$

Τα g_i, μ_{λ_i} *και* μ_i' εξαρτώνται μόνο από το f_{λ_ι} . Μένει μόνο να βρεθεί μια προσέγγιση του h_{ii}, η οποία θα γίνει με τη βοήθεια του leaving-out-one λήμματος.

4.10.5 ΤΟ LEAVING-OUT-ONE ΛΗΜΜΑ

Έστω $f_\lambda^{[-i]}$ ο ελαχιστοποιητής της σχέσης (4.10.4.1), τότε ο $f_\lambda^{[-i]}$είναι επίσης ο ελαχιστοποιητής της σχέσης $\frac{1}{n}\sum_{i=1}^{n} g\left(y_j f_j\right) + I_\lambda(f)$ με την προϋπόθεση ότι τα δεδομένα έχουν την παρακάτω μορφή

$$Y^{[-i]} = \left\{y_1, \dots, y_{i-1}, \mu f\left(\lambda_i^{[-i]}\right), y_{i+1}, \dots, y_n\right\}.$$

Δηλαδή, αν εξαιρέσουμε το i-οστό στοιχείο δεδομένων, γίνεται χρήση της συνάρτησης πρόβλεψης $\mu f\left(\lambda_i^{[-i]}\right)$ για να προβλεφθεί το y_i και να λυθεί το πρόβλημα μεταβολής με τη $\mu f\left(\lambda_i^{[-i]}\right)$ να υποκαθιστά το y_i έτσι ώστε να λαμβάνεται το $f\lambda_i^{[-i]}$ για τη λύση.

4.10.6 ΜΙΑ ΠΡΟΣΕΓΓΙΣΗ ΓΙΑ ΤΟ $h_{ii} = \frac{f_{\lambda_i} - f_\lambda^{[-i]}}{y_i - \mu_\lambda^{[-i]}}$

Έστω

$$f_\lambda = (f_{\lambda 1}, .., f_{\lambda n})' \qquad (4.10.6.1)$$

και $$f_\lambda^{[-i]} = \left(f_{\lambda 1}^{[-i]}, \dots, f_{\lambda n}^{[-i]}\right)' \qquad (4.10.6.2)$$

Στη συνέχεια αναπτύσσουμε το διάνυσμα $\frac{\partial I_\lambda}{\partial f}$ σε σειρά Taylor των f_λ και Y και καθώς η f_λ είναι ελαχιστοποιητής της $I_\lambda(f, Y)$ και από το Λήμμα Leaving-Out-One, καθώς επίσης η $f_\lambda^{[-i]}$ είναι ελαχιστοποιητής της $I_\lambda\left(f, Y^{[-i]}\right)$, έτσι έχουμε ότι:

$$\frac{\partial I_\lambda}{\partial f}\left(f_\lambda^{[-i]}, Y^{[-i]}\right) = \frac{\partial I_\lambda}{\partial f}(f_\lambda, Y) = 0 \qquad (4.10.6.3)$$

και συνεπώς

$$H_{ff}\left(f_\lambda - f_\lambda^{[-i]}\right) \approx -H_{fY}\left(Y - Y^{[-i]}\right) \qquad (4.10.6.4)$$

Άρα $h_{ii} \approx -\left(H_{ff}^{-1} H_{fY}\right)_{ii} = \widetilde{h_{ii}}$

όπου $H_{ff} = \frac{\partial^2 I_\lambda}{\partial f \partial f'}$ (4.10.6.5) και $H_{fy} = \frac{\partial^2 I_\lambda}{\partial f \partial Y'}$

Θέτοντας το $\widetilde{h_{ii}}$ στην (4.10.4.2) παίρνουμε την προσεγγιστική διασταυρωμένη επικύρωση ACV(λ) και η (4.10.4.1) γίνεται:

$$ACV(\lambda) = \frac{1}{n}\sum_{i=1}^{n} g\left(y_i\, f_{\lambda_i}\right) - \frac{1}{n}\sum_{i=1}^{n} g_i' \widetilde{h_{ii}} \frac{y_i' - \mu_{\lambda_i}}{1 - \mu_i' \widetilde{h_{ii}}} \qquad (4.10.6.5)$$

4.10.7 ΤΟ ΤΥΧΑΙΟΠΟΙΗΜΕΝΟ ΙΧΝΟΣ ΕΚΤΙΜΗΣΗΣ ΤΗΣ GACV ΓΙΑ ΤΟ $g(\tau) = [\varepsilon - \tau]_+$

Για $g(\tau) = [\varepsilon - \tau]_+$και δεδομένου ότι H_{ff}αντιστρέψιμος

(4.10.4.5)⇒ $$D(\lambda) \approx \frac{1}{n}\sum_{y_i f_{\lambda_i} - < \varepsilon} 2\widetilde{h_{ii}} + \frac{1}{n}\sum_{-\varepsilon < y_i f_{\lambda_i} < -\varepsilon} \widetilde{h_{ii}} \qquad (4.10.7.1)$$

$$= -\frac{2}{n} trace E_\varepsilon H_{ff}^{-1} H_{fy} \qquad (4.10.7.2)$$

Στη συνέχεια κάνουμε ένα εκφυλισμό των δεδομένων, απαιτούμε ο H_{ff} να είναι θετικά ορισμένος και έστω $Z = (z_1, \dots, z_n)'$, όπου τα z_i παράγονται από μια γεννήτρια τυχαίων αριθμών. Υποθέτουμε επίσης ότι οι $f_\lambda = f_\lambda^Y$και η f_λ^{Y+Z} είναι ελαχιστοποιητές των $I_\lambda(f, Y)$ και $I_\lambda(f, Y + Z)$ αντίστοιχα. Έτσι εκφυλίζεται το διάνυσμα Y με την προσθήκη του τυχαίου εκφυλισμού Z. Με τη χρήση του αναπτύγματος της σειράς Taylor και του ότι $EZ'AZ = \sigma_Z^2 trace A$ προκύπτει ότι:

$$H_{ff}\left(f_\lambda^{Y+Z} - f_\lambda^Y\right) \approx -H_{fy}\mathrm{Z} \Rightarrow \qquad (4.10.7.3)$$

$f_\lambda^{Y+Z} - f_\lambda^{Y} \approx -H_{fy}\mathrm{Z} \Rightarrow$ (4.10.7.4)

$\frac{2}{n}\frac{1}{\sigma_Z^2} E_e\left(f_\lambda^{Y+Z} - f_\lambda^{Y}\right)$ (4.10.7.5)

Η (4.10.7.5) αποτελεί εκτιμητή της (4.10.7.2) και λογικό τυχαιοποιημένο εκτιμητή της D(λ) στη (4.10.7.1). Τελικά η τυχαιοποιημένη συνάρτηση GACV ορίζεται ως εξής:

$$ranGACV(\lambda) = \frac{1}{n}\sum_{y_i f_{\lambda_i}}\left[\varepsilon - y_i f_{\lambda_i}^{Y}\right]_+ + \frac{2}{n}\frac{1}{\sigma_Z^2} Z' E_\varepsilon\left(f_\lambda^{Y+Z} - f_\lambda^{Y}\right) \quad (4.10.7.6)$$

Ο ελαχιστοποιητής της (4.10.7.6) για μεγάλα δείγματα μπορεί να είναι ένας καλός εκτιμητής για τη GCKL(λ).

	$y_i f_i < -\epsilon$	$-\epsilon < y_i f_i < \epsilon$	$\epsilon < y_i f_i$
$g(y_i f_i)$	$\epsilon - y_i f_i$		0
$\frac{\partial g}{\partial f_i}$	$-y_i$		0
$\frac{\partial^2 g}{\partial f_i \partial y_i}$	-1		0
$\frac{\partial g}{\partial f_i}(y_i - \mu_i)$	-2	-1	0
	$f_i < -\epsilon$	$-\epsilon < f_i < \epsilon$	$\epsilon < f_i$
$\mu(f_i)$	-1	0	1
$\frac{d\mu}{df_i}$	0	0	0

ΣΧΗΜΑ 4.10.1 : Συστατικά του $D(\lambda)$ για $g(t) = [\varepsilon - \tau]_+$

4.10.8 ΣΥΖΗΤΗΣΗ ΓΙΑ ΤΗ ranGACV

Στη συγκεκριμένη ενότητα, η leaving-out-one συνάρτηση γράφεται στη μορφή $-y_i f_{\lambda_i}^{[-i]} + b\left(f_{\lambda_i}\right), \text{όπου } y_i = 0 \text{ και } y_i = 1\text{, με πιθανότητες } p_i \text{ και } 1 - p_i, \text{αντίστοιχα}$ (δηλαδή στη μορφή της αρνητικής λογαριθμικής πιθανότητας $-yf + b(f)$, όπου $b(f) = \ln\left(1 + e^f\right)$). Το τελικό αποτέλεσμα που ονομάζεται ACV (βασίζεται στο ότι $y_i \in \{0,1\}$) συνεπάγεται ότι:

$$D(\lambda) = \sum_{i=1}^{n} \frac{y_i \widetilde{h_{ii}}\left(y_i - \mu_{\lambda_i}\right)}{1 - \mu_i' \widetilde{h_{ii}}}$$

Το D εγκαταστάθηκε από το D_{GACV} που ορίζεται ως εξής:

$$D_{GACV}(\lambda)=\bar{h}\frac{\sum_{i=1}^{n} y_i\left(y_i-\mu_{\lambda_i}\right)}{1-\left(\overline{\mu'}h\right)}$$

όπου $\bar{h}=\frac{1}{n}\sum_{i=1}^{n}\widetilde{h_{ii}}$ και $\overline{\mu'}h=\frac{1}{n}\sum_{i=1}^{n}\mu_i'\tilde{h}_{ii}$ και αποδείχτηκε ότι η προκύπτουσα $D_{GACV}(\lambda)$ αποτελεί μια εξαιρετική proxy για τη CKL.

4.11 ΕΝΔΕΙΞΗ ΥΠΟΛΟΓΙΣΜΕΝΗ ΜΕ ΤΗ ΜΕΘΟΔΟ LAPLACE

4.11.1 ΕΙΣΑΓΩΓΗ

Πρόσφατα οι επιστήμονες έχουν εφαρμόσει Μπεϋζιανές μεθόδους για να αντιμετωπίσουν τα προβλήματα της ακρίβειας των μεθόδων όσον αφορά τον υπολογιστικό χρόνο, τα δεδομένα εκπαίδευσης και την εύρεση καλών παραμέτρων και χαλαρών ορίων. Εν γένει η Μπεϋζιανή προσέγγιση είναι λογικά συνεπής, απλή και ευέλικτη. Σε σύγκριση με την παραδοσιακή προσέγγιση, οι Μπεϋζιανές μέθοδοι παρέχουν ένα ισχυρό πλαίσιο για την αυτόματη προσαρμογή των παραμέτρων συστηματοποίησης και πυρήνα στις σχεδόν βέλτιστες τιμές, χωρίς την ανάγκη να παραμεριστούν σε ένα σύνολο επικύρωσης. Επιπλέον, το Μπεϋζιανό πλαίσιο επιτρέπει την αντικειμενική σύγκριση μεταξύ των λύσεων που χρησιμοποιούν διαφορετικές αρχιτεκτονικές δικτύων. Οι Μπεϋζιανές τεχνικές επίσης προσφέρουν και κάποια άλλα σπουδαία χαρακτηριστικά. Για παράδειγμα, στα προβλήματα παλινδρόμησης οι γραμμές σφάλματος μπορούν να ανατεθούν σε προβλέψεις δικτύου.

Στα προβλήματα ταξινόμησης, με τη χρήση των μέσων αποτελεσμάτων, η τάση των συμβατικών προσεγγίσεων να κάνουν υπεραισιόδοξες προβλέψεις σε περιοχές αραιών δεδομένων μπορεί να αποφευχθεί. Μεταξύ άλλων οι Μπεϋζιανές τεχνικές έχουν επίσης χρησιμοποιηθεί για ενεργητική μάθηση και στο σχηματισμό μιας επιτροπής δικτύων. Συνεπώς είναι πολλά υποσχόμενο το να ενσωματωθούν τα SVMs με αυτές τις Μπεϋζιανές ιδέες.

Μια Μπεϋζιανή ερμηνεία των SVMs, έχει προταθεί από τον Smola. Συγκεκριμένα έχει δειχθεί ότι η χρήση των διαφορετικών πυρήνων στα SVMs μπορεί να θεωρηθεί ως ο ορισμός διαφορετικών εκ των προτέρων κατανομών πιθανότητας στο χώρο συναρτήσεων, καθώς $P[f] \propto exp\left(-\beta\left\|\hat{P}f\right\|^2\right)$. Εδώ το β>0 είναι μια σταθερά και το $\hat{P}$ ένας διαχειριστής συστηματικοποίησης που αντιστοιχεί στον επιλεγμένο πυρήνα.

4.11.2 ΤΟ ΠΛΑΙΣΙΟ ΤΗΣ ΕΝΔΕΙΞΗΣ ΚΑΙ Η ΕΦΑΡΜΟΓΗ ΣΤΟ SVM

Το πλαίσιο της ένδειξης είναι ένα Μπεϋζιανό πλαίσιο που προτάθηκε από τον Mckay και υπολογιστικά είναι ισοδύναμο με τη μέθοδο μεγίστης πιθανότητας τύπου II στη Μπεϋζιανή Στατιστική. Το πλαίσιο αυτό έχει εφαρμοστεί επιτυχώς σε προβλήματα συστηματικοποίησης και παλινδρόμησης.

Πρώτα εισάγονται κάποιες έννοιες, που χρησιμοποιούνται στο αποδεικτικό πλαίσιο. Ένα μοντέλο H με ένα διάνυσμα w , k-διάστατων παραμέτρων που αποτελείται από τη λειτουργική του μορφή f, την κατανομή $p(D|w,H)$, που το μοντέλο δημιουργεί για τα δεδομένα D και μια εκ των προτέρων παραμετρική κατανομή $p(w|H,\lambda)$ που συνήθως γράφεται στη μορφή:

$$p(w|H,\lambda) = \frac{exp\big(-\lambda E_w(w|H)\big)}{Z_w(\lambda)}$$

Εδώ το λ είναι μια παράμετρος συστηματικοποίησης και η ποσότητα

$$Z_w(\lambda) = \int d^k w exp(-\lambda E_w)$$

χρησιμοποιείται για κανονικοποίηση.

Το πλαίσιο ένδειξης διαιρείται σε τρία επίπεδα συμπερασμάτων. Για δοθείσα τιμή του λ, το πρώτο επίπεδο συμπερασμάτων συμπεραίνει μια εκ των υστέρων κατανομή του w. Το δεύτερο επίπεδο συμπερασμάτων καθορίζει την αξία του λ, μεγιστοποιώντας τη:

$$p(\lambda|D,H) \propto p(D|\lambda,H)p(\lambda|H)$$

Όταν $p(\lambda|H)$ είναι μια σταθερά εκ των προτέρων κατανομή, η $p(D|\lambda,H)$, η ένδειξη του λ, μπορεί να χρησιμοποιηθεί για να ορίσει μια προτίμηση σε εναλλακτικές τιμές του λ και στη συνέχεια μια σχεδόν βέλτιστη τιμή του λ μπορεί να ληφθεί με επαναληπτικό τρόπο. Το τρίτο επίπεδο συμπερασμάτων στην ένδειξη συμπερασμάτων ταξινομεί τα διάφορα μοντέλα εξετάζοντας τις εκ των υστέρων πιθανότητες $p(H|D) \propto p(D|H)p(H)$. Στη συνέχεια το πλαίσιο ένδειξής εφαρμόζεται στο SVM, υπολογίζεται το Εσσιανό[1] και προκύπτει ένα σύστημα ιδιοτιμών με χρονική πολυπλοκότητα $O(N^3)$ που μπορεί να είναι υπολογιστικά ακριβές για μεγάλο N (όπου N=# ιδιοδιανυσμάτων). Στη συνέχεια το δεύτερο επίπεδο συμπερασμάτων για τα SVMs καθορίζει τη τιμή της λ, μεγιστοποιώντας τη $p(D|\lambda,H)$ και για να αποκτηθεί η ένδειξη του μοντέλου $p(D|H)$ στο τρίτο επίπεδο συμπερασμάτων ή οι επαναλήψεις της λ στο δεύτερο επίπεδο, πρέπει να υπολογιστεί ο αποδοτικός αριθμός παραμέτρων.

4.11.3 ΠΡΟΣΟΜΟΙΩΣΗ

Θα αναφερθούν, τα αποτελέσματα σχετικά με την εφαρμογή του αποδεικτικού πλαισίου για τα SVMs, με τον πολυωνυμικό και το Γκαουσιανό πυρήνα. Γίνεται χρήση τριών συνόλων δεδομένων για τους πειραματισμούς και το n τίθεται ίσο με 100. Ο διαχωρισμός των συνόλων δεδομένων σε σύνολο εκπαίδευσης και δοκιμής γίνεται ως εξής για τα τρία σύνολα δεδομένων στα οποία διεξάγεται η προσομοίωση:

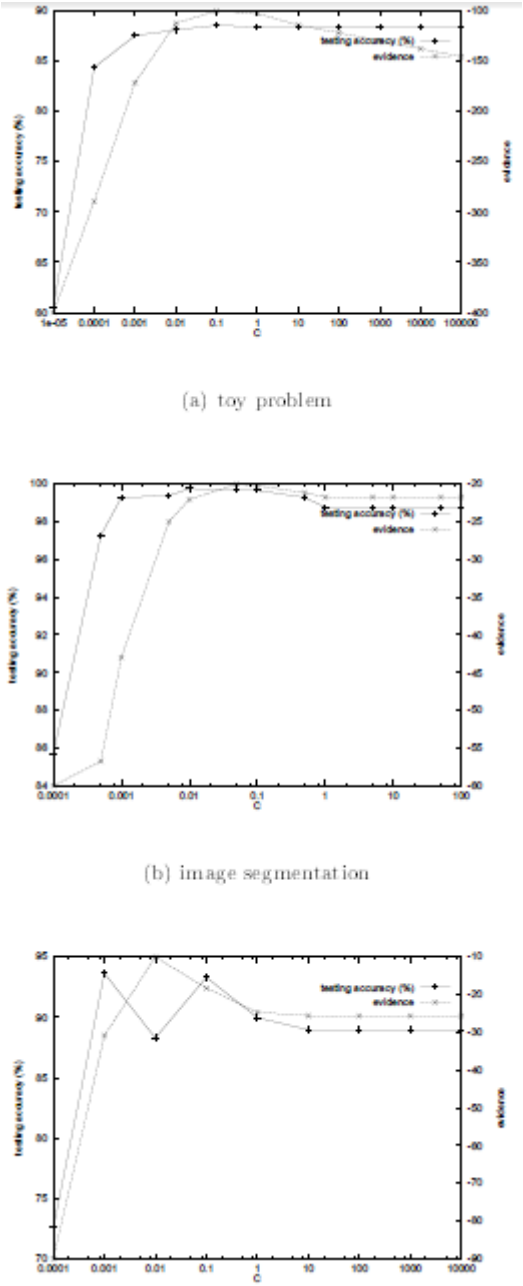
ΣΥΝΟΛΟ ΔΕΔΟΜΕΝΩΝ	# ΠΡΟΤΥΠΩΝ ΤΟΥ ΣΥΝΟΛΟΥ ΕΚΠΑΙΔΕΥΣΗΣ	# ΠΡΟΤΥΠΩΝ ΤΟΥ ΣΥΝΟΛΟΥ ΔΟΚΙΜΗΣ
1	500	10^4
2	210	2100
3	150	299

Στη συνέχεια γίνεται επιλογή των παραμέτρων συστηματοποίησης και πυρήνα. Γίνεται χρήση της ένδειξης για το λ, *και το* $p(D|\lambda,H)$ υπολογίζεται για να ταξινομηθούν οι διαφορετικές τιμές του λ. Στις περισσότερες περιπτώσεις η ένδειξη για το λ ακολουθεί την ακρίβεια ελέγχου πολύ στενά. Μια σχεδόν βέλτιστη τιμή για το λ μπορεί να αποκτηθεί. Η δοκιμή ακρίβειας για το SVM που λαμβάνεται στον επαναληπτικό τρόπο είναι πολύ κοντά στην ακρίβεια δοκιμών για το καλύτερο SVM στο φάσμα των εξεταζόμενων τιμών. Τα

[1] A= $\nabla^2 M = \nabla^2(\lambda E_w + \sum_{i=1}^{N} \xi_i)$,όπου $\xi_i = [1 - y_i a_i]_+$

αποτελέσματα είναι ιδιαιτέρως ευνοϊκά για τους Γκαουσιανούς πυρήνες. Επιπλέον, ο αριθμός των επαναλήψεων που απαιτούνται για την εύρεση της σχεδόν βέλτιστης τιμής είναι αρκετά μικρός. Στη συνέχεια γίνεται εκτίμηση των παραμέτρων πυρήνα που περιλαμβάνουν τον πολυωνυμικό βαθμό d για τους πολυωνυμικούς πυρήνες και του εύρους σ για τους τους Γκαουσιανούς .Η παράμετρος συστηματοποίησης C (για σταθερά d ή σ) εκτιμάται επαναληπτικά. Έπειτα σχεδιάζεται το $p(D|H)$ και το ποσοστό των σωστών συστηματικοποιήσεων στο σύνολο δοκιμής, για διαφορετικές τιμές του d όταν γίνεται χρήση ενός πολυωνυμικού πυρήνα. Η ένδειξη αν και όχι τέλεια ακόμα ακολουθεί στενά την ακρίβεια ελέγχου.

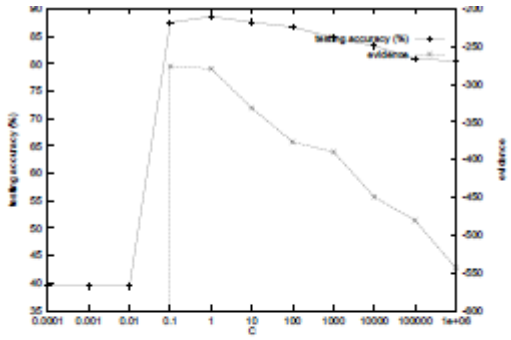
N



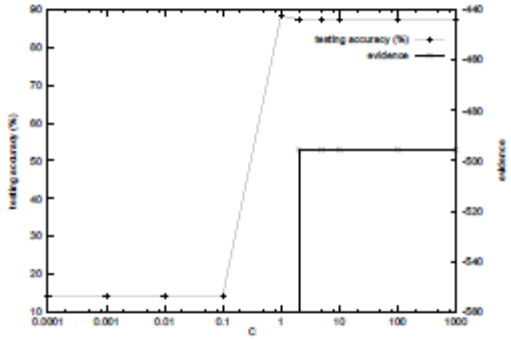
(a) toy problem

(b) image segmentation

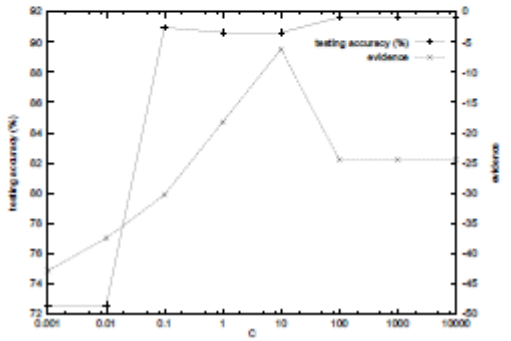
ΣΧΗΜΑ 4.11.1: Αποτελέσματα της χρήσης διαφορετικών αξιών του C (πολυωνυμικός πυρήνας)



(a) toy problem



(b) image segmentation



ΣΧΗΜΑ 4.11.2: Αποτελέσματα της χρήσης διαφορετικών αξιών του C (Γκαουσιανός πυρήνας)

TABLE I

RESULTS OBTAINED BY ITERATING LEVELS 1 AND 2 (POLYNOMIAL KERNEL).

	C obtained	testing accuracy (%)	number of iterations	best C in range tested	best testing accuracy (%)
toy problem	6.39	88.4	2	0.1	88.6
image segmentation	0.91	98.8	7	0.01	99.8
breast cancer	0.125	92.3	5	0.001	93.6

ΣΧΗΜΑ 4.11.3: Αποτελέσματα από τα επαναλαμβανόμενα επίπεδα 1 και 2 (πολυωνυμικός πυρήνας)

TABLE II
RESULTS OBTAINED BY ITERATING LEVELS 1 AND 2 (GAUSSIAN KERNEL).

	C obtained	testing accuracy (%)	number of iterations	best C in range tested	best testing accuracy (%)
toy problem	0.55	88.7	5	1	88.6
image segmentation	1.71	87.5	2	1.71	87.5
breast cancer	80.7	91.6	5	80.7	91.6

ΣΧΗΜΑ 4.11.4: Αποτελέσματα από τα επαναλαμβανόμενα επίπεδα 1 και 2 (Γκαουσιανός πυρήνας)

4.12 ξα-ΕΚΤΙΜΗΣΗ

4.12.1 ΕΙΣΑΓΩΓΗ

Ο στόχος αυτής της μεθόδου είναι η ανάπτυξη επιχειρησιακών εκτιμητών απόδοσης που είναι πραγματικής χρήσης όταν εφαρμόζονται στα SVMs. Αυτό προϋποθέτει ότι οι εκτιμητές είναι αποδοτικοί και υπολογιστικά επαρκείς. Οι συγκεκριμένοι εκτιμητές (ξα-estimators) ξεπερνούν τα προβλήματα (π.χ. το να είναι δύχρηστοι σε πρακτικές εφαρμογές ή το να είναι υπολογιστικά ανεπαρκείς) επεκτείνοντας τα αποτελέσματα των Vapnik (1998) και Jaakola και Haussler (1999) στη γενική μορφή των SVMs και είναι ακριβείς και επαρκείς για τον υπολογισμό.

4.12.2 ΤΑΞΙΝΟΜΗΣΗ ΚΕΙΜΕΝΟΥ

Η ταξινόμηση κειμένου είναι το κύριο πεδίο εφαρμογής για τη συγκεκριμένη μέθοδο και έχει ως στόχο της την εκχώρηση σε ένα σταθερό αριθμό σημασιολογικών κατηγοριών. Η χρήση ενός δυαδικού ταξινομητή για κάθε τέτοια κατηγορία οδηγεί στο εξής πρόβλημα μάθησης με επίβλεψη:
Δεδομένου ενός ανεξάρτητου και ισόνομου δείγματος εκπαίδευσης

$$S_n: (\overrightarrow{x_1}, y_1); \dots; (\overrightarrow{x_n}, y_n)$$

μεγέθους n, από μια δεδομένη αλλά άγνωστη κατανομή $Pr(\vec{x}, y)$, που περιγράφει το καθήκον μάθησης. Το $\overrightarrow{x_i}$ αναπαριστά το περιεχόμενο του αρχείου και το $y_i \in \{-1,1\}$ υποδεικνύει την κλάση. Ο L (learner), στοχεύει στην εύρεση ενός κανόνα απόφασης $h_L: \vec{x} \to \{-1,1\}$, βασισμένος στο S_n που ταξινομεί τα νέα αρχεία με όσο το δυνατόν μεγαλύτερη ακρίβεια. Γίνεται ευρεία χρήση, των σκορ που βασίζονται στην ακρίβεια (precision)[2] και στην ανάκληση (recall)[3] . Η ανάκληση, Rec(h), ενός κανόνα απόφασης h, ορίζεται ως η πιθανότητα ένα έγγραφο με ετικέτα y=1 να ταξινομηθεί σωστά. Η ακρίβεια, Prec(h), ενός κανόνα απόφασης h, είναι η πιθανότητα ένα έγγραφο που ταξινομείται ως $h(\vec{x})$=1, να ταξινομηθεί σωστά. Μεταξύ υψηλής ακρίβειας και υψηλής ανάκλησης υπάρχει

[2] Precision=$\frac{\text{αριθμός ανακτήμενων εγγράφων που είναι σχετικά}}{\text{ολικός αριθμός εγγράφων που ανακτώνται}}$

[3] Recall= $\frac{\text{αριθμός ανακτήμενων εγγράφων που είναι σχετικά}}{\text{ολικός αριθμός εγγράφων που είναι σχετικά}}$

ένας συμβιβασμός. Μια δημοφιλής μέθοδος, για να συνδυαστούν και τα δύο σε ένα και μόνο μέσο απόδοσης είναι το F_1, που ορίζεται ως ο γεωμετρικός ακρίβειας και ανάκλησης.

4.12.3 ΕΚΤΙΜΗΤΕΣ ΕΠΑΡΚΟΥΣ ΑΠΟΔΟΣΗΣ ΓΙΑ ΤΑ SVMS

Οι ξα-εκτιμητές βασίζονται στην ιδέα της leave-one-out εκτίμησης. Ο leave-one-out εκτιμητής του ποσοστού σφάλματος εξελίσσεται ως εξής. Από το δείγμα εκπαίδευσης

$$S_n: \left((\overrightarrow{x_1}, y_1); \dots; (\overrightarrow{x_n}, y_n)\right)$$

το πρώτο παράδειγμα αφαιρείται. Το προκύπτον δείγμα $S^{/1} = \left((\overrightarrow{x_2}, y_2); \dots; (\overrightarrow{x_n}, y_n)\right)$, χρησιμοποιείται για εκπαίδευση, οδηγώντας σε ένα κανόνα ταξινόμησης $h_L^{/1}$, ο οποίος ελέγχεται στο παρακρατημένο παράδειγμα $(\overrightarrow{x_1}, y_1)$, αν το παράδειγμα ταξινομείται λανθασμένα παράγει ένα leave-one-out σφάλμα. Αυτή η διαδικασία επαναλαμβάνεται για όλα τα παραδείγματα εκπαίδευσης. Ο learner όμως θα πρέπει να επικαλεστεί n φορές για ένα σύνολο εκπαίδευσης μεγέθους n, γεγονός που αποτελεί μειονέκτημα του leave-one-out εκτιμητή καθώς τον καθιστά υπολογιστικά ανεπαρκή και ακριβό για πρακτικά προβλήματα.

Οι ξα-εκτιμητές βασίζονται στη leave-one-out μέθοδο, αλλά δεν απαιτούν στην πραγματικότητα διεξαγωγή επαναδειγματισμού και επανεκπαίδευσης, αλλά μπορούν να εφαρμοστούν απευθείας, μετά από εκπαίδευση του learner στο S_n. Οι εισαγωγές στους εκτιμητές είναι το διάνυσμα $\vec{\alpha}$ που λύνει το δυικό πρόβλημα εκπαίδευσης SVM (OP_2) και το διάνυσμα $\vec{\xi}$ από τη λύση του πρωταρχικού προβλήματος εκπαίδευσης SVM (OP_1).

4.12.4 ΠΟΣΟΣΤΟ ΣΦΑΛΜΑΤΟΣ

Για τα προαναφερθέντα διανύσματα $\vec{\alpha}$ και $\vec{\xi}$ ο ξα-εκτιμητής του ποσοστού σφάλματος $Err_{\xi\alpha}^n(h_L)$ ορίζεται ως εξής:

<u>Ορισμός 1</u>: Για σταθερά SVMs χαλαρού περιθωρίου, ο ξα-εκτιμητής του ποσοστού σφάλματος είναι:

$$Err_{\xi\alpha}^n(h_L) = \frac{d}{n} \text{ , με } = \left|\{i: \rho_{\alpha_i} R_\Delta^2 + \xi_i\} \geq 1\right| \qquad (4.12.4.1)$$

όπου το ρ=2, το R_Δ^2 είναι ένα άνω φράγμα του $K(\vec{x}, \vec{x}) - K(\vec{x}, \overrightarrow{x'})$,για όλα τα $\overrightarrow{x}$ και $\overrightarrow{x'}$. Το ρ υποθέτουμε με βάση τα θεωρητικά αποτελέσματα ότι ισούται με 2 (επίσης το ρ=1 ενδείκνυται για την ταξινόμηση κειμένου), το d μέτρο του αριθμού των παραδειγμάτων εκπαίδευσης για τα οποία η ανισότητα $\left(\rho_{\alpha_i} R^2 + \xi_i\right) \geq 1$ ισχύει. Εάν ένα παράδειγμα $(\overrightarrow{x_i}, y_i)$ από ένα SVM που είναι εκπαιδευμένο στο υπόδειγμα $S_n^{\setminus i}$, τότε θα πρέπει να πληροί την ανισότητα (4.12.4.1) για ένα SVM που είναι εκπαιδευμένο στο πλήρες δείγμα S_n

<u>Λήμμα 1</u>: Ο αριθμός των leave-one-out σφαλμάτων $\sum_{i=1}^n \left(Lh_L^{\setminus i}(x_i), y_i\right)$ των σταθερών SVMs χαλαρού περιθωρίου, σε ένα σύνολο εκπαίδευσης S_n, φράσσεται από:

$$\sum_{i=1}^n \left(Lh_L^{\setminus i}(x_i), y_i\right) \leq |\{(i: 2\alpha_i R^2 + \xi_i) \geq 1\}| \qquad (4.12.4.2)$$

Ενώ το παραπάνω Λήμμα ισχύει για όλες συναρτήσεις πυρήνα που επιστρέφουν θετικές αξίες, είναι πιο αυστηρό όταν η ελάχιστη αξία είναι μηδενική.

Λήμμα 2: Το SVM χαλαρού περιθωρίου είναι αμετάβλητο υπό την προσθήκη μιας πραγματικής αξίας c, στη συνάρτηση πυρήνα.

Θεώρημα 1: Ο ξα-εκτιμητής του ποσοστού σφάλματος αρνητικά μεροληπτικός κατά την ακόλουθη έννοια:

$$\mathcal{E}\left(Err_{\xi\alpha}^{n}(h_L)\right) \geq \mathcal{E}\left(Err^{n-1}(h_L)\right) \qquad (4.12.4.3)$$

Η προσδοκία στην αριστερή πλευρά είναι η υπερεκπαίδευση συνόλου μεγέθους n, ενώ αυτή στη δεξιά πλευρά είναι η υπερεκπαίδευση μεγέθους n-1, δηλαδή ο ξα-εκτιμητής τείνει να υπερεκτιμήσει το πραγματικό ποσοστό σφάλματος. Δεδομένης μιας χαμηλής διακύμανσης της εκτίμησης, μπορεί να διασφαλιστεί ότι το πραγματικό σφάλμα είναι χαμηλότερο από την εκτίμηση.

4.12.5 ΑΝΑΚΛΗΣΗ, ΑΚΡΙΒΕΙΑ ΚΑΙ F_1

Με παρόμοιους ισχυρισμούς, όπως για το ποσοστό σφάλματος, εξετάζονται οι ξα-εκτιμητές για την ανάκληση, την ακρίβεια και το F_1 όπως περιγράφονται παρακάτω.

Ορισμός 2 :Για σταθερά SVMs χαλαρού περιθωρίου, οι ξα-εκτιμητές για την ανάκληση, την ακρίβεια και το F_1 είναι:

$$Rec_{\xi\alpha}^{n}(h_L) = 1 - \frac{d_+}{n_+} \qquad (4.12.5.1)$$

$$Prec_{\xi\alpha}^{n}(h_L) = \frac{n_+ - d_+}{n_+ - d_+ + d_-} \qquad (4.12.5.2)$$

$$F_{1\,\xi\alpha}^{\;n}(h_L) = \frac{2n_+ - 2d_+}{2n_+ - d_+ + d_-} \qquad (4.12.5.3)$$

με τη χρήση των

$$d_+ = \left|\left\{i: y_i = 1 \wedge \left(i: \rho_{\alpha_i} R_\Delta^2 + \xi_i\right) \geq 1\right\}\right| \qquad (4.12.5.4)$$

$$d_- = \left|\left\{i: y_i = -1 \wedge \left(i: \rho_{\alpha_i} R_\Delta^2 + \xi_i\right) \geq 1\right\}\right| \qquad (4.12.5.5)$$

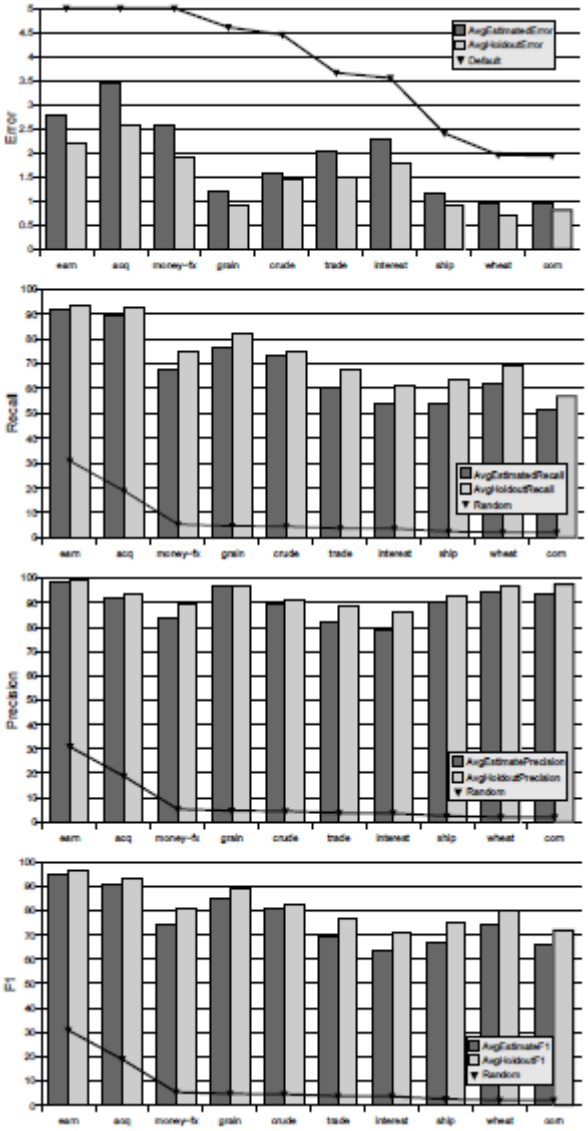
$$n_+ = \left|\{i: y_i = 1\}\right| \qquad (4.12.5.5)$$

όπου d_+ (d_-) είναι ο αριθμός των θετικών (αρνητικών) δειγμάτων εκπαίδευσης για τα οποία η ανισότητα (4.12.4.1) ισχύει και ένα άνω φράγμα του αριθμού των θετικών (αρνητικών) παραδειγμάτων που παράγουν ένα leave-one-out σφάλμα. Το n_+ είναι ο αριθμός των θετικών παραδειγμάτων στο δείγμα εκπαίδευσης.

4.12.6 ΠΕΙΡΑΜΑΤΑ

Διεξάγονται πειράματα για να διαπιστωθεί το κατά πόσο οι ξα-εκτιμητές δουλεύουν στην πράξη. Γίνεται αξιολόγηση για το Reuters-21578 σύνολο δεδομένων (εργασία ταξινόμησης κειμένων) και συγκεκριμένα χρήση του υποσυνόλου ''Mod Apte'' (περιλαμβάνει ένα σύνολο 12.902 εγγράφων) και χρησιμοποιούνται 10 κατηγορίες θεμάτων. Στην ξα-εκτίμηση

θέτουμε ρ≈ 1. Τα αποτελέσματα που προκύπτουν είναι μέσοι όροι 10 τυχαίων διαιρέσεων εκπαίδευσης. Τα σύνολα δοκιμής και εκπαίδευσης είναι ίσου μεγέθους έτσι ώστε να συγκρίνουν εκτιμήσεις της διακύμανσης. Το σύνολο δοκιμής χρησιμοποιείται για τη λήψη μιας εκτίμησης ως μία προσέγγιση της πραγματικής παραμέτρου. Επίσης θέτουμε τις παραμέτρους C=0,5 και R_{Δ}=1. Στη συνέχεια γίνεται σύγκριση των αποδόσεων των μέσων ξα-εκτιμήσεων του σφάλματος, της ακρίβειας, της ανάκλησης και του μέτρου F_1, με τα αντίστοιχα μέσα πραγματικά, μετρημένων σε hold-out σύνολο δεδομένων του ιδίου μεγέθους με το σύνολο εκπαίδευσης για ρ=1 και ρ=2. Οι ξα-εκτιμητές υπερεκτιμούν ελαφρώς το πραγματικό σφάλμα και υποτιμούν την ακρίβεια, την ανάκληση και το F_1. Για ρ=2, οι ξα-εκτιμητές είναι ουσιαστικά πιο προκατειλημμένοι από ότι για ρ=1. Επίσης οι τυπικές αποκλίσεις για τις ξα-εκτιμήσεις είναι παρόμοιες με αυτές για τις hold-out εκτιμήσεις. Στη συνέχεια γίνεται σύγκριση των μέσων ξα και hold-out εκτιμήσεων για μικρότερα σύνολα δεδομένων και συμπεραίνεται ότι δεν υπάρχει ισχυρή συστηματική σύνδεση μεταξύ της μεροληψίας και του μεγέθους του συνόλου εκπαίδευσης και οι μέσες εκτιμήσεις είναι σχεδόν ίσες. Όσον αφορά τη διακύμανση, το μέγεθος του συνόλου εκπαίδευσης έχει ισχυρή επιρροή στις ξα και hold-out εκτιμήσεις. Συγκεκριμένα η διακύμανση και το μέγεθος είναι αντιστρόφως ανάλογα ποσά. Εν γένει, οι εκτιμήσεις ξα με ρ=1, είναι προτιμητέες σε σχέση με αυτές για ρ=2. Επίσης οι ξα εκτιμητές είναι συντηρητικοί όσον αφορά τη μεροληψία, η οποία είναι χαμηλή, ενώ η διακύμανση είναι ουσιαστικά τόσο χαμηλή όσο και για τους hold-out εκτιμητές.



ΣΧΗΜΑ 4.12.1: Διαγράμματα που συγκρίνουν τις μέσες ξα-εκτιμήσεις (ρ=1) του σφάλματος, της ανάκλησης, της ακρίβειας και του μέτρου F_1 μετρημένων σε ένα σύνολο hold-out για τις 10 πιο συχνές κατηγορίες Reuters.

$\rho = 1$	$Err_{\xi\alpha}(h_{\mathcal{L}})$	$Err(h_{\mathcal{L}})$	$Rec_{\xi\alpha}(h_{\mathcal{L}})$	$Rec(h_{\mathcal{L}})$
earn	2.68 ± 0.08	1.87 ± 0.15	92.3 ± 0.3	94.8 ± 0.5
acq	3.54 ± 0.16	2.53 ± 0.20	88.1 ± 0.8	92.5 ± 0.5
money-fx	2.67 ± 0.12	1.92 ± 0.14	64.9 ± 2.0	73.6 ± 2.4
grain	1.23 ± 0.09	0.82 ± 0.11	74.4 ± 1.7	82.7 ± 2.1
crude	1.58 ± 0.08	1.30 ± 0.10	72.5 ± 1.7	76.6 ± 2.6
trade	1.99 ± 0.11	1.42 ± 0.10	60.4 ± 2.1	70.0 ± 2.4
interest	2.20 ± 0.18	1.67 ± 0.20	54.0 ± 5.2	63.8 ± 4.3
ship	1.23 ± 0.10	0.96 ± 0.13	47.5 ± 5.6	61.2 ± 3.9
wheat	0.91 ± 0.07	0.65 ± 0.08	62.8 ± 1.8	71.1 ± 2.3
corn	0.90 ± 0.05	0.71 ± 0.06	52.4 ± 4.1	61.8 ± 1.7
$\rho = 2$	$Err_{\xi\alpha}(h_{\mathcal{L}})$	$Err(h_{\mathcal{L}})$	$Rec_{\xi\alpha}(h_{\mathcal{L}})$	$Rec(h_{\mathcal{L}})$
earn	6.79 ± 0.21	1.87 ± 0.15	84.4 ± 0.4	94.8 ± 0.5
acq	12.99 ± 0.28	2.53 ± 0.20	59.3 ± 1.4	92.5 ± 0.5
money-fx	5.38 ± 0.22	1.92 ± 0.14	38.4 ± 2.0	73.6 ± 2.4
grain	3.20 ± 0.19	0.82 ± 0.11	48.1 ± 2.0	82.7 ± 2.1
crude	3.69 ± 0.20	1.30 ± 0.10	45.3 ± 2.1	76.6 ± 2.6
trade	3.80 ± 0.15	1.42 ± 0.10	34.7 ± 1.8	70.0 ± 2.4
interest	4.19 ± 0.26	1.67 ± 0.20	27.5 ± 4.2	63.8 ± 4.3
ship	2.21 ± 0.08	0.96 ± 0.13	18.2 ± 2.6	61.2 ± 3.9
wheat	1.79 ± 0.14	0.65 ± 0.08	43.2 ± 1.8	71.1 ± 2.3
corn	1.72 ± 0.12	0.71 ± 0.06	28.1 ± 2.1	61.8 ± 1.7

$Prec_{\xi\alpha}(h_{\mathcal{L}})$	$Prec(h_{\mathcal{L}})$	$F1_{\xi\alpha}(h_{\mathcal{L}})$	$F1(h_{\mathcal{L}})$
98.8 ± 0.1	99.1 ± 0.1	95.4 ± 0.1	96.9 ± 0.2
92.1 ± 0.4	93.6 ± 0.7	90.0 ± 0.5	93.1 ± 0.4
83.4 ± 1.4	89.8 ± 1.9	73.0 ± 1.4	80.9 ± 1.4
98.2 ± 0.8	98.5 ± 0.5	84.7 ± 1.2	89.9 ± 1.2
90.9 ± 1.0	92.7 ± 1.2	80.6 ± 1.0	83.9 ± 1.3
83.5 ± 1.3	89.0 ± 1.5	70.0 ± 1.7	78.3 ± 1.6
80.2 ± 2.9	87.2 ± 2.8	64.5 ± 4.4	73.6 ± 3.4
93.3 ± 1.9	93.8 ± 2.5	62.7 ± 5.0	74.0 ± 2.9
95.0 ± 1.5	97.8 ± 1.0	75.6 ± 1.2	82.3 ± 1.4
97.3 ± 1.8	99.1 ± 0.6	68.1 ± 3.6	76.1 ± 1.3
$Prec_{\xi\alpha}(h_{\mathcal{L}})$	$Prec(h_{\mathcal{L}})$	$F1_{\xi\alpha}(h_{\mathcal{L}})$	$F1(h_{\mathcal{L}})$
92.7 ± 0.5	99.1 ± 0.1	88.3 ± 0.4	96.9 ± 0.2
66.0 ± 1.2	93.6 ± 0.7	62.5 ± 1.2	93.1 ± 0.4
52.1 ± 1.8	89.8 ± 1.9	44.2 ± 1.8	80.9 ± 1.4
72.8 ± 1.8	98.5 ± 0.5	57.9 ± 2.0	89.9 ± 1.2
62.8 ± 2.4	92.7 ± 1.2	52.6 ± 2.1	83.9 ± 1.3
51.2 ± 2.2	89.0 ± 1.5	41.4 ± 1.9	78.3 ± 1.6
40.8 ± 5.0	87.2 ± 2.8	32.8 ± 4.6	73.6 ± 3.4
49.4 ± 5.2	93.8 ± 2.5	26.6 ± 3.5	74.0 ± 2.9
65.8 ± 2.7	97.8 ± 1.0	52.2 ± 1.9	82.3 ± 1.4
57.1 ± 3.2	99.1 ± 0.6	37.6 ± 2.2	76.1 ± 1.3

ΣΧΗΜΑ 4.12.2: Πίνακας που συγκρίνει τις μέσες ξα-εκτιμήσεις με τις μέσες πραγματικές αποδόσεις για τις 10 πιο συχνές κατηγορίες Reuters κατηγορίες. Το πάνω μισό δείχνει τις εκτιμήσεις για ρ=1, ενώ το κάτω μισό για ρ=2. Οι "πραγματικές" αξίες εκτιμώνται από ένα hold-out σύνολο του ιδίου μεγέθους με το σύνολο εκπαίδευσης. Για όλες τις αξίες

βρίσκονται οι μέσοι όροι για 10 τυχαίους διαχωρισμούς εκπαίδευσης/δοκιμής εκθέτοντας τις τυπικές αποκλίσεις μετά από κάθε μέσο.

4.13 ΜΕΘΟΔΟΣ ΕΛΑΧΙΣΤΟΠΟΙΗΣΗΣ ΕΓΓΥΗΜΕΝΟΥ ΚΙΝΔΥΝΟΥ (GUARANTEED RISK MINIMIZATION)

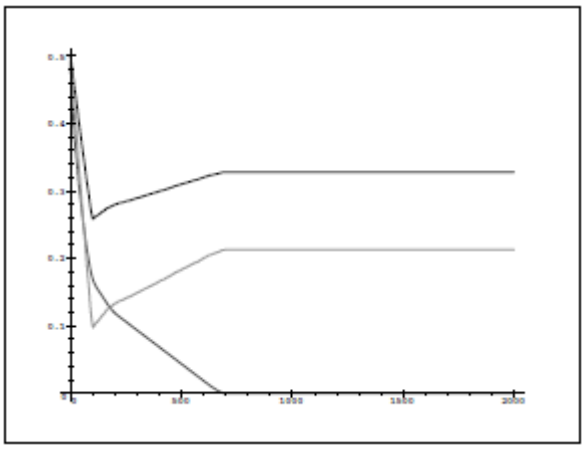
4.13.1 ΕΙΣΑΓΩΓΗ

Στο πρόβλημα επιλογής μοντέλου πρέπει να εξισορροπηθεί η πολυπλοκότητα του στατιστικού μοντέλου με την καλή προσαρμογή του στα δεδομένα εκπαίδευσης. Ο στόχος είναι η ελαχιστοποίηση του προκύπτοντος σφάλματος γενίκευσης. Στη συγκεκριμένη ενότητα θα γίνει παρουσίαση της μεθόδου GRM που ανήκει στην κατηγορία των penalty-based μεθόδων και κάποιων πειραματικών αποτελεσμάτων.

4.13.2 ΟΡΙΣΜΟΙ

Θα παρουσιαστεί μια λίστα ποικίλων μέτρων σφάλματος, τα οποία στη συνέχεια θα χρησιμοποιηθούν εκτενώς:

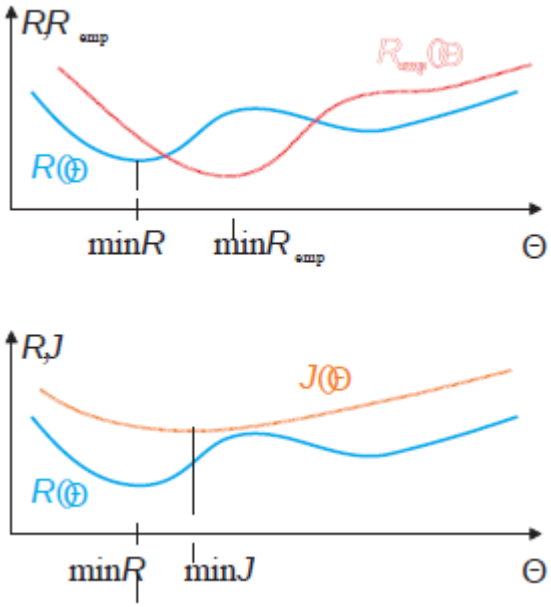
1. Το ε(h) δηλώνει το σφάλμα γενίκευσης μιας υπόθεσης h σύμφωνα με τη συνάρτηση στόχου f και την κατανομή D. Δηλαδή:
$$\varepsilon(h) \stackrel{\text{def}}{=} Pr_{x\in D}[h(x) \neq f(x)].$$
Ομοίως για το ποσοστό θορύβου n>0:
$$\varepsilon^n(h) \stackrel{\text{def}}{=} Pr_{x\in D,c}[h(x) \neq f(x) \oplus c]$$
όπου c, είναι ένα δυαδικό ψηφίο, το οποίο είναι 1 με πιθανότητα n και 0 με πιθανότητα 1-n.
2. Το $\hat{\varepsilon}(h)$ είναι το σφάλμα εκπαίδευσης του h στο δείγμα S. Δηλαδή:
$\hat{\varepsilon}(h) = \hat{\varepsilon}_S(h) \stackrel{\text{def}}{=} |\{\langle x_i, b_i \in S: h(x_i \neq b_i)\rangle\}|/m$,
όπου m είναι το μέγεθος του δείγματος S.
3. Το $\varepsilon_{opt}(d)$ είναι το ελάχιστο σφάλμα γενίκευσης που λαμβάνεται υπό όλες τις υποθέσεις στο F_d.Δηλαδή:
$$\varepsilon_{opt}(d) \stackrel{\text{def}}{=} \varepsilon(h_d)$$
όπου
$$h_d \stackrel{\text{def}}{=} arg \min_{h\in F_d}\{\varepsilon(h)\}$$
και
$$\varepsilon^n_{opt}(d) \stackrel{\text{def}}{=} \varepsilon^n(h_d)$$
4. Το $\varepsilon(d) = \varepsilon_{L,S}(d) \stackrel{\text{def}}{=} \varepsilon(\widetilde{h_d})$
και
$$\hat{\varepsilon}(d) = \hat{\varepsilon}_{L,S}(d) \stackrel{\text{def}}{=} \hat{\varepsilon}(\widetilde{h_d})$$
όπου, τα $\widetilde{h_d} = L(S,d)$ και $\varepsilon^n(d)$, ορίζονται ανάλογα.



ΣΧΗΜΑ 4.13.2: Πειραματικά γραφήματα του σφάλματος γενίκευσης ε(d) (κατώτερη καμπύλη με τοπικό ελάχιστο), $\varepsilon^n(d)$ (ανώτερη καμπύλη με τοπικό ελάχιστο) και $\hat{\varepsilon}(d)$ (μονότονα φθίνουσα καμπύλη) σε αντιδιαστολή με την πολυπλοκότητα d για συνάρτηση στόχου 100 εναλλασσόμενων διαστημάτων, μεγέθους δείγματος 2000 και ποσοστού θορύβου n=0,2. Κάθε στοιχείο δεδομένων αναπαριστά ένα μέσο 10 προσπαθειών

4.13.3 Η ΜΕΘΟΔΟΣ ΔΙΑΡΘΡΩΤΙΚΗΣ ΕΛΑΧΙΣΤΟΠΟΙΗΣΗΣ ΚΙΝΔΥΝΟΥ (STRUCTURAL RISK MINIMIZATION) ΚΑΙ ΑΝΩ ΦΡΑΓΜΑ ΕΝΟΣ ΠΡΑΓΜΑΤΙΚΟΥ ΚΙΝΔΥΝΟΥ (TRUE RISK)

Η γενική ιδέα είναι η προσθήκη ενός ρυθμιστή (regularizer), ο οποίος προτιμά μια απλούστερη στρατηγική. Η θεωρία πληροφορίας των Vapnik-Chervonenkis εισάγει την έννοια του guaranteed risk $J(\theta)$, με πιθανολογική εμπιστοσύνη n, $0 \leq n \leq 1$. Το άνω φράγμα μπορεί να είναι τόσο μεγάλο με την απαισιόδοξη έννοια που μπορεί να είναι άχρηστο όπως φαίνεται και στα δύο γραφήματα του σχήματος (4.13.1)



ΣΧΗΜΑ 4.13.1

$$R(\theta) < J(\theta) = Remp(\theta) + \sqrt{\frac{h\left(\log\left(\frac{2L}{h}\right)\right) - \log\left(\frac{n}{4}\right)}{L}}$$

(4.13.3.1)

όπου L είναι το μήκος του πολλαπλού συνόλου εκπαίδευσης, h η VC-διάσταση της κλάσης στρατηγικών q(x, θ). Να σημειωθεί ότι το παραπάνω άνω φράγμα είναι ανεξάρτητο της πραγματικής $p(x, y)$.

4.13.4 ΔΙΑΣΤΑΣΗ VAPNIK- CHERVONENKIS

Η διάσταση V-C είναι ένας αριθμός που χαρακτηρίζει τη στρατηγική απόφασης, μέτρο της χωρητικότητας ενός αλγορίθμου στατιστικής ταξινόμησης (προσδιορίζει την πολυπλοκότητά του) και μια από τις κεντρικές έννοιες της θεωρίας πληροφορίας των Vapnik-Chervonenkis.

Ορισμός: Ας θεωρήσουμε ένα σύνολο διχοτομικών στρατηγικών $q(x, \theta) \in Q$. Το σύνολο που αποτελείται από h στοιχεία δεδομένων, μπορεί να επισημανθεί με 2^h δυνατούς τρόπους. Υπάρχει μια στρατηγική $q \in Q$ που αναθέτει ετικέτες σωστά σε όλους τους πιθανούς σχηματισμούς. Η διαδικασία εύρεσης όλων των πιθανών σχηματισμών με σωστή ανάθεση ετικετών λέγεται θραύση (shattering). Η V-C διάσταση είναι ο μέγιστος αριθμός h, των παρατηρήσεων που μπορούν να θρυμματιστούν. Η V-C διάσταση για μια γραμμική στρατηγική σε ένα n-διάστατο χώρο, είναι μια γενίκευση σε n διαστάσεις, για γραμμικούς ταξινομητές. Συγκεκριμένα, ένα υπερεπίπεδο σε χώρο $\mathbb{R}^n$, σπάει κάθε σύνολο h=n+1 γραμμικά ανεξάρτητων σημείων, συνεπώς η V-C διάσταση στρατηγικών γραμμικής απόφασης είναι h=n+1.

4.13.5 ΔΙΑΡΘΡΩΤΙΚΗ ΕΛΑΧΙΣΤΟΠΟΙΗΣΗ ΚΙΝΔΥΝΟΥ (STRUCTURAL RISK MINIMIZATION)

Η SRM ελαχιστοποιεί το εγγυημένο ρίσκο και το άνω φράγμα $J(\theta)$ (4.13.3.1) . Συγκεκριμένα για κάθε μοντέλο i στη λίστα των υποθέσεων :

(i) Υπολογίζουμε τη V-C διάσταση h_i, (ii) $\theta_i^* = arg\min_{\theta_i} Remp(\theta_i)$, (iii) Υπολογίζουμε τα $J_i(\theta_i^*, h_i)$ και τέλος επιλέγουμε το μοντέλο με τη χαμηλότερη τιμή $J_i(\theta_i^*, h_i)$. Κατά προτίμηση βελτιστοποιούμε απευθείας το (θ^*, h^*)=$arg\min_{\theta,h} J(\theta, h)$.

4.13.6 ΕΛΑΧΙΣΤΟΠΟΙΗΣΗ ΕΜΠΕΙΡΙΚΟΥ ΚΙΝΔΥΝΟΥ (EMPIRICAL RISK MINIMIZATION)

Η θεωρία που υποδεικνύει τη μάθηση με προοδευτικά πιο πολύπλοκες κλάσεις των στρατηγικών απόφασης Q και δικαιολογεί τη χρήση της ERM στις κλασεις συναρτήσεων με λογική V-C διάσταση. Το ERM είναι υπολογιστικά δύσκολο. Οι περισσότερες κλάσεις συναρτήσεων απόφασης Q, για τις οποίες το ERM (τουλάχιστον τοπικά) μπορεί να είναι επαρκώς οργανωμένο, μπορεί να είναι χρήσιμες.

4.13.7 ΑΛΓΟΡΙΘΜΟΣ ΕΠΙΛΟΓΗΣ ΜΟΝΤΕΛΟΥ GRM

Ο αλγόριθμος GRM ανήκει σε μια γενική κατηγορία στην οποία ανήκουν οι αλγόριθμοι βασισμένοι στην ποινή. Η γενική ιδέα σε αυτούς τους αλγορίθμους είναι η προσπάθεια κατασκευής μιας προσέγγισης του ε(d) μόνο όσον αφορά το σφάλμα εκπαίδευσης και την πολυπλοκότητα του d, συχνά με την προσπάθεια διόρθωσης του $\hat{\varepsilon}(d)$, κατά το ποσό που υποτιμά το ε(d), μέσω της προσθήκης ενός όρου ''ποινής πολυπλοκότητας''. Στο GRM, το $\tilde{d}$, επιλέγεται σύμφωνα με τον κανόνα:

$$\tilde{d} = arg\min_d \left\{\hat{\varepsilon}(d) + \left(\frac{d}{m}\right)\left(1 + \sqrt{1 + \hat{\varepsilon}(d)\, m/d}\right)\right\}$$
(4.13.7.1)

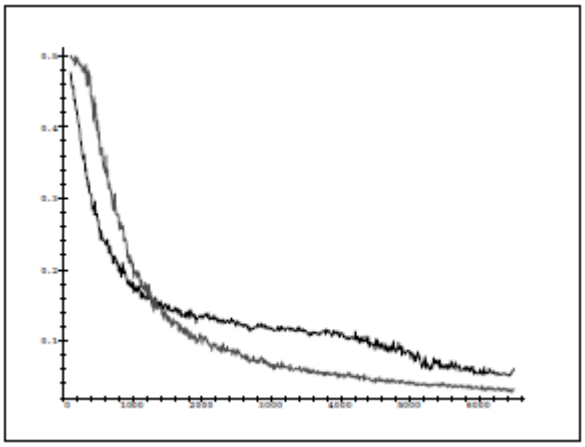
όπου έχουμε υποθέσει ότι η d είναι η V-C διάσταση της κλάσης F_d. Το γνήσιο GRM του Vapnik στην πραγματικότητα πολλαπλασιάζει το δεύτερο όρο μέσα στο argmin $\{\cdot\}$, με ένα λογαριθμικό παράγοντα που προορίζεται για την προφύλαξη από τις επιλογές χειρότερων καταστάσεων και έτσι έχει την εξής μορφή:

$$arg\min_d \left\{\hat{\varepsilon}(d) + \frac{d\left(\frac{\ln 2m}{d}+1\right)}{m}\left(\sqrt{1 + \frac{\hat{\varepsilon}(d)m}{d\left(\frac{\ln 2m}{d}+1\right)}}\right)\right\} \qquad (4.13.7.2)$$

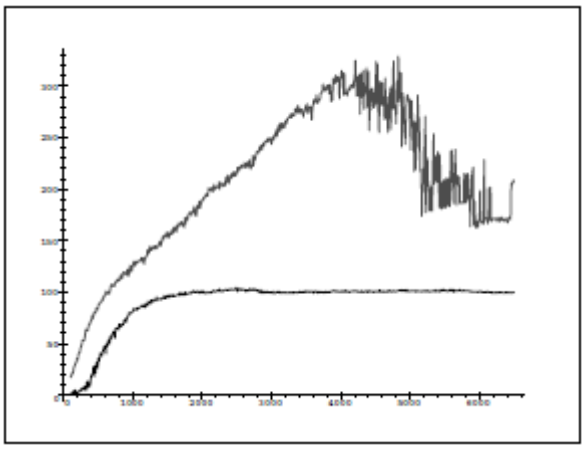
Εντούτοις, έχουμε βρει ότι οι λογαριθμικοί παράγοντες καθιστούν το GRM μη ανταγωνιστικό και έτσι προτιμάται ο τροποποιημένος και πιο ανταγωνιστικός κανόνας της εξίσωσης (4.13.7.1) της οποίας το πνεύμα είναι το ίδιο.

4.13.8 ΠΕΙΡΑΜΑΤΑ

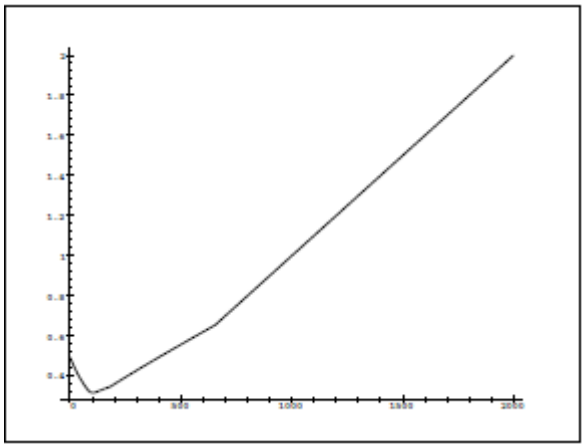
Δίνονται τα γραφήματα της συνολικής GRM ποινής για τα τρία ίδια μεγέθη δείγματος και ποσοστά σφάλματος. Η συμπεριφορά του είναι αρκετά ελεγχόμενη. Για κάθε μέγεθος δείγματος, η συνολική ποινή έχει το ίδιο σχήμα (λεκάνης) με ένα και μόνο ελάχιστο, με το ελάχιστο να ξεκινά από τα αριστερά του d=100 (το ελάχιστο να πετυχαίνεται για περίπου d=40 και m=500) και σταδιακά υποχωρεί για d=100 και οξύνεται για μεγάλα m. Στη συνέχεια εξετάζουμε μια τροποποιημένη εκδοχή του GRM, που προκύπτει πολλαπλασιάζοντας τον όρο πολυπλοκότητας της ποινής στον κανόνα GRM (το δεύτερο όρο μέσα στο argmin $\{\cdot\}$ της εξίσωσης (4.13.7.1)) με μια σταθερά μικρότερη του ένα. Στη συνέχεια συγκρίνεται η απόδοση του πραγματικού GRM με αυτή της τροποποιημένης εκδοχής στο οποίο η ποινή πολυπλοκότητας πολλαπλασιάζεται με 0,5. Όμως, παρατηρείται ότι ενώ η τροποποιημένη εκδοχή κωδικοποιεί πιο γρήγορα και έτσι μειώνει το μικρό m σφάλμα γενίκευσης, παράλληλα γίνεται υπερεκτίμηση με ένα αντίστοιχο εκφυλισμό του σφάλματος γενίκευσης. Το GRM δείχνει μεροληψία είτε υπέρ είτε κατά της κωδικοποίησης που υπερνικάται από τις εγγενείς ιδιότητες των στοιχείων ασυμπτωτικά, αλλά που μπορεί να έχει επίσης μεγάλη επίδραση για μικρά και μεσαία μεγέθη δείγματος. Αυτή η μεροληψία δε μπορεί να υπερνικηθεί απλά από την προσαρμογή των σχετικών βαρών των ποινών σφάλματος και πολυπλοκότητας, χωρίς την αντιστροφή της μεροληψίας του προκύπτοντος κανόνα και χωρίς να υποφέρει αυξημένο σφάλμα γενίκευσης για κάποιο εύρος του m.



ΣΧΗΜΑ 4.13.3: Πειραματικά γραφήματα του σφάλματος γενίκευσης $\varepsilon_{GRM}(m)$, με τη χρήση πολλαπλασιαστών πολυπλοκότητας ποινής 1.0 (χαμηλή αρχική μείωση) και 0.5 (γρήγορη αρχική μείωση) στον όρο πολυπλοκότητας ποινής $(d/m)\left(1+\sqrt{1+\hat{\varepsilon}(d)\,m/d}\right)$ σε αντιδιαστολή με το μέγεθος δείγματος m σε ένα στόχο 100 εναλλασσόμενων διαστημάτων και ποσοστό σφάλματος n=0.2. Κάθε στοιχείο δεδομένων αναπαριστά ένα μέσο 10 δοκιμών.



ΣΧΗΜΑ 4.13.4: Πειραματικά γραφήματα του μήκους υπόθεσης $\widetilde{d}_{GRM}(m)$, με τη χρήση πολλαπλασιαστών πολυπλοκότητας ποινής 1.0 (χαμηλή αρχική μείωση) και 0.5 (γρήγορη αρχική μείωση) στον όρο πολυπλοκότητας ποινής $(d/m)\left(1+\sqrt{1+\hat{\varepsilon}(d)\,m/d}\right)$ σε αντιδιαστολή με το μέγεθος δείγματος m σε ένα σε ένα στόχο 100 εναλλασσόμενων διαστημάτων και ποσοστό σφάλματος n=0.2. Κάθε στοιχείο δεδομένων αναπαριστά ένα μέσο 10 δοκιμών.



ΣΧΗΜΑ 4.13.5: GRM ολικής ποινής $\hat{\varepsilon}(d) + (d/m)\left(1 + \sqrt{1 + \hat{\varepsilon}(d)\, m/d}\right)$ σε αντιδιαστολή με την πολυπλοκότητα d για ένα και μοναδικό τρέξιμο για 2000 παραδείγματα για μια συνάρτηση στόχου 100 εναλλασσόμενων διαστημάτων και ποσοστού σφάλματος 0,2.

4.14 ΜΗΚΟΣ ΕΛΑΧΙΣΤΗΣ ΠΕΡΙΓΡΑΦΗΣ (MINIMUM DESCRIPTION LENGTH)

Η προσέγγιση του μήκους ελαχίστης περιγραφής (MDL) δίνει ένα κριτήριο επιλογής που έχει ως κίνητρο τη βέλτιστη κωδικοποίηση. Γίνεται πρώτα μια ανασκόπηση της θεωρίας κωδικοποίησης για τα δεδομένα συμπίεσης και έπειτα μια εφαρμογή της στην επιλογή μοντέλου. Θεωρούμε το δεδομένο μας z ως ένα μήνυμα που θέλουμε να κωδικοποιήσουμε και να στείλουμε σε κάποιον άλλο (''το δέκτη''). Θεωρούμε το μοντέλο ως ένα τρόπο να κωδικοποιήσουμε το δεδομένο και να επιλέξουμε το πιο φειδωλό μοντέλο, που είναι ο συντομότερος κώδικας, για τη μετάδοση.

Ας υποθέσουμε πρώτα ότι τα πιθανά μηνύματα, τα οποία μπορεί να θέλουμε να μεταδώσουμε είναι: $z_1, \dots, z_m$. Ο κώδικάς μας χρησιμοποιεί ένα πεπερασμένο αλφάβητο μήκους A. Θα περιοριστούμε στους στιγμιαίους κώδικες προθέματος, δηλαδή στους κώδικες στους οποίους κανένας δεν είναι το πρόθεμα κανενός άλλου και ο δέκτης (ο οποίος γνωρίζει όλους τους πιθανούς κώδικες) γνωρίζει ακριβώς πότε το μήνυμα έχει ολοκληρωτικά σταλεί. Αποφασίζουμε ποιον κώδικα να χρησιμοποιήσουμε ανάλογα με το πόσο συχνά στέλνουμε κάθε μήνυμα. Εν γένει χρησιμοποιούμε την εξής στρατηγική, πιο σύντομοι κώδικες, για πιο συχνά μηνύματα , έτσι ώστε το μέσο μήκος μηνύματος να είναι μικρότερο.

Εν γένει αν τα μηνύματα στέλνονται με πιθανότητες $Pr(z_i)$ $i = 1,2, \dots ,4$, σύμφωνα με το θεώρημα του Shannon, θα μπορούσαμε να χρησιμοποιούσαμε μήκη κωδίκων $l_i = -\log_2 Pr(z_i)$ και το μέσο μήκος μηνύματος να ικανοποιεί την:

$$E(length) \geq \sum Pr(z_i) \log_2 Pr(z_i) \qquad (4.14.1)$$

Το δεξί μέλος της (4.14.1), επίσης ονομάζεται εντροπία της κατανομής $Pr(z_i)$. Η ανισότητα είναι ισότητα όταν οι πιθανότητες ικανοποιούν την $p_i = A^{-l_i}$. Γενικά το κατώτερο όριο δεν μπορεί να επιτευχθεί, με εξαίρεση διαδικασίες όπως το σύστημα κωδικοποίησης του Huffman. Να σημειωθεί ότι με ένα άπειρο σύνολο μηνυμάτων, η εντροπία αντικαθίσταται από $-\int Pr(z) \log_2 Pr(z) dz$. Από αυτό το αποτέλεσμα συμπεραίνουμε το εξής:

Για να μεταδώσουμε μια τυχαία μεταβλητή z,που έχει συνάρτηση πυκνότητας πιθανότητας $Pr(z)$, απαιτούμε περίπου $-\log_2 Pr(z)$ bits πιθανότητας. Έχουμε πλέον αλλάξει συμβολισμό από $\log_2 Pr(z)$ σε $\log Pr(z)$=$\log_e Pr(z)$, αυτό γίνεται για ευκολία και απλά εισάγει μια ασήμαντη πολλαπλασιαστική σταθερά.

Τώρα εφαρμόζουμε αυτό το αποτέλεσμα στο πρόβλημα επιλογής μοντέλου. Έχουμε ένα μοντέλο M με παραμέτρους θ και δεδομένα z=(X, y) που αποτελούνται από εισόδους και εξόδους. Έστω (υπό όρους) η πιθανότητα των εξόδων σύμφωνα με το μοντέλο ότι είναι $Pr(y|\theta, M, X)$ και ας υποθέσουμε ότι ο δέκτης γνωρίζει όλες τις εισόδους και θέλουμε να μεταδώσει τις εξόδους. Τότε το μήκος μηνύματος που απαιτείται για να μεταδοθούν οι έξοδοι είναι:

$$length = -\log Pr(y|\theta, M, X) - \log Pr(\theta|M) \qquad (4.14.2)$$

Η λογαριθμική πιθανότητα (log-probability) των τιμών-στόχων, δεδομένων των εισόδων. Ο δεύτερος όρος είναι το μέσο μήκος κώδικα για τη μετάδοση των παραμέτρων θ του μοντέλου, ενώ ο πρώτος είναι το μέσο μήκος κώδικα για τη μετάδοση της διαφοράς μεταξύ του μοντέλου και των πραγματικών τιμών-στόχων.

Σύμφωνα με την αρχή του MDL πρέπει να επιλέγουμε το μοντέλο που ελαχιστοποιεί τη (4.14.2). Αναγνωρίζουμε τη (4.14.2) ως την αρνητική λογαριθμική εκ των υστέρων κατανομή και ως εκ τούτου η ελαχιστοποίηση του μήκους περιγραφής είναι ισοδύναμη με τη μεγιστοποίηση της εκ των υστέρων πιθανότητας.

Να σημειωθεί ότι έχουμε αγνοήσει την ακρίβεια με την οποία μια τυχαία μεταβλητή Z κωδικοποιείται. Με ένα πεπερασμένο μήκος κώδικα δε μπορούμε να κωδικοποιήσουμε ακριβώς μια συνεχή μεταβλητή. Εντούτοις, αν κωδικοποιήσουμε τη z με κάποια ανοχή δ_z , το μήκος μηνύματος είναι ο λογάριθμος της πιθανότητας στο διάστημα $[z, z + \delta_z]$, το οποίο προσεγγίζεται καλά από το $\delta_z Pr(z)$ αν το δ_z είναι μικρό. Από τη στιγμή που $\log \delta_z Pr(z)$=$\log \delta_z + \log Pr(z)$, αυτό σημαίνει ότι μπορούμε απλά να αγνοήσουμε τη σταθερά $\log \delta_z$ και να χρησιμοποιήσουμε τη $\log Pr(z)$ ως το μέτρο του μήκους μηνύματος. Η προηγούμενη άποψη του MDL για την επιλογή μοντέλου υποστηρίζει ότι πρέπει να επιλέξουμε το μοντέλο με την υψηλότερη εκ των υστέρων πιθανότητα. Διαφορετικά θα μπορούσαμε να βγάλουμε συμπεράσματα από τη δειγματοληψία της εκ των υστέρων κατανομής.

4.15 ΜΕΘΟΔΟΙ BOOTSTRAP

Η μέθοδος bootstrap είναι ένα γενικό εργαλείο για την αξιολόγηση της στατιστικής ακρίβειας. Πρώτα περιγράφουμε τη bootstrap γενικά και έπειτα δείχνουμε πως μπορεί να χρησιμοποιηθεί για να εκτιμήσουμε το σφάλμα πρόβλεψης εκτός δείγματος. Η bootstrap επιδιώκει να εκτιμήσει το κατά συνθήκη σφάλμα Err_T, αλλά συνήθως εκτιμά καλά μόνο την αναμενόμενη πρόβλεψη σφάλματος Err.

Ας υποθέσουμε ότι έχουμε ένα μοντέλο που προσαρμόζεται σε ένα σύνολο δεδομένων εκπαίδευσης Z=$(z_1, \dots, z_N)$, όπου $z_i = (x_i, y_i)$. Η βασική ιδέα είναι να επιλέξουμε τυχαία σύνολα δεδομένων με αντικατάσταση από τα δεδομένα εκπαίδευσης, κάθε δείγμα του ιδίου μεγέθους όπως το αρχικό σύνολο εκπαίδευσης. Αυτό γίνεται B φορές, παράγοντας B bootstrap σύνολα δεδομένων. Έπειτα επαναπροσαρμόζουμε το μοντέλο σε κάθε ένα από τα bootstrap σύνολα δεδομένων και εξετάζουμε τη συμπεριφορά στις B επαναλήψεις. Μια προσέγγιση για την εφαρμογή της bootstrap έτσι ώστε να εκτιμήσουμε το σφάλμα πρόβλεψης θα μπορούσε να είναι η προσαρμογή του εν λόγω μοντέλου σε ένα σύνολο

bootstrap δειγμάτων, και στη συνέχεια να παρακολουθήσουμε πόσο καλά προβλέπει το αρχικό σύνολο εκπαίδευσης. Εάν η $\hat{f}^{*b}(x_i)$, είναι η προβλεπόμενη τιμή της x_i, από το μοντέλο που είναι προσαρμοσμένο στο b-οστό σύνολο δεδομένων bootstrap, η εκτίμησή μας είναι:

$$\widehat{Err}_{boot} = \frac{1}{B}\frac{1}{N}\sum_{b=1}^{B}\sum_{i=1}^{N} L\left(y_i, \hat{f}^{*b}(x_i)\right) \quad (4.15.1)$$

Εντούτοις είναι εύκολο να διαπιστώσουμε ότι το Err_{boot}, δεν παρέχει μια καλή εκτίμηση εν γένει. Ο λόγος είναι ότι τα σύνολα δεδομένων bootstrap δρουν όπως και τα δείγματα εκπαίδευσης, ενώ το αρχικό σύνολο εκπαίδευσης δρα όπως το δείγμα δοκιμής και αυτά τα δύο δείγματα έχουν κοινές παρατηρήσεις. Αυτή η επικάλυψη μπορεί να δημιουργήσει τις υπερπροσαρμοσμένες προβλέψεις να φαίνονται εξωπραγματικά καλές και για αυτό το λόγο η διασταυρωμένη επικύρωση χρησιμοποιεί ρητά μη επικαλυπτόμενα δεδομένα για τα δείγματα εκπαίδευσης και δοκιμής.

Μιμούμενο τη διασταυρωμένη επικύρωση, μια καλύτερη εκτίμηση bootstrap μπορεί να επιτευχθεί. Για κάθε παρατήρηση παρακολουθούμε μόνο τις προβλέψεις από τα bootstrap δείγματα τα οποία δεν περιλαμβάνουν αυτή την παρατήρηση. Η leave-one-out bootstrap εκτίμηση του σφάλματος πρόβλεψης ορίζεται από:

$$\widehat{Err}^{(1)} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|C^{-i}|}\sum_{b\in C^{-i}} L\left(y_i, \hat{f}^{*b}(x_i)\right) \quad (4.15.2)$$

Εδώ το C^{-i} , είναι ένα σύνολο δεικτών των bootstrap δειγμάτων b που δεν περιέχουν την παρατήρηση i και $|C^{-i}|$ είναι ο αριθμός των εν λόγω δειγμάτων. Στον υπολογισμό του $\widehat{Err}^{(1)}$, είτε θα πρέπει να επιλέξουμε το B να είναι αρκετά μεγάλο για να εξασφαλίσουμε ότι όλα τα $|C^{-i}|$ είναι μεγαλύτερα του 0 , είτε θα πρέπει να αφήσουμε τους όρους στην (4.15.2) που αντιστοιχούν στο $|C^{-i}|$ να είναι 0.

Ορίζουμε το γ να είναι το χωρίς-πληροφορία ποσοστό σφάλματος: αυτό είναι το ποσοστό σφάλματος του κανόνα πρόβλεψής μας, εάν οι είσοδοι και οι ετικέτες κλάσης ήταν ανεξάρτητες. Μια εκτίμηση της γ αποκτάται αξιολογώντας τον κανόνα πρόβλεψης σε όλους τους πιθανούς συνδυασμούς των στόχων y_i και των predictors x_i' ως

$$\hat{\gamma} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{i'=1}^{N} L\left(y_i, \hat{f}(x_{i'})\right). \quad (4.15.3)$$

Ορίζουμε το σχετικό υπερπροσαρμοσμένο ποσοστό να είναι ίσο με

$$\hat{R} = \frac{\widehat{Err}^{(1)} - \overline{err}}{\hat{\gamma} - \overline{err}} \quad (4.15.4)$$

Μια ποσότητα που κυμαίνεται από 0 , αν δεν υπάρχει υπερπροσαρμογή ($\widehat{Err}^{(1)} = \overline{err}$) μέχρι 1, εάν η υπερπροσαρμογή ισούται με τη χωρίς πληροφορία αξία $\hat{\gamma} - \overline{err}$.

5

ΝΕΑ ΚΡΙΤΗΡΙΑ ΠΛΗΡΟΦΟΡΙΑΣ

5.1 ΚΡΙΤΗΡΙΑ ΠΛΗΡΟΦΟΡΙΑΣ ΓΙΑ ΜΗΧΑΝΕΣ ΠΥΡΗΝΑ

5.1.1 ΕΙΣΑΓΩΓΗ

Τα τελευταία χρόνια έχουν διεξαχθεί πολλές έρευνες πάνω στις μηχανές διανυσματικής υποστήριξης (SVMs) και στη λογιστική παλινδρόμηση πυρήνα (KLR). Συγκεκριμένα η επιλογή των παραμέτρων και του πυρήνα αποτελεί ένα πολύ σημαντικό ζήτημα, διότι έχει πολύ σοβαρές επιπτώσεις στις επιδόσεις των ταξινομήσεων που προκύπτουν από την KLR και τα SVMs.Τόσο η KLR όσο και τα SVMs έχουν δύο είδη παραμέτρων να συντονίσουν, την ποινή εσφαλμένης ταξινόμησης (συχνά συμβολίζεται με C) και τις παραμέτρους που προσδιορίζουν τη συνάρτηση του πυρήνα.

Σε αυτή την ενότητα, θα ασχοληθούμε κυρίως με τη βελτιστοποίηση της ποινής εσφαλμένης ταξινόμησης C. Η παράμετρος C στα SVMs και στην KLR αναγνωρίζεται ως η παράμετρος συστηματοποίησης στην αντίστοιχη στατιστική κανονικοποίηση μοντέλων και η τιμή της καθορίζει την ανταλλαγή μεταξύ του σφάλματος εκπαίδευσης και της πολυπλοκότητας του μοντέλου.

Σε αυτή την ενότητα, δύο διαφορετικά μοντέλα συστηματοποίησης καθορίζονται ως τα στατιστικά μοντέλα για τα SVMs. Αυτά τα μοντέλα αντιστοιχούν στα Μπεϋζιανά μοντέλα, τα λογιστικά Μπεϋζιανα μοντέλα και το μοντέλο του Sollich. Στα μοντέλα συστηματοποίησης, η παράμετρος συστηματοποίησης αντιστοιχεί στην υπερπαράμετρο (hyperparameter) των Μπεϋζιανών μοντέλων. Στη συνέχεια εισάγουμε ένα κριτήριο, το κριτήριο πληροφορίας συστηματοποίησης πυρήνα (KRIC) για τη ρύθμιση της παραμέτρου συστηματοποίησης .

Το KRIC αντιστοιχεί στο κριτήριο πληροφορίας συστηματοποίησης (RIC) στο χώρο των χαρακτηριστικών. Είναι γνωστό ότι το RIC είναι ένα αποτελεσματικό κριτήριο για τη ρύθμιση παραμέτρων στη συστηματοποίηση των μοντέλων. Η κύρια ιδέα του RIC είναι να ελαχιστοποιήσει την προσδοκία της Kullback-Leibler απόκλισης από την πραγματική

συνάρτηση κατανομής στην εκτιμώμενη συνάρτηση κατανομής με το μοντέλο παλινδρόμησης. Το υπολογιστικό κόστος της επιλογής μοντέλου από το RIC είναι συχνά χαμηλότερο από ότι άλλων μεθόδων επειδή δε χρειάζεται καμιά αναδειγματοληψία δεδομένων ή αναδρομικές βελτιστοποιήσεις. Στη συνέχεια θα γίνει μείωση των υπολογισμών του RIC για τα SVMs σε ένα πρόβλημα ιδιοτιμών και εισαγωγή του KRIC.Το KRIC μπορεί να χρησιμοποιηθεί για τη ρύθμιση των παραμέτρων συστηματοποίησης στη λογιστική παλινδρόμηση πυρήνα (KLR). Προτείνεται να ρυθμιστούν οι παράμετροι συστηματοποίησης για την KLR με τη χρήση του KRIC. Το αποτέλεσμα μπορεί να χρησιμοποιηθεί απευθείας στα SVMs, καθώς τα Μπεϋζιανα μοντέλα για την KLR είναι παρόμοια με τα λογιστικά Μπεϋζιανα μοντέλα για τα SVMs .

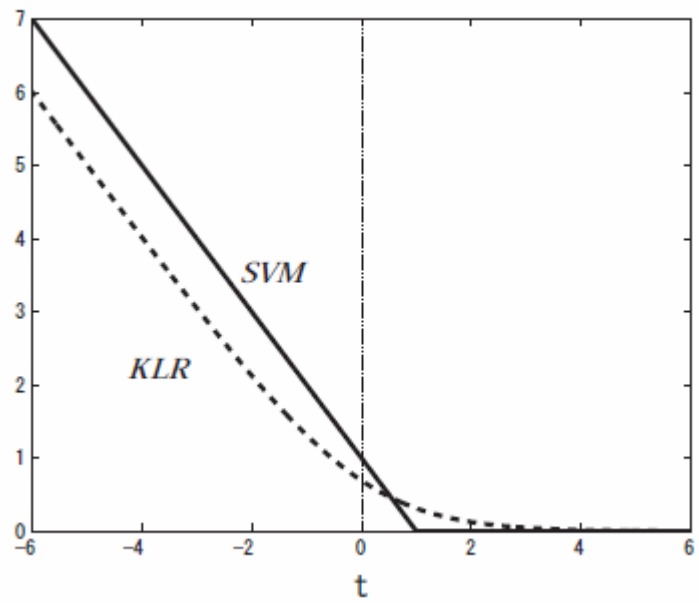
5.1.2 Η ΛΟΓΙΣΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ ΠΥΡΗΝΑ (KLR)

Η KLR λύνει τα ακόλουθα προβλήματα βελτιστοποίησης :

- **Το πρόβλημα βελτιστοποίησης της KLR :**

$$\min \frac{1}{2}\|w\|^2 \quad + c\sum_{i=1}^{l} \xi_i$$

υπό τους περιορισμούς $\xi_i = g(y_i(w \times zi - b)\,)$ για i=1,…,l



ΣΧΗΜΑ 5.1.1 : Σύγκριση των συναρτήσεων απώλειας των SVMs και της KLR

όπου $g(t) = \log(1 + e^{-t})$

Η KLR χρησιμοποιεί τη συνάρτηση απώλειας $g(t)$ αντί της συνάρτησης απώλειας $[1 - t]_+$. Το ΣΧΗΜΑ 1 δείχνει τη διαφορά των δύο συναρτήσεων απώλειας.

Έστω $G(\delta) = \delta \log(\delta) + (1 - \delta)\log(1 - \delta)$. Το δυϊκό πρόβλημα του Wolfe για την KLR παρουσιάζεται παρακάτω.

- **Το δυϊκό πρόβλημα για την KLR**

$$\max -c\sum_{i=1}^{l} G\left(\frac{a_i}{C}\right) - \sum_{i,j=1}^{l} a_i a_j y_i y_j\, K(x_i x_j)$$

υπό τους περιορισμούς : $\sum_{i=1}^{l} a_i y_i = 0$

5.1.3 ΤΟ ΛΟΓΙΣΤΙΚΟ ΜΠΕΥΖΙΑΝΟ ΠΛΑΙΣΙΟ ΓΙΑ ΤΑ SVMS ΚΑΙ ΤΗΝ KLR

Υποθέτουμε ότι κάθε δείγμα είναι ανεξάρτητα και πανομοιότυπα κατανεμημένο. Επίσης υποθέτουμε ότι η p(x_i) είναι ανεξάρτητη των w και π . Θεωρούμε μια λογιστική πυκνότητα πρόβλεψης :

$$p(y_i|x_i, b, \pi, K)=\frac{1}{1+exp(-na_i y_i)} \qquad (5.1.3.1)$$

όπου η n είναι μία καθορισμένη θετική σταθερά. Στην αριστερή πλευρά το K αντιπροσωπεύει τη συνάρτηση πυρήνα και το π αντιπροσωπεύει τις παραμέτρους του πυρήνα. Παραλείπουμε το K στα ακόλουθα. Στη συνέχεια, υποθέτουμε μια κανονική εκ των προτέρων πυκνότητα :

$$p(w, b|\lambda, \pi) \propto exp\left(-\frac{\lambda}{2}\|w\|^2\right), \ (\lambda>0)$$

όπου η λ είναι μια θετική υπερπαράμετρος. Από τον κανόνα του Bayes η εκ των υστέρων κατανομή είναι η

$$p(w, b|D, \lambda, \pi) \propto p(w, b|\lambda, \pi)p(D|w, b, \pi)$$
$$= p(w, b|\lambda, \pi) \prod_i p(y_i|x_i, w, b, \pi)p(x_i)$$

Ως εκ τούτου η αρνητική λογαριθμική εκ των υστέρων κατανομή είναι:

$$\begin{aligned} -\log(w|D, \lambda, \pi) &= \frac{\lambda}{2}\|w\|^2 \\ &+ \sum_i \log\{1 + exp(-na_i y_i)\} \\ &- \sum_i \log p(x_i) + constant \ \ (\alpha\nu\varepsilon\xi\acute{\alpha}\rho\tau\eta\tau\eta \ \tau o\upsilon \ w) \end{aligned}$$

5.1.4 ΤΟ ΜΠΕΥΖΙΑΝΟ ΠΛΑΙΣΙΟ ΤΟΥ SOLLICH ΓΙΑ ΤΑ SVMS

Σε αυτή την ενότητα συνοψίζουμε το Μπεϋζιανό μοντέλο του Sollich. Συντομεύουμε το $\alpha(\cdot, w, b) \ \sigma\tau o \ a$. Όπως στη Γκαουσιανή διαδικασία, θεωρούμε μία εκ των προτέρων κατανομή της συνάρτησης α. Έστω $Q(x)$ μια πραγματική συνάρτηση πυκνότητας για κάθε δείγμα x_i. Τότε οι συναρτήσεις πιθανότητας του Μπεϋζιανού μοντέλου δίνονται από τη σχέση

$$p(x|a) = Q(x)\, v\big(a(x)\big)/N(a),$$
$$p(y|x, a) = exp(-C[1 + ya(x)]_+)\, \tau(C)/v\big(a(x)\big),$$

όπου $$N(a) := \int_x Q(x) v\big(a(x)\big)\, dx \quad , \qquad (5.1.4.1)$$

$$v\big(a(x)\big) := \tau(C)[exp\{-C[1 - a(x)]_+\} + exp\{-C[1 + a(x)]_+\}]$$

και $\tau(C) = \{1 + exp(-2C)\}^{-1}$

Η ολοκληρωσιμότητα της δεξιάς πλευράς της (5.1.4.1) υποτίθεται. Ο συντελεστής $\tau(C)$ έχει οριστεί για να διασφαλιστεί ότι $v\big(a(x)\big)$<1. Η εκ των προτέρων συνάρτηση πιθανότητας ορίζεται ως εξής :

$$\pi(\alpha) \propto \left(-\frac{1}{2}\int \alpha(x)K^{-1}(x,\acute{x})a(\acute{x})\,dx\,d\acute{x}\right)N^{l}(a) \qquad (5.1.4.2)$$

Ας υποθέσουμε ότι για το παραπάνω μοντέλο, ο λογάριθμος της εκ των υστέρων πιθανότητας είναι ίσος με

$$\log p(a|D) = \sum_{i} \log p(y_i|x_i, a) + \log p(x_i|a) + \log \pi(a) + const$$
$$= -\frac{1}{2}\int \alpha(x)K^{-1}(x,\acute{x})a(\acute{x})\,dx\,d\acute{x} - C\sum_{i=1}^{l}[1 - y_i\theta(x_i)]_+ + const$$
$$= -\frac{1}{2}\|w\|^2 + \frac{1}{2}b^2\int K^{-1}(x,\acute{x})\,dx\,d\acute{x} - C\sum_{i=1}^{l}[1 - y_i a(x_i)]_+ + const$$

Υιοθετούμε την εκ των προτέρων Γκαουσιανή διαδικασία στο α(x)+b=wφ(x) αντί για το

$$\pi(\alpha) \propto exp\left\{-\frac{1}{2}\int (a(x) + b)K^{-1}(x,\acute{x})(a(\acute{x}) + b)\,dx\,d\acute{x}\right\}N^{l}(a) \qquad (5.1.4.3)$$

Τότε η εκ των υστέρων πιθανότητα γίνεται

$$\log p(a|D) = -\frac{1}{2}\|w\|^2 - C\sum_i\,[1 - y_i a(x_i)]_+ + const.$$

Ως εκ τούτου, η μεγιστοποίηση της εκ των υστέρων πιθανότητας αντιστοιχεί στη βελτιστοποίηση στο τετραγωνικό πρόβλημα των SVMs. Στο άρθρο του Sollich προτείνεται να μεγιστοποιηθεί το

$$-\frac{1}{2}\|w\|^2 - \frac{1}{2}b^2B^{-2} - C\sum_{i=1}^{l}[1 - y_i a(x_i)]_+ - C\sum_i\,[1 - y_i a(x_i)]_+ \qquad (5.1.4.4)$$

για ένα επαρκώς επιλεγμένο B>0 .

5.1.5 ΤΟ ΚΡΙΤΗΡΙΟ ΣΥΣΤΗΜΑΤΟΠΟΙΗΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ (RIC)

Το RIC για ένα μοντέλο συστηματοποίησης

$$L(D;\theta;\lambda) = \log p(D|\theta) - \lambda k(\theta)$$

Ορίζεται από τη σχέση

$$RIC := -2[-\log p(D|\theta^*) + trace[IJ^{-1}]] \qquad (5.1.5.1)$$

όπου

$$I := E_{p_0}\left[\frac{\partial^2 L(D;\theta;\lambda)}{\partial\theta\partial\theta^T}|\theta = \theta^*\right],$$

$$\mathrm{J} := Var_{p_0}\left[\frac{\partial L(D;\theta;\lambda)}{\partial\theta}|\theta = \theta^*\right],$$

όπου p_0 είναι το άγνωστο πραγματικό μέτρο πιθανότητας του D και τo $\theta^* = \theta^*(D,\lambda)$ μεγιστοποιεί το $L(D;\theta;\lambda)$. Για τον υπολογισμό του I και του J, αντί των προσδοκιών σε σχέση με την άγνωστη πραγματική πυκνότητα p_0, συνήθως χρησιμοποιούνται οι μέσες τιμές δείγματος p_o . Αποδεικνύεται ότι

$$\int p_0(\acute{D})KL\left(p_0(D)||p\left(D\middle|\theta^*(\acute{D},\lambda)\right)\right)$$
$$= \int p_0(D)\log p_0(D) + \frac{1}{2}RIC + o(1) \qquad (5.1.5.2)$$

όπου η KL(p||q) αναπαριστά την απόκλιση Kullback-Leibler από το p στο q. Ο πρώτος όρος της δεξιάς πλευράς δεν εξαρτάται από το λ . Έτσι η ελαχιστοποίηση του RIC όσον αφορά την παράμετρο λ, ανστιστοιχεί στην ελαχιστοποίηση της προσδοκίας της Kullback-Leibler απόκλισης από το πραγματικό μέτρο πιθανότητας στο εκτιμώμενο μέτρο πιθανότητας στο μοντέλο συστηματοποίησης (5.1.5.1) με την παράμετρο λ .

5.1.6 ΤΟ ΚΡΙΤΗΡΙΟ ΣΥΣΤΗΜΑΤΟΠΟΙΗΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ ΠΥΡΗΝΑ (KRIC)

Σε αυτή την ενότητα αναγνωρίζουμε την KLR και τα SVMs ως μοντέλα συστηματοποίησης και προτείνουμε ένα καινούριο κριτήριο για τη ρύθμιση των παραμέτρων, το κριτήριο συστηματοποίησης πληροφορίας του πυρήνα (KRIC). Υποθέτουμε ότι ο πυρήνας αναπαραγωγής του χώρου Hilbert είναι πεπερασμένης διάστασης. Πρώτον θεωρούμε το λογιστικό Μπεϋζιανό μοντέλο για τα SVMs. Στο λογιστικό Μπεϋζιανό πλαίσιο, μεγιστοποιούμε το

$$L(D;w;b;\lambda;\pi) = \log(D|w;b;\pi) - \frac{\lambda}{2}\|w\|^2 \qquad (5.1.6.1)$$

Αυτό το μοντέλο μπορεί να αναγνωριστεί ως ένα μοντέλο συστηματοποίησης. Ως εκ τούτου, μπορούμε να χρησιμοποιήσουμε το RIC προκειμένου να βελτιστοποιήσουμε την παράμετρο συστηματοποίησης λ .

Αξιολογούμε την τιμή του RIC. Έστω $\widetilde{w}^T = [w^T\gamma^{-1}b]$ $\kappa\alpha\iota$ $\widetilde{z_i}^T = [z_i{}^T - \gamma]$, όπου γ μια θετική σταθερά διαιτησίας. Τότε τα I και J στο RIC αξιολογούνται ως ακολούθως.

$$\mathrm{J} = E_{p_0}\left[\frac{\partial^2}{\partial\widetilde{w}\partial\widetilde{w}^T}\left\{-\sum_i \log p(x_i)p(y_i|x_i,\widetilde{w},\pi) + \frac{\lambda}{2}\|w\|^2\right\}|\widetilde{w} = \widetilde{w}^T\right]$$

$$= \sum_i t_i\tilde{z}_i\,\tilde{z}_i{}^T\left(1 + O\left(l^{-\frac{1}{2}}\right)\right) + \lambda\tilde{I}_d \qquad (5.1.6.2)$$

όπου

$$\tilde{I}_d := \begin{bmatrix} I_d & 0 \\ 0^T & 0 \end{bmatrix} \qquad (5.1.6.3)$$

και

$$t_i := \frac{\partial^2}{\partial a_i{}^2}\log p(y_i = 1|x_i,\widetilde{w}^*,\pi) \qquad (5.1.6.4)$$

Ομοίως

$$I = Var_{p_0}\left[\frac{\partial}{\partial \widetilde{w}}\left\{-\sum_i \log p(y_i|x_i,w,\pi) + \frac{\lambda}{2}\|w\|^2\right\}|_{w=\widetilde{w}^T}\right]$$
$$= \left(\sum_i m_i{}^2\, \widetilde{z}_i\widetilde{z}_i{}^T - \frac{1}{l}\sum_i m_i\widetilde{z}_i \sum_i m_j z_j{}^T\right)\left(1 + O\left(l^{-\frac{1}{2}}\right)\right) \qquad (5.1.6.5)$$

όπου

$$m_i := \frac{\partial}{\partial a_i}\log p\left(y_i = 1\middle|x_i,\ \widetilde{w}^*,\pi\right) \qquad (5.1.6.6.)$$

Στις σχέσεις (5.1.6.2) και (5.1.6.5) χρησιμοποιείται ο νόμος των μεγάλων αριθμών. Επειδή τα I και J εξαρτώνται από το άγνωστο πραγματικό μέτρο πιθανότητας p_0 , αντικαθιστούμε τα J και I με τα παρακάτω

$$\hat{J} := \sum_I t_i\widetilde{z}_i\widetilde{z}_i{}^T + \lambda\widetilde{I}_d$$
$$\hat{I} := \sum_I m_i{}^2\widetilde{z}_i\widetilde{z}_i{}^T - \frac{1}{l}\sum_i m_i\,\widetilde{z}_i \sum_j m_j\widetilde{z}_j{}^T \qquad (5.1.6.7)$$

Για το λογιστικό Μπεϋζιανό μοντέλο των SVMs , η p($y_i|x_i$,$\widetilde{w}$,π) ορίζεται όπως η (5.1.3.1). Έτσι , τα t_i και m_i παίρνουν τη μορφή

$$t_i := n^2\frac{exp(-na_iy_i)}{\{1+exp(-na_iy_i)\}^2}$$
$$m_i := -n\frac{y_i exp(-na_iy_i)}{1+exp(-na_iy_i)} \quad , \quad \text{για i =1,…,l} \qquad (5.1.6.8)$$

Ο άμεσος υπολογισμός των πινάκων $\hat{J}$ και $\hat{I}$ χρειάζεται την αξιολόγηση του κάθε $\widetilde{z}_i\widetilde{z}_j{}^T$. Ωστόσο, είναι δύσκολο να υπολογίσουμε την τιμή του κάθε $\widetilde{z}_i\widetilde{z}_j{}^T$ διότι τα $\tilde{z}_i$ και $\widetilde{z}_J$ δεν μπορούν να υπολογιστούν αναλυτικά (καθώς κάθε t_i στη (5.1.6.8) είναι θετικά ορισμένο και ως εκ τούτου αντιστρέψιμο). Με σκοπό την επίλυση αυτής της δυσκολίας, χρησιμοποιούμε τις ιδιοτιμές του πίνακα $\hat{J}^{-1}\hat{I}$, που μπορεί να υπολογιστεί αναλυτικά.
Έστω ότι $\{\rho_k\}$ και $\{\widetilde{v_k}\}$, είναι οι ιδιοτιμές και τα αντίστοιχα κανονικοποιημένα ιδιοδιανύσματα του $\hat{J}^{-1}\hat{I}$, αντίστοιχα. Στη συνέχεια έχουμε ότι

$$\hat{J}^{-1}\hat{I}\tilde{v}_k = \rho_k\tilde{v}_\kappa$$

και (5.1.6.9)

$$\hat{I}\tilde{v}_k = \rho_k\hat{J}\tilde{v}_\kappa$$

Αντικαθιστώντας τις (5.1.6.2) και (5.1.6.5) στην (5.1.6.9) παίρνουμε

$$\sum_i \left\{m_i^2\left(\tilde{z}_i^T\tilde{v}_k\right) - \frac{1}{l}m_i\sum_j m_j\left(\tilde{z}_j^T\tilde{v}_k\right)\right\}\tilde{z}_i = \sum_i t_i\left(\tilde{z}_i{}^T\tilde{v}_k\right)\tilde{z}_i + \lambda\tilde{I}_d\tilde{v}_k.$$

Επειδή το $\tilde{\upsilon}_k$είναι κανονικοποιημένο, όταν παίρνουμε γ→0 ,η εξίσωση (5.1.6.10) γίνεται

$$\sum_i \left\{m_i^2\left(z_i^T \upsilon_k\right) - \frac{1}{l} m_i \sum_j m_j\left(z_i^T \upsilon_k\right)\right\} z_i = \sum_i t_i\left(z_i^T \upsilon_k\right) z_i + \lambda I_d \upsilon_k \qquad (5.1.6.10)$$

Ως εκ τούτου, το υ_k αναπαρίσταται ως ένας γραμμικός συνδυασμός των z_i . Θέτουμε $\upsilon_k = \sum_i \mu_{k_i} z_i$. Παίρνουμε τα εσωτερικά προϊόντα κάθε μέλους της (5.1.6.10) και ένα συγκεκριμένο z_q. Το αριστερό μέλος της εξίσωσης είναι ίσο με

$$\sum_i \mu_{k_i}\left(\sum_j m_j^2 K_{ij} K_{jq} - \frac{1}{l}\sum_j m_j K_{ji} \sum_p m_p K_{pq}\right)$$

όπου το $K_{ij} := z_i z_j^T$ και K=(K_{ij}). Το δεξί μέλος της εξίσωσης είναι ίσο με

$$\rho_k \sum_i \mu_{k_i} \sum_j t_j K_{ji} K_{iq} + \lambda K_{kq}.$$

Έτσι έχουμε

$$K\left(diag(m) - \frac{1}{l} mm^T\right) K_{\mu k} = \rho_k (K diag(t) + \lambda I_l) K_{\mu k}.$$

Έστω $\acute{\mu}_k = K_{\mu k}$. Καθώς ο $(K diag(t) + \lambda I_l)$ είναι αντιστρέψιμος[4], παίρνουμε την ακόλουθη εξίσωση ιδιοτιμών

$$(K diag(t) + \lambda I_l)^{-1}\left(K diag(m)^2 - \frac{1}{l} K mm^T\right) \acute{\mu}_\kappa = \rho_k \acute{\mu}_k \,. \qquad (5.1.6.11)$$

Λύνοντας την εξίσωση ιδιοτιμών (5.1.6.11) ,έχουμε ότι

$$trace\left(\hat{J}^{-1}\hat{I}\right) = \sum_i \rho_i$$

$$trace\left[(K diag(t) + \lambda I_l)^{-1}\left(K diag(m)^2 - \frac{1}{l} K mm^T\right)\right] \qquad (5.1.6.12)$$

Αντικαθιστώντας την (5.1.6.12) στην (5.1.6.1) και προσεγγίζοντας τον w^{*T}με τον $\left[\hat{w}^T \gamma^{-1} \hat{b}\right]$, επιλεγμένο από το SVM, παίρνουμε το κριτήριο συστηματοποίησης πληροφορίας πυρήνα (KRIC).

<u>Το KRIC για το λογιστικό Μπεϋζιανό μοντέλο για τα SVMs</u>

[4] $|K diag(t) + \lambda I_l| = |diag(t)||K + \lambda diag(t)^{-1}| > 0$. *Έτσι ο* $(K diag(t) + \lambda I_l)$ είναι αντιστρέψιμος .

$$KRIC = 2\left[\sum_i \log\{1 + exp(-na_i y_i)\} + trace\left\{(Kdiag(t) + \lambda I_l)^{-1}\left(Kdiag(m)^2 - \frac{1}{l}Kmm^T\right)\right\}\right]$$

(5.1.6.13)

όπου τα t_i και m_i δίνονται από τις (5.1.6.4) και (5.1.6.6). Το KRIC για την KLR είναι το ίδιο όπως αυτό για το λογιστικό Μπεϋζιανό μοντέλο (5.1.6.13) με n=1 στις (5.1.6.4) και (5.1.6.6).

Στη συνέχεια, κατασκευάζουμε το KRIC για το Μπεϋζιανό μοντέλο του Sollich για τα SVMs. Το Μπεϋζιανό μοντέλο αντιστοιχεί στο ακόλουθο μοντέλο συστηματοποίησης :

$$L_s(D; w, \lambda, \pi) = \log p_S(D|w, \pi) + k_S(w; \lambda)$$

όπου

$$\log p_S(D|w, \pi) \log p(D|w, \pi) - l \log N(a)$$

και

$$k_S(w, \lambda) = \frac{\lambda}{2}\|w\|^2 + l \log N(a)$$

Δεδομένου ότι $L_s(D; w, \lambda, \pi) = L(D; w, \lambda, \pi)$ οι τιμές των $\hat{\jmath}$ και $\hat{I}$ είναι ίδιες με εκείνες του λογιστικού Μπεϋζιανου μοντέλου. Ως εκ τούτου, η διαφορά μεταξύ του KRIC για το μοντέλο του Sollich και αυτού για το λογιστικό μοντέλο είναι μόνο ο όρος πιθανότητας.

Δεδομένου ότι το $p_0(x)$ είναι το άγνωστο πραγματικό μέτρο πιθανότητας, χρησιμοποιούμε $\widehat{N}(a) = \sum_{i=1}^{l} \nu\big(\alpha(x_i)\big)$ αντί του N(α). Συνεπώς , το KRIC γίνεται ως εξής:

KRIC= $\Big[2\sum_i \log\{1 + exp(-na_i y_i)\} - l \log \sum_{i=1}^{l} \nu\big(\alpha(x_i)\big) + trace\Big\{(Kdiag(t) +$
λIl−1Kdiagm2−1lKmmT

όπου τα ti και m_i δίνονται από τις (5.1.6.4) και (5.1.6.6).

5.1.7 Η ΠΡΟΣΕΓΓΙΣΤΙΚΗ ΜΕΘΟΔΟΣ ΤΟΥ NYSTROM ΓΙΑ ΤΟΝ ΥΠΟΛΟΓΙΣΜΟ ΤΟΥ KRIC

Σε αυτήν την ενότητα, παρουσιάζουμε μια προσέγγιση για το KRIC με τη μέθοδο του Nystrom. Το υπολογιστικό κόστος για το KRIC είναι $O(l^3)$, διότι επιλύει την εξίσωση ιδιοτιμών για πίνακες l×l . Όταν χρησιμοποιούμε τον αλγόριθμο SMO, το υπολογιστικό κόστος για τη βελτιστοποίηση των SVMs είναι μικρότερο από $O(l^3)$. Έτσι, το υπολογιστικό κόστος για την M-πλάσια διασταυρωμένη επικύρωση είναι μικρότερο από $O(Ml^3)$ για ένα καθορισμένο M. Αυτό σημαίνει ότι αν το μέγεθος των δεδομένων εκπαίδευσης γίνεται μεγάλο, το KRIC απαιτεί υψηλότερο υπολογιστικό κόστος από ότι η M-πλάσια

διασταυρωμένη επικύρωση. Εφαρμόζουμε τη μέθοδο Nystrom στους υπολογισμούς για το KRIC και επιλύουμε το πρόβλημα.

Στη μέθοδο Nystrom , ο πίνακας Gram $K = K_{ij}$ προσεγγίζεται από ένα μειωμένης τάξης πίνακα τον $\widetilde{K}$. Έστω ότι το $(i_1, \dots, i_m)$ επιλέγεται τυχαία, m δείκτες όπως $1 \leq i_1, i_2, \dots, i_m \leq l$. Ορίζουμε τους πίνακες $K_{m,m}$ και $K_{l,m}$ ως $K_{m,m}(j,k) = K(i_j, i_k)$ για ένα $1 \leq j, k \leq m$ και $K_{l,m}(i,k) = K(i, i_k)$ για $1 \leq i \leq l$ και $1 \leq k \leq m$. Στη συνέχεια ο $K_{m,m}$ είναι ο περιορισμός του Gram πίνακα K στις τυχαία επιλεγμένες m γραμμές και στήλες και ο $K_{l,m}$ είναι ο $l \times m$ περιορισμός του K, με τις ίδιες γραμμές όπως ο $K_{m,m}$.

Έστω $\lambda_i^{(m)}$και $u_i^{(m)}$ να είναι η n-οστή μεγαλύτερη ιδιοτιμή και το αντίστοιχο ιδιοδιάνυσμα. Προσεγγίζουμε το K με

$$\widetilde{K} = \sum_{\iota=1}^{p} \tilde{\lambda}_i^{(l)} u_i^{(l)} \left(u_i^{(l)}\right)^T$$

όπου

$$\tilde{\lambda}_i^{(l)} := \frac{l}{m} \lambda_\iota^{(m)}$$

και

$$\tilde{u}_i^{(l)} := \sqrt{\frac{m}{l}} \frac{1}{\lambda_\iota(m)} K_{l,m} u_i(m)$$

Έστω $\widetilde{U}$ ένας $l \times p$ πίνακας του οποίου οι στήλες διανύσματα είναι $\tilde{u}_i^{(l)}$ $i = (1, \dots, p)$ και $\tilde{\Lambda}$ είναι ένας $p \times p$ διαγώνιος πίνακας του οποίου το (i,i) στοιχείο είναι $\tilde{\Lambda}_i^{(l)}$. Τότε

$$\widetilde{K} = \widetilde{U} \tilde{\Lambda} \widetilde{U}^T$$

Χρησιμοποιώντας τον $\widetilde{K}$ σαν προσέγγιση του K, η τιμή του $trace\left(\hat{J}^{-1}\hat{I}\right)$ γίνεται ως εξής :

$$\begin{aligned} trace\left(\hat{J}^{-1}\hat{I}\right) &= trace\left[(Kdiag(t) + \lambda I_l)^{-1}\left(Kdiag(m) - \frac{1}{l}Kmm^T\right)\right] \\ &\cong trace\left[\left(\widetilde{K}diag(t) + \lambda I_l\right)^{-1}\left(\widetilde{K}diag(m) - \frac{1}{l}\widetilde{K}mm^T\right)\right] \end{aligned}$$

(5.1.7.1)

Από το λήμμα της αντιστροφής του πίνακα, το πιο δεξί μέλος της (5.1.7.1) περιγράφεται χρησιμοποιώντας τους $\widetilde{U}$ *και* $\tilde{\Lambda}$ ως εξής :

$$trace\left[\left(\widetilde{K}diag(t) + \lambda I_l\right)^{-1}\left(\widetilde{K}diag(m) - \frac{1}{l}\widetilde{K}mm^T\right)\right] = trace\left[\left(\widetilde{U}^T diag(t)\widetilde{U} + \right.\right.$$

$\lambda\Lambda{-}1{-}1UTdiagm2{-}1lmmTU$ (5.1.7.2)

Αυτή είναι η προσέγγιση του δεύτερου όρου του KRIC με τη μέθοδο Nystrom. Συνεπώς, το KRIC από τη μέθοδο Nystrom γίνεται όπως περιγράφεται παρακάτω.

<u>Το KRIC για τα SVMs από τη μέθοδο Nystrom:</u>

$$KRIC = 2\Big[\log p\big(D|\widehat{w},\hat{b},\pi\big) + trace\Big[\big(\widetilde{U}^T diag(t)\widetilde{U} + \lambda\tilde{\Lambda}^{-1}\big)^{-1}\widetilde{U}^T\Big(diag(m)^2 - \frac{1}{l}mm^T\Big)\widetilde{U}\Big]\Big]$$

Το υπολογιστικό κόστος για τον υπολογισμό του KRIC με τη μέθοδο του Nystrom είναι $O(ml^2)$. Αν θέσουμε $m \ll l$,το υπολογιστικό κόστος για το KRIC γίνεται πολύ μικρότερο από ότι αυτό για την m-πλάσια διασταυρωμένη επικύρωση ειδικά όταν το μέγεθος δείγματος l είναι μεγάλο.

Data property			
Name	Dim	# train.	# test
Bld	6	230	115
Cra	6	133	67
Hea	13	180	90
Ion	33	234	117
Rsy	2	250	1000
Snr	60	138	70

ΣΧΗΜΑ 5.1.2: Ιδιότητες κάθε συνόλου δεδομένων

5.1.8 ΠΕΙΡΑΜΑΤΑ

Συγκρίνουμε την απόδοση του KRIC και τρία άλλα κριτήρια, την 10-πλάσια διασταυρωμένη επικύρωση (10-fold cross validation),την ένδειξη υπολογισμένη με τη μέθοδο του Laplace (evidence calculated by Laplace's method (Kwok, 2000)) και την ξα-εκτίμηση (ξα-estimation (Joachims, 2000)) .

Όταν υπολογίζουμε την ένδειξη υποθέτουμε το λογιστικό Μπεϋζιανό μοντέλο για τα SVMs. Η ένδειξη γίνεται

$$\log\int p(D|w)\,p(w|\lambda)d\theta \cong \log p(D|\widehat{w}) - \frac{1}{2}\log\Big|\frac{1}{\lambda}diag(t)^{\frac{1}{2}}Kdiag(t)^{\frac{1}{2}} + I_l\Big| + (independent\ of\ \lambda)$$

όπου $\widehat{w}$ είναι η βέλτιστη τιμή υπολογισμένη με τα SVMs και την KLR και τα t και K είναι τα ίδια που χρησιμοποιούνται στην τιμή του KRIC .

Η ξα-εκτίμηση ορίζεται από τη σχέση

$$Err_{\xi\alpha}^{l} = \frac{1}{l}\big|\{i;\big(\rho_{\alpha_i}R_{\Delta}^2\big) \geq 1\}\big|$$

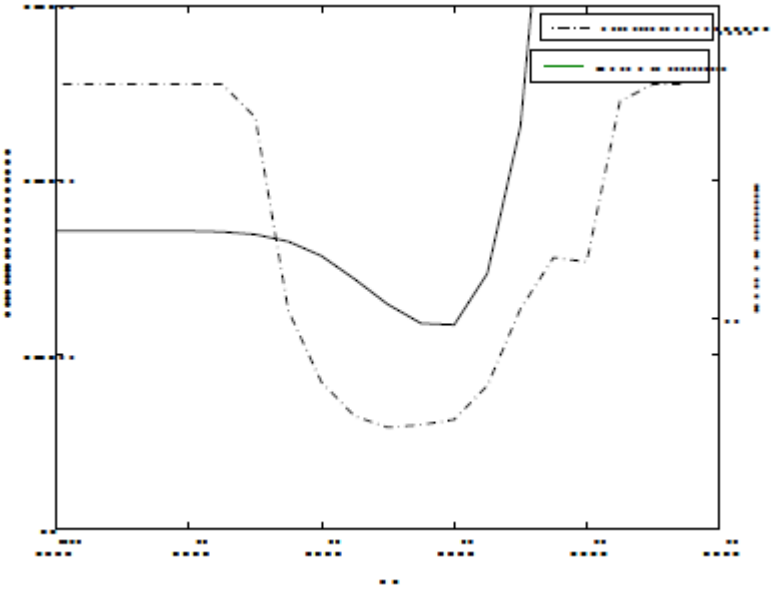
όπου ρ= 2 , $R_{\Delta}^{2} = 1$.

Στο πείραμα χρησιμοποιείται ο Γκαουσιανός πυρήνας:

$$K(x_i, x_j) = \exp\left(-\left\|x_i - x_j\right\|^2 / 2\sigma^2\right)$$

Η παράμετρος συστηματοποίησης C ρυθμίζεται από κάθε κριτήριο. Χρησιμοποιούμε τις συλλογές δεδομένων του ΣΧΗΜΑΤΟΣ 5.1.2 . Η ιδιότητα κάθε συλλογής δεδομένων είναι στο ΣΧΗΜΑ 5.1.2 . Τα ''dim'' ,''# train.'' ,''# test'' υποδηλώνουν τη διάσταση εισαγωγής, το μέγεθος των δεδομένων εκπαίδευσης και το μέγεθος των δεδομένων δοκιμής αντίστοιχα.

Κάθε συλλογή δεδομένων χωρίζεται σε συλλογές δεδομένων εκπαίδευσης $(2/3)$ και σε συλλογές δεδομένων δοκιμής ($1/3$), εκτός από τη συλλογή δεδομένων rsy, η οποία είχε αρχικά διαιρεθεί σε 250 συλλογές δεδομένων εκπαίδευσης και σε 1000 συλλογές δεδομένων δοκιμής τυχαία. Επαναλάβαμε τον τυχαίο χωρισμό 100 φορές και δημιουργούμε 100 διαφορετικές συλλογές δεδομένων. Κάθε συλλογή δεδομένων $\{x_i = (x_{i1}, \dots, x_{in})^T\}_{l=1}^{L}$ είναι κανονικοποιημένη καθώς $\sum_{l=1}^{l} x_{ij}^2 = 1$ για j=1,…,n όπου L είναι το άθροισμα των αριθμών των συλλογών των δεδομένων εκπαίδευσης και δοκιμής.



ΣΧΗΜΑ 5.1.3: Ένα παράδειγμα της ρύθμισης της παραμέτρου συστηματοποίησης από το KRIC. Χρησιμοποιούμε το rsy σύνολο δεδομένων. Η συνεχής και η διακεκομμένη γραμμή αντιπροσωπεύουν την τιμή του KRIC και το ποσοστό σφάλματος δοκιμής, αντίστοιχα. Η παράμετρος κλίμακας στον Γκαουσιανό πυρήνα σ, καθορίζεται στο 10^{-1}.

Θέτουμε τον παραμετρικό χώρο ως $C = \left\{10^{k/2-2}\right\}_{k=0}^{19}$. Επιλέγουμε την καλύτερη τιμή του C στο σύνολο C με την ελαχιστοποίηση ή τη μεγιστοποίηση κάθε κριτηρίου. Η παράμετρος κλίμακας στο Γκαουσιανό πυρήνα, σ ορίζεται ως 10.0 .

Οι πίνακες των σχημάτων (5.1.4) και (5.1.5) είναι τα αποτελέσματα της σύγκρισης των επιδόσεων των κριτηρίων για τα SVMs και την KLR, αντίστοιχα. Συγκρίνουμε το KRIC (το λογιστικό μοντέλο) , τη GACV, τη 10-πλάσια διασταυρωμένη επικύρωση, το KRIC (το μοντέλο του Sollich), την ένδειξη υπολογισμένη με τη μέθοδο του Laplace και την ξα-εκτίμηση. Τα δύο είδη του KRIC και η ένδειξη είναι υπολογισμένα με την προσέγγιση Nystrom. Στην προσέγγιση του Nystrom, το μέγεθος των τυχαία επιλεγμένων υποπινάκων ορίζεται ως m=50. Ο αριθμός των επιλεγμένων κύριων συνιστωσών ορίζεται ως p=30. Το ERROR (σφάλμα) δηλώνει το μέσο ποσοστό σφάλματος δοκιμής μαζί με την τυπική απόκλιση. Το TIME (χρόνος) δηλώνει το μέσο χρόνο CPU σε δευτερόλεπτα που

καταναλώνεται με υπολογισμό του κάθε κριτηρίου για τη ρύθμιση της παραμέτρου συστηματοποίησης μαζί με την τυπική απόκλιση.

Η μέθοδος SVM και η 10-πλάσια διασταυρωμένη επικύρωση αξιολογήθηκε με τον OSU-SVM αλγόριθμο και η KLR αξιολογήθηκε με τον SMO αλγόριθμο .

Υπολογίσαμε την τιμή p-value για τα t-tests ανά ζεύγη για τη διαφορά σφάλματος μεταξύ του KRIC με τη μέθοδο Nystrom και των άλλων κριτηρίων. Επίσης υπολογίσαμε τη τιμή p-value για τη διαφορά του χρόνου CPU. Ένα(κατώτατο) όριο για την p-value ίσο με 0,05 χρησιμοποιήθηκε για να αποφασίσουμε τη σημαντική διαφορά. Στους πίνακες (5.1.4) και (5.1.5) κάθε σημάδι * (ή @) στα αριστερά της p-value του κριτηρίου αναπαριστά ότι η απόδοση του κριτηρίου είναι σημαντικά καλύτερη (ή χειρότερη από αυτή του KRIC).

Από αυτό το αποτέλεσμα, μπορούμε να πούμε ότι δεν υπάρχει κριτήριο του οποίου το μέσο σφάλμα δοκιμής είναι σημαντικά χαμηλότερο από αυτό του KRIC. Το ποσοστό σφάλματος του KRIC για το λογιστικό Μπεϋζιανό μοντέλο είναι χαμηλότερο από αυτό της GACV και της ξα-εκτίμησης. Το μέσο υπολογιστικό κόστος του KRIC είναι χαμηλότερο από αυτό της 10-πλάσιας διασταυρωμένης επικύρωσης και τον υπολογισμό της ένδειξης με τη μέθοδο του Laplace. Από την άλλη πλευρά, ο υπολογισμός της ξα-εκτίμησης και της GACV είναι γρηγορότερος από αυτόν του KRIC. Το μέσο σφάλμα δοκιμής που δίδεται από το KRIC για το Μπεϋζιανό μοντέλο του Sollich είναι σημαντικά χειρότερο από αυτό για το λογιστικό Μπεϋζιανό μοντέλο για κάποιο σύνολο δεδομένων.

Σημειώνουμε ότι το ποσοστό σφάλματος δοκιμής δοσμένο από το KRIC για την KLR είναι καλύτερο από εκείνο για τα SVMs σε σύγκριση με άλλα κριτήρια. Ειδικότερα το μέσο ποσοστό σφάλματος που δίνεται από το KRIC είναι καλύτερο από αυτό που δίνεται από τη 10-πλή διασταυρωμένη επικύρωση και τα αποδεικτικά στοιχεία για τα περισσότερα από τα σύνολα δεδομένων αν και η διαφορά δεν είναι σημαντική στο t-test. Το σχήμα 5.1.3 είναι ένα παράδειγμα ρύθμισης της παραμέτρου συστηματοποίησης από το KRIC.

data set		bld		cra	
		ERROR or TIME	p-value	ERROR or TIME	p-value
optimal	ERROR	0.2821 ± 0.0344	-	0.0000 ± 0.0000	-
	TIME(s)	1.6829 ± 1.1600	-	0.1070 ± 0.1428	-
KRIC	ERROR	0.2888 ± 0.0355	-	0.0009 ± 0.0036	-
(logistic)	TIME(s)	0.2456 ± 0.0128	-	0.1088 ± 0.0080	-
GACV	ERROR	0.4131 ± 0.0381	@0.0456	0.5391 ± 0.0285	@ 0.0000
	TIME(s)	0.1575 ± 0.0040	*0.0000	0.0621 ± 0.0026	* 0.0000
10-fold	ERROR	0.2908 ± 0.0377	0.4892	0.0004 ± 0.0013	0.4559
CV	TIME(s)	13.2987 ± 4.2776	@0.0012	0.8745 ± 0.5552	0.0870
KRIC	ERROR	0.3891 ± 0.0458	0.1085	0.4582 ± 0.0818	@ 0.0000
(Sollich)	TIME(s)	0.2465 ± 0.0164	0.4883	0.1037 ± 0.0076	0.3705
evidence	ERROR	0.3204 ± 0.0383	0.3340	0.0009 ± 0.0036	0.5000
(Lap.)	TIME(s)	0.5258 ± 0.0201	@0.0000	0.1933 ± 0.0063	@ 0.0000
$\xi\alpha$-est.	ERROR	0.4131 ± 0.0381	@0.0456	0.0009 ± 0.0036	0.5000
	TIME(s)	0.1869 ± 0.0041	*0.0003	0.0826 ± 0.0017	* 0.0034

data set		hea		ion	
		ERROR or TIME	p-value	ERROR or TIME	p-value
optimal	ERROR	0.2109 ± 0.0307	-	0.0855 ± 0.0221	-
	TIME(s)	2.2332 ± 1.6830	-	0.4544 ± 0.2921	-
KRIC	ERROR	0.2340 ± 0.0371	-	0.1342 ± 0.0193	-
(logistic)	TIME(s)	0.1959 ± 0.0144	-	0.1450 ± 0.0184	-
GACV	ERROR	0.4622 ± 0.0645	@0.0124	0.1598 ± 0.0221	0.2680
	TIME(s)	0.1474 ± 0.0058	*0.0083	0.2369 ± 0.0208	@ 0.0097
10-fold	ERROR	0.2305 ± 0.0349	0.4806	0.1024 ± 0.0324	0.2696
CV	TIME(s)	17.9704 ± 8.2644	@0.0159	2.9289 ± 1.3944	@ 0.0244
KRIC	ERROR	0.3983 ± 0.1020	0.1188	0.1342 ± 0.0193	0.5000
(Sollich)	TIME(s)	0.1975 ± 0.0159	0.4786	0.1424 ± 0.0174	0.4714
evidence	ERROR	0.2321 ± 0.0341	0.4894	0.1111 ± 0.0248	0.3007
(Lap.)	TIME(s)	0.4804 ± 0.0293	@0.0000	0.6176 ± 0.0746	@ 0.0000
$\xi\alpha$-est.	ERROR	0.4233 ± 0.0980	0.0807	0.1205 ± 0.0195	0.3624
	TIME(s)	0.1845 ± 0.0054	0.2828	0.0599 ± 0.0056	* 0.0000

data set		rsy		snr	
		ERROR or TIME	p-value	ERROR or TIME	p-value
optimal	ERROR	0.1072 ± 0.0069	-	0.2057 ± 0.0345	-
	TIME(s)	7.3261 ± 8.1287	-	0.0793 ± 0.0136	-
KRIC	ERROR	0.1112 ± 0.0060	-	0.2429 ± 0.0517	-
(logistic)	TIME(s)	0.1956 ± 0.0173	-	0.1090 ± 0.0277	-
GACV	ERROR	0.4835 ± 0.0755	@ 0.0000	0.3071 ± 0.0828	0.3164
	TIME(s)	0.1062 ± 0.0044	* 0.0000	0.1027 ± 0.0197	0.4472
10-fold	ERROR	0.1100 ± 0.0073	0.4639	0.2371 ± 0.0358	0.4740
CV	TIME(s)	38.4810 ± 14.2007	@ 0.0035	0.5163 ± 0.0670	@0.0000
KRIC	ERROR	0.2858 ± 0.0706	@ 0.0113	0.2443 ± 0.0454	0.4941
(Sollich)	TIME(s)	0.2113 ± 0.0095	0.2788	0.1069 ± 0.0182	0.4822
evidence	ERROR	0.1138 ± 0.0065	0.4171	0.2229 ± 0.0382	0.4120
(Lap.)	TIME(s)	0.3859 ± 0.0184	@ 0.0000	0.2975 ± 0.0436	@0.0041
$\xi\alpha$-est.	ERROR	0.1080 ± 0.0076	0.4067	0.2430 ± 0.0516	0.4994
	TIME(s)	0.1461 ± 0.0024	* 0.0058	0.2403 ± 0.0099	*0.0002

ΣΧΗΜΑ 5.1.4 :Σύγκριση των κριτηρίων για τη ρύθμιση των παραμέτρων συστηματοποίησης για τα SVMs (Γκαουσιανός πυρήνας , σ=10.0). Συγκρίνουμε το KRIC (για το λογιστικό μοντέλο), τη GACV, τη 10-πλάσια διασταυρωμένη επικύρωση, το KRIC (για το μοντέλο του Sollich) την ένδειξη υπολογισμένη με τη μέθοδο του Laplace και την ξα-εκτίμηση. Η προσέγγιση του Nystrom χρησιμοποιείται στον υπολογισμό του KRIC και της ένδειξης.

data set		bld		cra	
		ERROR or TIME	p-value	ERROR or TIME	p-value
optimal	ERROR	0.3374 ± 0.0426	-	0.0060 ± 0.0077	-
	TIME(s)	12.2935 ± 0.4063	-	8.1325 ± 0.4327	-
KRIC	ERROR	0.3409 ± 0.0430	-	0.0060 ± 0.0077	-
(logistic)	TIME(s)	0.1519 ± 0.0102	-	0.0881 ± 0.0036	-
GACV	ERROR	0.4200 ± 0.0318	0.1448	0.5313 ± 0.0316	@0.0000
	TIME(s)	0.8097 ± 0.5817	0.1332	0.0893 ± 0.0047	0.4396
10-fold	ERROR	0.3452 ± 0.0382	0.4786	0.0164 ± 0.0192	0.3489
CV	TIME(s)	108.8243 ± 7.3829	@0.0000	71.6293 ± 5.6273	@0.0000
KRIC	ERROR	0.4217 ± 0.0326	0.1422	0.3940 ± 0.0581	@0.0000
(Sollich)	TIME(s)	0.1517 ± 0.0093	0.4971	0.0941 ± 0.0168	0.3844
evidence	ERROR	0.3417 ± 0.0420	0.4959	0.0060 ± 0.0077	0.5000
(Lap.)	TIME(s)	0.3304 ± 0.0238	@0.0000	0.1311 ± 0.0111	@0.0017
$\xi\alpha$-est.	ERROR	0.4217 ± 0.0326	0.1422	0.5313 ± 0.0316	@0.0000
	TIME(s)	0.1016 ± 0.0072	*0.0019	0.0327 ± 0.0043	*0.0000

data set		hea		ion	
		ERROR or TIME	p-value	ERROR or TIME	p-value
optimal	ERROR	0.2011 ± 0.0430	-	0.1077 ± 0.0352	-
	TIME(s)	10.9960 ± 0.5419	-	14.9930 ± 0.8140	-
KRIC	ERROR	0.2167 ± 0.0511	-	0.1103 ± 0.0355	-
(logistic)	TIME(s)	0.1515 ± 0.0139	-	0.2555 ± 0.0370	-
GACV	ERROR	0.2878 ± 0.1476	0.3603	0.1863 ± 0.0516	0.1910
	TIME(s)	0.2352 ± 0.0278	@0.0224	1.0713 ± 0.1496	@0.0000
10-fold	ERROR	0.2181 ± 0.0342	0.4931	0.1098 ± 0.0351	0.4976
CV	TIME(s)	98.4162 ± 8.6210	@ 0	129.8559 ± 17.2304	@0.0000
KRIC	ERROR	0.2456 ± 0.0543	0.3920	0.1265 ± 0.0471	0.4221
(Sollich)	TIME(s)	0.1526 ± 0.0167	0.4858	0.2557 ± 0.0365	0.4991
evidence	ERROR	0.2167 ± 0.0511	0.5000	0.1128 ± 0.0351	0.4855
(Lap.)	TIME(s)	0.2826 ± 0.0308	@0.0017	1.3124 ± 0.2161	@0.0000
$\xi\alpha$-est.	ERROR	0.4344 ± 0.1036	0.0797	0.3530 ± 0.0472	@0.0017
	TIME(s)	0.0642 ± 0.0057	*0.0000	0.5334 ± 0.0868	*0.0124

data set		rsy		snr	
		ERROR or TIME	p-value	ERROR or TIME	p-value
optimal	ERROR	0.1206 ± 0.0089	-	0.2143 ± 0.0243	-
	TIME(s)	20.6412 ± 0.5797	-	9.3232 ± 0.7657	-
KRIC	ERROR	0.1235 ± 0.0098	-	0.2343 ± 0.0279	-
(logistic)	TIME(s)	0.3786 ± 0.0199	-	0.1755 ± 0.0201	-
GACV	ERROR	0.2593 ± 0.1740	0.2301	0.4671 ± 0.0738	@0.0110
	TIME(s)	0.8985 ± 0.0943	@0.0000	0.2693 ± 0.0488	0.0868
10-fold	ERROR	0.1233 ± 0.0112	0.4968	0.2529 ± 0.0261	0.3656
CV	TIME(s)	88.7065 ± 5.6848	@0.0000	84.3182 ± 10.6565	@0.0000
KRIC	ERROR	0.2535 ± 0.0430	@0.0070	0.2729 ± 0.0353	0.2709
(Sollich)	TIME(s)	0.3917 ± 0.0176	0.3628	0.1665 ± 0.0218	0.4145
evidence	ERROR	0.1283 ± 0.0129	0.4163	0.2371 ± 0.0271	0.4793
(Lap.)	TIME(s)	0.6113 ± 0.0193	@0.0000	0.5545 ± 0.0746	@0.0000
$\xi\alpha$-est.	ERROR	0.4697 ± 0.1159	@0.0030	0.4643 ± 0.0817	@0.0180
	TIME(s)	0.1487 ± 0.0030	*0.0000	0.1640 ± 0.0160	0.3745

ΣΧΗΜΑ 5.1.5 :Σύγκριση των κριτηρίων για τη ρύθμιση της παραμέτρου συστηματοποίησης της KLR (Γκαουσιανός πυρήνας , σ=10.0). Συγκρίνουμε το KRIC (το λογιστικό μοντέλο), τη GACV, τη 10-πλάσια διασταυρωμένη επικύρωση, το KRIC (για το μοντέλο του Sollich), την ένδειξη υπολογισμένη από τη μέθοδο του Laplace και την ξα-εκτίμηση. Τα δύο είδη του KRIC και η ένδειξη είναι υπολογισμένα από τη προσέγγιση του Nystrom. Η p-value είναι για το ζευγαρωτό t-test στο ποσοστό σφάλματος και το χρόνο CPU σύμφωνα με το KRIC (για λογιστικό μοντέλο).

5.1.9 ΣΥΜΠΕΡΑΣΜΑΤΑ

Το KRIC εισήχθη για τη ρύθμιση της παραμέτρου συστηματοποίησης στα SVMs και την KLR. Το κριτήριο θα μπορούσε να υπολογιστεί με χαμηλό υπολογιστικό κόστος, αν γινόταν χρήση κάποιων μεθόδων προσέγγισης.

Πρώτα κατασκευάσαμε τα στατιστικά μοντέλα των SVMs και της KLR. Θεωρήσαμε επίσης δύο μοντέλα συστηματοποίησης που να αντιστοιχούν σε καθένα από τα Μπεϋζιανά μοντέλα για τα SVMs. Το RIC για αυτά τα μοντέλα δεν μπορεί να αξιολογηθεί άμεσα, επειδή ο παραμετρικός χώρος, είναι ο χώρος Hilbert αναπαραγωγής πυρήνα. Για να προκύψει το RIC, εισάγουμε ένα πρόβλημα ιδιοτιμών και αποδεικνύουμε ότι το RIC υπολογίζεται από τις ιδιοτιμές. Καλούμε το RIC που υπολογίζεται με αυτό τον τρόπο ως το RIC πυρήνα (KRIC).

Στη συνέχεια παρουσιάστηκε μια προσέγγιση του KRIC με τη χρήση της μεθόδου Nystrom. Στη μέθοδο Nystrom ο Gram πίνακας προσεγγίζεται από έναν υποπίνακα και τις κύριες συνιστώσες. Ως εκ τούτου, το υπολογιστικό κόστος για τον υπολογισμό του KRIC με τη μέθοδο Nystrom γίνεται πολύ χαμηλό. Στη συνέχεια η επιλογή του μοντέλου από το KRIC με την προσέγγιση γίνεται πολύ γρηγορότερα από αυτή με τη διασταυρωμένη επικύρωση όταν το μέγεθος του δείγματος είναι μεγάλο.

Από το πειραματικό αποτέλεσμα, το μέσο ποσοστό σφάλματος δοκιμής με την παράμετρο συστηματοποίησης συντονισμένη από το KRIC για το λογιστικό Μπεϋζιανό μοντέλο είναι χαμηλότερο από αυτό από τη GACV και την ξα-εκτίμηση. Το μέσο ποσοστό σφάλματος είναι συγκρίσιμο από αυτό από τη 10-πλάσια διασταυρωμένη επικύρωση και την ένδειξη που είναι αξιολογημένη με τη μέθοδο του Laplace. Το υπολογιστικό κόστος για τον υπολογισμό του KRIC είναι πολύ χαμηλότερο από τα αντίστοιχα για αυτά τα δύο κριτήρια.

Στην προκειμένη περίπτωση, το KRIC χρησιμοποιείται για τη ρύθμιση της παραμέτρου συστηματοποίησης. Εντούτοις, μόνο οι παράμετροι πυρήνα και η συνάρτηση πυρήνα μπορούν να ρυθμιστούν με το KRIC.

Το KRIC, στη συγκεκριμένη ενότητα, χρησιμοποιείται μόνο για τη ρύθμιση της παραμέτρου συστηματοποίησης. Εντούτοις, οι παράμετροι πυρήνα και η συνάρτηση πυρήνα μπορούν να ρυθμιστούν από το KRIC. Εξηγούμε το λόγο, το KRIC μπορεί να αξιολογήσει το αριστερό μέλος της (5.1.5.2) , όπου $p_0(d)$ είναι η πραγματική συνάρτηση πυκνότητας και η $\theta^*(\acute{D},\lambda)$ μεγιστοποιεί την $L(\acute{D};\theta;\lambda)$. Αφού η τιμή της $L(\acute{D};\theta;\lambda)$ εξαρτάται από την παράμετρο πυρήνα π και τη συνάρτηση πυρήνα K, η $\theta^*(\acute{D},\lambda)$ επίσης εξαρτάται από τα π και K. Αυτός είναι ο λόγος για τον οποίο το KRIC μπορεί να χρησιμοποιηθεί για την επιλογή των π και K επίσης.

Διεξάγουμε κάποια πειράματα για την αξιολόγησης της ρύθμισης της παραμέτρου πυρήνα από το KRIC. Χρησιμοποιούμε τα ίδια σύνολα δεδομένων όπως αυτά που χρησιμοποιήθηκαν στην ενότητα 5.1.8 , για την επιλογή και της παραμέτρου κλίμακας σ στο Γκαουσιανό πυρήνα και της παραμέτρου συστηματοποίησης C . Η παράμετρος κλίμακας σ είναι καλά ρυθμισμένη από το KRIC όταν περιορίζουμε το σ σε $\sigma \leq 10^{-4}$, το KRIC παίρνει ολικό ελάχιστο σε κάθε τέτοια περιοχή του σ. Εντούτοις, το ολικό ελάχιστο της δοκιμής λάθους δεν επιτυγχάνεται στην περιοχή.

Μια απλοϊκή λύση του προβλήματος, είναι να αποκλείσουμε την περίπτωση που το σ είναι μικρότερο από μία συγκεκριμένη τιμή (π.χ. 10^{-4} στα παραπάνω πειράματα). Τουλάχιστον στα πειράματα οι βέλτιστες παράμετροι (με την έννοια του ελαχίστου σφάλματος δοκιμής) δε βρίσκονται στην αποκλεισμένη περιοχή, διότι το σ είναι τόσο μικρό ώστε να λαμβάνει χώρα υπερπροσαρμογή. Έτσι η ρύθμιση της περιοχής σ είναι λογική. Επιπλέον η οριακή τιμή 10^{-4} θα έπρεπε να λαμβάνεται εξαρτώμενη από το μέγεθος δείγματος.

Επίσης δεν έχει υποτεθεί καμιά παράμετρος μεροληψίας b στο Μπεϋζιανό μοντέλο για τα SVMs. Εντούτοις, τα αποτελέσματά μας εύκολα μπορούν να εφαρμοστούν στα τροποποιημένα SVMs που έχουν τετραγωνική ποινή για τη μεροληψία. Θεωρείστε την τροποποιημένη τετραγωνική συνάρτηση (5.1.2.4) . Στη συνέχεια η πιθανότητα συστηματοποίησης (5.1.7.1) αλλάζει σε :

$$L(D;w,b;\lambda,\pi)\log p(D|w,b,\pi) - \frac{\lambda}{2}\|w\|^2 - \frac{\lambda B^{-2}}{2}b^2$$

Έστω $\widetilde{w}^T = [w^T B^{-1} b]$ και $\tilde{z}^T = [z^T - B]$. Το KRIC για αυτό το μοντέλο δίνεται με την αντικατάσταση του

$$\tilde{I}_d = I_{d+1}$$

στις σχέσεις (5.1.6.2) και (5.1.6.7) αντί της (5.1.6.3). Το παραπάνω μοντέλο συστηματοποίησης έχει δύο παραμέτρους συστηματοποίησης. Αν συντονίσουμε και τις δύο παραμέτρους συστηματοποίησης, η ελάχιστη τιμή του KRIC γίνεται μικρότερη από αυτή που δίνεται με τη ρύθμισης μόνο της μίας παραμέτρου συστηματοποίησης. Αυτό σημαίνει ότι δε μπορούμε να συγκρίνουμε μοντέλα που να έχουν διαφορετικούς αριθμούς παραμέτρων συστηματοποίησης από το KRIC. Με άλλα λόγια το KRIC δε μπορεί να αποφύγει την υπερπροσαρμογή σε σχέση με τις παραμέτρους συστηματοποίησης.

5.2 ΕΝΑ ΚΡΙΤΗΡΙΟ ΠΛΗΡΟΦΟΡΙΑΣ ΓΙΑ ΕΠΙΛΟΓΗ ΜΕΤΑΒΛΗΤΩΝ ΣΤΙΣ ΜΗΧΑΝΕΣ ΔΙΑΝΥΣΜΑΤΙΚΗΣ ΥΠΟΣΤΗΡΙΞΗΣ

5.2.1 ΕΙΣΑΓΩΓΗ

Σε πολλές στατιστικές εφαρμογές και ειδικότερα στην ανάλυση παλινδρόμησης για προγνωστικούς σκοπούς, συνίσταται η επιλογή ενός υποσυνόλου μεταβλητών για να μοντελοποιηθούν τα διαθέσιμα δεδομένα εκπαίδευσης. Οι λόγοι για αυτό είναι ποικίλοι. Καταρχάς, το ότι πρέπει να εκτιμηθούν πολλές παράμετροι, οδηγεί σε ένα κακώς τοποθετημένο μοντέλο για τα δεδομένα. Ένας άλλος καλός λόγος για τη χρήση μεταβλητών επιλογής είναι οι ΄΄φτωχές΄΄ προβλέψεις που προκύπτουν από την προσαρμογή ενός μοντέλου με πάρα πολλές μεταβλητές. Αυτό είναι αληθές, ειδικά σε προβλήματα ταξινόμησης, καθώς η πιθανότητα της επίτευξης ενός τέλειου διαχωρισμού των δεδομένων εκπαίδευσης αυξάνει με τον αριθμό των μεταβλητών πρόβλεψης. Επιπλέον προσθήκη περισσοτέρων μεταβλητών οδηγεί σε αυξημένη μεταβλητότητα.

Μπορεί να υποστηριχθεί ότι η επιλογή μεταβλητών δε θα ήταν πραγματικά αναγκαία στην τοποθέτηση των SVMs, δεδομένου ότι καταφέρνει να παρακάμψει τη λεγόμενη ΄΄κατάρα των διαστάσεων΄΄(curse of dimensionality). Αυτή η συλλογιστική είναι ωστόσο αληθής σε κάποιο βαθμό. Ενώ η προσέγγιση SVM αποφεύγει την προσαρμογή ενός αριθμού παραμέτρων ίσου με τη διάσταση του χώρου εισαγωγής, η υψηλή πιθανότητα ενός τέλειου διαχωρισμού σε προβλήματα υψηλής διάστασης παραμένει. Ως εκ τούτου, ο κίνδυνος απόκτησης ενός κανόνα απόφασης με κακές ιδιότητες γενίκευσης (δηλαδή υψηλό ποσοστό σφάλματος εντός δείγματος) δε μπορεί να αποφευχθεί.

Μια συχνά χρησιμοποιούμενη τεχνική για την επιλογή μεταβλητών είναι με τη χρήση ενός κριτηρίου πληροφορίας. Αυτό αποδίδει μια αξία ΄΄καλής προσαρμογής΄΄ σε κάθε υποσύνολο των υπό εξέταση μεταβλητών και επιλέγει εκείνο με την καλύτερη τιμή για το κριτήριο, ως το καταλληλότερο μοντέλο. Παραδείγματα τέτοιων κριτηρίων είναι το κριτήριο πληροφορίας ΑΚΑΙΚΕ (AIC) και το Μπεϋζιανό κριτήριο πληροφορίας (BIC) για τη (γενικευμένη) γραμμική παλινδρόμηση. Για τα SVMs, όμως, διαπιστώνουμε ότι μέχρι τώρα μόνο μετρικά κριτήρια πληροφορίας έχουν αναπτυχθεί. Ένα από αυτά τα υπάρχοντα κριτήρια πληροφορίας είναι το KRIC, το οποίο είχε αρχικά προταθεί για τη ρύθμιση των παραμέτρων του SVM, αντί για την επιλογή μεταβλητών. Παρακάτω παρουσιάζονται δύο νέα κριτήρια πληροφορίας, το ένα μοιράζεται κάποιες ιδιότητες με το AIC, ενώ το άλλο με το BIC. Γίνεται η σύγκριση των επιδόσεών τους με ποικίλες άλλες τεχνικές επιλογής μεταβλητών για τα SVMs. Πιο συγκεκριμένα είναι αναγκαίο αυτά τα νέα κριτήρια να επιλέγουν μοντέλα με καλές ιδιότητες πρόβλεψης και τα ταυτόχρονα είναι αναγκαίο τα κριτήρια να είναι εύκολα υπολογίσιμα.

Στη συνέχεια, ορίζονται τα νέα κριτήρια πληροφορίας και τονίζονται τα πλεονεκτήματα τα οποία αυτά τα νέα κριτήρια έχουν σε σχέση με αυτά που ήδη έχουν αναπτυχθεί. Έπειτα, περιγράφονται τα αποτελέσματα μιας μελέτης προσομοίωσης για τη σύγκριση της απόδοσης των νέων κριτηρίων σε σχέση με τις μεθόδους για την επιλογή μεταβλητών στα SVMs και στη συνέχεια συγκρίνονται οι διάφορες τεχνικές σε μερικά σύνολα δεδομένων αναφοράς στον πραγματικό κόσμο.

5.2.2 ΥΠΑΡΧΟΥΣΕΣ ΤΕΧΝΙΚΕΣ ΕΠΙΛΟΓΗΣ ΜΕΤΑΒΛΗΤΗΣ

Το ζήτημα επιλογής μεταβλητών ή χαρακτηριστικών στα SVMs έχει ερευνηθεί αρκετά και οι περισσότερες τεχνικές που έχουν αναπτυχθεί έως τώρα βασίζονται είτε σε

μετασχηματισμό του χώρου εισόδου, είτε στη χρήση μιας διαφορετικής αντικειμενικής συνάρτησης στο σχηματισμό του SVM. Παρακάτω γίνεται αναφορά στις τεχνικές που χρησιμοποιούν τα δεδομένα ως έχουν. Μια από τις πιο κοινές τεχνικές συνίσταται από την κατάταξη των μεταβλητών κατά τέτοιο τρόπο ανάλογα με τη σημασία τους και από την επιλογή των πιο σημαντικών, όπου ο αριθμός των διατηρούμενων μεταβλητών προσδιορίζεται εκ των προτέρων. Ένα μειονέκτημα αυτής της μεθόδου είναι ότι οι εξωτερικές πληροφορίες απαιτούνται για τον καθορισμό του αριθμού των απαιτούμενων μεταβλητών.

Είναι φυσικά προτιμότερο να γίνει επιλογή μεταβλητών με χρήση μόνο πληροφοριών που περιέχονται στα δεδομένα. Μία σύγκριση της χρήσης των διασταυρωμένα επικυρωμένου ποσοστού σφάλματος (CV error rate) και της εγγυημένης ελαχιστοποίησης κινδύνου (GRM) αποδεικνύει ότι κανένα κριτήριο δεν είναι σταθερά καλύτερο από τα άλλα, καθώς το μέγεθος δείγματος n ποικίλει.

5.2.3 ΤΕΧΝΙΚΕΣ ΚΑΤΑΤΑΞΗΣ ΜΕΤΑΒΛΗΤΗΣ

Ένα σημαντικό πρόβλημα είναι η εκθετική αύξηση του αριθμού των υπό εξέταση μοντέλων με την αύξηση του αριθμού των μεταβλητών. Για αυτό το λόγο μια πλήρης έρευνα υποσυνόλου είναι υπολογιστικά ανέφικτη ακόμα και για προβλήματα με ένα μόνο μικρό αριθμό διαστάσεων. Για να αποφευχθεί αυτό το πρόβλημα, διαφορετικές τεχνικές έχουν θεσπιστεί για να μειώσουν δραματικά τον αριθμό των υπό εξέταση μοντέλων, ενώ επιλέγουν ακόμα ένα μοντέλο το οποίο είναι ΄΄σχεδόν΄΄ το καλύτερο.

Κάποιες τεχνικές (γενετικός αλγόριθμος και στρατηγική της από κοινού προς τα πίσω απαλοιφής /προς τα εμπρός επιλογής (combined backward elimination/forward selection strategy)) που προτάθηκαν στη συνέχεια έπασχαν από την πιθανότητα ελέγχου ενός πολύ μεγάλου μοντέλου πριν την εύρεση μίας λύσης, για αυτό το λόγο ακολουθείται η τεχνική της κατάταξης μεταβλητής (the technique of variable ranking). Αυτή αποτελείται από την ανάθεση μιας΄΄ τιμής σημαντικότητας΄΄ σε κάθε μεταβλητή και το κρησάρισμα των μεταβλητών ανάλογα με τη σημαντικότητά τους. Αυτό έχει ως αποτέλεσμα μια σειρά από p εμφωλευμένα μοντέλα, και ως τέτοια, μόνο p αξιολογήσεις του κριτηρίου επιλογής μεταβλητής χρειάζεται. Ο πιο ευρέως χρησιμοποιημένος αλγόριθμος είναι η SVM τεχνική αναδρομικής εξάλειψης/κατάργησης χαρακτηριστικών (SVM-RFE).

Για ένα γραμμικό SVM, προτείνεται η ταξινόμηση των μεταβλητών από το w_j^2 , όπου w_j είναι η j-οστή συνιστώσα του διανύσματος βάρους w. Να σημειωθεί ότι αυτή η μέθοδος ταξινόμησης έχει νόημα μόνο όταν οι μεταβλητές έχουν τυποποιηθεί έτσι ώστε να έχουν μέση τιμή μηδέν και διακύμανση ίση με ένα. Διαφορετικά είναι πιθανό ότι μία άσχετη μεταβλητή θα επιλεχθεί ως η πιο σημαντική όταν η τυπική της απόκλιση είναι αρκετά μεγέθη υψηλότερη από τις τυπικές αποκλίσεις των άλλων μεταβλητών. Αυτό το κριτήριο κατάταξης επεκτείνεται, για να γίνει εφαρμόσιμο στα SVMs με ένα μη γραμμικό πυρήνα και προτείνονται διάφορα άλλα κριτήρια κατάταξης. Επιπροσθέτως προτείνεται η χρήση του

$$S_j = \left|w_j\left(m_{j,+1} - m_{j,-1}\right)\right|,$$

όπου $m_{j,+1}$ και $m_{j,-1}$ είναι οι εντός κλάσης μέσοι της μεταβλητής j . Τέλος προτείνεται μία μέθοδος κατάταξης η οποία χρησιμοποιεί το score του Fisher

$$S_j = \frac{\left(m_{j,+1} - m_{j,-1}\right)}{\sqrt{\sigma_{j,+1}^2 + \sigma_{j,-1}^2}}$$

για ένα γραμμικό SVM, όπου τα $\sigma_{j,+1}^2$ και $\sigma_{j,-1}^2$ είναι οι εντός κλάσης διακυμάνσεις της μεταβλητής j. Το κύριο πλεονέκτημα της χρήσης του score του Fisher είναι ότι δεν είναι

αναγκαίο να εκπαιδευτεί οποιοδήποτε SVM για να ταξινομήσει τις μεταβλητές με αυτό το κριτήριο. Το score του Fisher καθώς επίσης και ο SVM-RFE αλγόριθμος χρησιμοποιούνται για να ταξινομήσουν τις μεταβλητές, με μεταβλητή επιρροής

$$\Delta\|w_S\|^2_{(j)} \stackrel{\text{def}}{=} \left|\|w_S\|^2 - \|w_{S\setminus\{j\}}\|\right|.$$

Στη συνέχεια παρουσιάζεται ο SVM-RFE αλγόριθμος :

Βήμα 1: Αρχικοποίησε το S={1,…,p}, το υποσύνολο των μη ταξινομημένων χαρακτηριστικών και r=() η λίστα των ταξινομημένων χαρακτηριστικών

Βήμα 2: Επανέλαβε τα ακόλουθα βήματα μέχρι το S=Ø

Βήμα 2a: Εκπαίδευσε ένα SVM σε (x_i, S, y_i) και έστω $\|w_S\|^2 = a_S' Q_S a_S$ η τετραγωνική νόρμα του διανύσματος βάρους

Βήμα 2b: Για κάθε $j \in S$ εκπαίδευσε ένα νέο SVM στο $\left(x_{i,S\setminus\{j\}}, y_i\right)$. Αυτό δίνει $\left\|w_{S/(j)}\right\|^2 = a_{S/\{j\}} Q_{S/\{j\}} a_{S/\{j\}}$ για κάθε $j \in S$.

Βήμα 2c: Δήλωσε

$$m = arg\min_j \left|\|w_S\|^2 - \left\|w_{S,(j)}\right\|^2\right|$$

και έστω S=S\{m} και r=(m,r).

Το διάνυσμα r περιέχει τις ταξινομημένες μεταβλητές, με πρώτο στοιχείο την πιο σημαντική μεταβλητή.

5.2.4 ΤΑ ΝΕΑ ΚΡΙΤΗΡΙΑ ΠΛΗΡΟΦΟΡΙΑΣ

Σε αυτήν την ενότητα παρουσιάζονται δύο νέα κριτήρια πληροφορίας και θα δοθεί ένα κίνητρο για τη χρήση τους. Η αξιολόγηση του διασταυρωμένα επικυρωμένου ποσοστού λάθους (CV error rate) και του KRIC, ενός συγκεκριμένου SVM, απαιτεί μεγάλο αριθμό πρόσθετων υπολογισμών. Για αυτό το λόγο προτείνεται ένα νέο κριτήριο το οποίο χρησιμοποιεί πληροφορίες ήδη διαθέσιμες στο SVM, χωρίς πρόσθετους περίπλοκους υπολογισμούς. Αυτό το κριτήριο βασίζεται στο πόσο άσχημα το SVM παραβιάζει τους περιορισμούς του περιθωρίου οι οποίοι γράφονται ως εξής:

$$\sum_{i=1}^{n} \xi_{i,S}$$

όπου $\xi_{i,S}$ είναι το χαλαρό περιθώριο της παρατήρησης i στο SVM, που είναι εκπαιδευμένο στις μεταβλητές με δείκτες στο S, το οποίο S είναι ένα υποσύνολο του {1,…,p}.

Γενικά (αλλά όχι πάντα) το $\sum \xi_{i,S}$ θα μειωθεί καθώς περισσότερες μεταβλητές προστίθενται. Ως εκ τούτου, θα πρέπει επίσης να προστεθεί ένας όρος ποινής που να σχετίζεται με τον αριθμό των μεταβλητών που περιλαμβάνονται στο μοντέλο. Προτείνεται η προσθήκη ενός γραμμικού όρου ποινής έτσι ώστε να προκύψει ένα κριτήριο πληροφορίας της μορφής

$$IC(S) = \sum_{i=1}^{n} \xi_{i,S} + C(n)|S|,$$

όπου το S είναι ένα σύνολο μεταβλητών που περιλαμβάνονται στο μοντέλο.

Η πρώτη επιλογή για το C(n) είναι C(n)=2, μια σταθερά. Αυτό οδηγεί στο πρώτο κριτήριο πληροφορίας για τα SVMs (SVMIC) :

SVMICα(S)= $\sum_{i=1}^{n} \xi_i + C(n)|S|$ (5.2.4.1)

Το SVMICα είναι μια εύκολα υπολογίσιμη προσέγγιση του KRIC (μέχρι σταθερούς παράγοντες) για το γραμμικό SVM.

Ως εκ τούτου, ο πρώτος όρος του KRIC μπορεί να προσεγγιστεί (μέχρι ένα σταθερό παράγοντα) με $\sum_i \ \xi_i$. Για την προσέγγιση του δεύτερου όρου, να σημειωθεί πρώτα ότι

$$(Q_S diag(t_S) + \lambda I_n)^{-1} Q_S (diag(m_S)^2 - n^{-1} m_S m_S^T)=$$

$$(Q_S diag(t_S) + \lambda I_n)^{-1} Q_S diag(t_S) diag(t_S)^{-1} (diag(m_S)^2 - n^{-1} m_S m_S^T)$$
$$\approx V diag(t_S)^{-1} (diag(m_S)^2 - n^{-1} m_S m_S^T),$$

με V=$(Q_S diag(t_S) + \lambda I_n)^{-1} Q_S diag(t_S)$ ένα θετικά ημιορισμένο πίνακα. Ο πίνακας V μπορεί στη συνέχεια να γραφτεί ως $V = U_1 L_1 U_1^{'}$, όπου ο U_1 είναι ένας ορθογώνιος πίνακας και L_1ένας διαγώνιος πίνακας που αποτελείται από τις ιδιοτιμές του V. Από τη στιγμή που ο Q_S είναι τάξεως $|S|$ για $|S| < n$, προκύπτει ότι

$$V = (Q_S diag(t_S) + \lambda I_n)^{-1} Q_S diag(t_S) \approx \left(Q_S diag(t_S)\right)^{-} Q_S diag(t_S)$$

Για λ μικρό, με A^- συμβολίζουμε το ψευδοαντίστροφο του πίνακα A.
Ως αποτέλεσμα,

$$L_1 \approx \begin{pmatrix} I_{|S|} & 0 \\ 0 & 0_{n-|S|} \end{pmatrix}$$

Ως εκ τούτου, προκύπτει ότι:

$$trace\left((Q_S diag(t_S) + \lambda I_n)^{-1} Q_S (diag(m_S)^2 - n^{-1} m_S m_S^T)\right)$$
$$\approx trace\left[U_1 \begin{pmatrix} I_{|S|} & 0 \\ 0 & 0_{n-|S|} \end{pmatrix} U_1^{'} diag(t_S)^{-1} (diag(m_S)^2 - n^{-1} m_S m_S^T)\right]$$
$$\approx \ trace\left[\begin{pmatrix} I_{|S|} & 0 \\ 0 & 0_{n-|S|} \end{pmatrix} U_1^{'} diag(t_S)^{-1} (diag(m_S)^2 - n^{-1} m_S m_S^T) U_1\right]$$

Ο πίνακας επιλογής του πρώτου παράγοντα υποδεικνύει ότι μόνο τα πρώτα $|S|$ διαγώνια στοιχεία προστίθενται. Ως εκ τούτου, η παραπάνω έκφραση μπορεί να προσεγγιστεί από τη σχέση

$$\frac{S}{n} trace\left(U_1^{'} diag(t_S)^{-1} (diag(m_S)^2 - n^{-1} m_S m_S^T) U_1\right)$$

ή από τη σχέση

$$\frac{S}{n} trace\left(diag(t_S)^{-1} (diag(m_S)^2 - n^{-1} m_S m_S^T)\right)$$

Δεδομένου ότι το ίχνος είναι αμετάβλητο σε ένα ορθογώνιο μετασχηματισμό ένας γρήγορος υπολογισμός αποκαλύπτει ότι το i-οστό διαγώνιο στοιχείο του πίνακα είναι ίσο με

$$\frac{n-1}{n} t_{i,S}^{-1} m_{i,S}^2 = \frac{n-1}{n} exp\left(-na_{i,S} y_i\right).$$

Για την περαιτέρω διευκόλυνση των υπολογισμών, αυτό αντικαθίσταται από το 1. Παρά το γεγονός ότι αυτός ο υπολογισμός θα μπορούσε να είναι αδρός για ένα και μόνο όρο, διαπιστώνεται εμπειρικά ότι λειτουργεί καλά για την άθροιση πάνω σε ολόκληρο το σύνολο εκπαίδευσης. Ως εκ τούτου, προκύπτει ότι

$$\frac{S}{n} trace\left(diag(t_S)^{-1}(diag(m_S)^2 - n^{-1} m_S m_S^T)\right) \approx \frac{|S|}{n} trace(I_n) = |S|$$

ο οποίος είναι ο γραμμικός όρος ποινής στο SVMICα. Το νέο προτεινόμενο κριτήριο SVMICα για τα SVMICα SVMs μοιράζεται τη μορφή της ποινής με το κριτήριο πληροφορίας ΑΚΑΙΚΕ (AIC) .Το AIC ορίζεται ως μείον δύο φορές την τιμή της μεγιστοποιημένης λογαριθμικής πιθανότητας του μοντέλου, συν δύο φορές τον αριθμό των προς εκτίμηση παραμέτρων (δηλαδή, $2|S|$). Επειδή η ποινή δεν εξαρτάται από το μέγεθος του δείγματος n , αναμένουμε ότι και τα δύο κριτήρια μοιράζονται ορισμένες ιδιότητες, όπως για παράδειγμα ότι έχουν την τάση να μην επιλέγουν το πιο φειδωλό μοντέλο.

Το δεύτερο προτεινόμενο κριτήριο ακολουθεί το πνεύμα του Μπεϋζιανού κριτηρίου πληροφορίας BIC. Αυτό το κριτήριο είναι παρόμοια ορισμένο με το AIC, αλλά αντί της ποινής $2|S|$, χρησιμοποιεί το $\log(n)|S|$.Το BIC έχει αποδειχθεί συνεπές, αυτό σημαίνει ότι αν το πραγματικό μοντέλο περιέχεται στη λίστα αναζήτησης, το κριτήριο (στο όριο για $n \to \infty$) επιλέγει το σωστό μοντέλο.

Για αυτό λαμβάνεται $C(n) = \log(n)$ και ορίζεται το δεύτερο κριτήριο:

$$SVMICb(S) = \sum_{i=1}^{n} \xi_i + \log(n)|S| \qquad (5.2.4.2)$$

Είναι αμέσως προφανές, ότι το υπολογιστικό κόστος και των δύο SVMICs είναι πολύ χαμηλότερο από αυτό του διασταυρωμένα επικυρωμένου ποσοστού σφάλματος (10 επιπλέον SVMs προς εκπαίδευση για τη 10-πλάσιας διασταυρωμένη επικύρωση) και από αυτό του KRIC (το οποίο χρειάζεται υπολογισμούς της τάξης $O(n^3)$ λόγω της αντιστροφής του πίνακα). Στην καλύτερη περίπτωση τα $\xi_{i,S}$ είναι άμεσα διαθέσιμα. Σε αυτή την περίπτωση, ο υπολογισμός των SVMICs είναι ένας υπολογισμός τάξης $O(n)$ και συνήθως ακόμα μικρότερης (τάξης) όταν εφαρμόζεται η ιδιότητα ότι

$$\xi_{i,S} \neq 0 \leftrightarrow a_{i,S} = 1.$$

Όταν μόνο τα a_S και Q_S είναι διαθέσιμα, τα $\xi_{i,S}$ πρέπει να υπολογιστούν με τη χρήση της σχέσης

$$\xi_{i,S} = \left[1 - y_i \sum_{\substack{j=1 \\ a_{j,S}>0}}^{n} a_{j,S}[Q_S]_{ij}\right]_+$$

Αυτό σημαίνει ότι στη χειρότερη περίπτωση, ο υπολογιστικός χρόνος των SVMICs είναι $O(n^2)$, που εξακολουθεί να είναι ταχύτερος από ότι χρησιμοποιώντας είτε το CV ποσοστό σφάλματος είτε το KRIC.

5.2.5 ΜΟΤΙΒΟ ΠΡΟΣΟΜΟΙΩΣΗΣ

Σε αυτό το πείραμα εκτελούνται M=100 τρεξίματα προσομοίωσης με τις ακόλουθες ρυθμίσεις. Παράγουμε $n \in \{25,50,100,200\}$ ανεξάρτητες παρατηρήσεις x_i , $1 \leq i \leq n$, διάστασης $p \in \{25,50,100,200\}$, με κατανομή $N(0, \sigma^2 I_p)$ όπου $\sigma^2 = 1$. Για κάθε παρατήρηση, παράγεται επίσης μία ετικέτα κλάσης $y_i \in \{-1,1\}$, με $P(y_i = 1) = \frac{1}{2}$.
Τελικά, ορίζονται το $\mu = \left(+\frac{1}{2}, -\frac{1}{2}, -\frac{1}{2}, +\frac{1}{2}, 0, \dots, 0\right)$ διάστασης p, και το $x_{i=}x_i + y_i\mu$ για το διαχωρισμό των δύο κλάσεων σε κάποιο βαθμό. Αυτό συνεπάγεται ότι το βέλτιστο διαχωριστικό υπερεπίπεδο θα μπορούσε να είναι x'μ=0 , τέτοιο ώστε $\hat{y} = +1$, αν x'μ>0. Αυτό το υπερεπίπεδο θα μπορούσε να δώσει ένα εκτός δείγματος ποσοστό εσφαλμένης ταξινόμησης του

$$P(Y = +1)P\left(X'\mu + \|\mu\|_2^2 < 0\right) + P(Y = -1)P\left(X'\mu - \|\mu\|_2^2 > 0\right)$$
$$= \frac{1}{2}P(\sigma\|\mu\|_2 Z < -\|\mu\|_2^2) + \frac{1}{2}P(\sigma\|\mu\|_2 Z > \|\mu\|_2^2)$$
$$= \Phi\left(-\frac{\|\mu\|_2}{\sigma}\right),$$

όπου το Z είναι μία τυπική κανονικά κατανεμημένη τυχαία μεταβλητή. Στο παράδειγμα, με σ=1 και $\|\mu\|_2 = 1$, βρίσκεται ένα βέλτιστο λάθος εσφαλμένης ταξινόμησης εκτός δείγματος.

Κατά τη διάρκεια κάθε τρεξίματος, πρώτα τυποποιούνται οι μεταβλητές για να βελτιωθεί η αριθμητική απόδοση του SVM αλγορίθμου. Η τυποποίηση είναι επίσης αναγκαία για το βήμα της ταξινόμησης μεταβλητών που έπεται. Στη συνέχεια, οι μεταβλητές ταξινομούνται με τη χρήση είτε του score του Fisher είτε της επιρροής μεταβλητής. Για κάθε ένα από τα εμφωλευμένα μοντέλα (nested models) που παράγονται στο βήμα της ταξινόμησης μεταβλητών, υπολογίζεται το SVMICa και το SVMICb. Συγκρίνονται οι αποδόσεις των δύο κριτηρίων με το κριτήριο του διασταυρωμένα επικυρωμένου ποσοστού σφάλματος (CV error rate), χρησιμοποιώντας ένα σχήμα 10-πλάσιας διασταυρωμένης επικύρωσης και το GRM του Vapnik. Θα χρησιμοποιηθεί η CV2 μέθοδος που περιλαμβάνει τη διαδικασία επιλογής χαρακτηριστικών στη διασταυρωμένη επικύρωση. Στη CV2 μέθοδο, η ταξινόμηση των μεταβλητών γίνεται σε κάθε τμήμα της 10-πλής διασταυρωμένης επικύρωσης και μπορεί να ποικίλει από το ένα μέρος στο άλλο. Τέλος συγκρίνονται τα νέα κριτήρια επιλογής με τα KRICs. Το κύριο μειονέκτημα αυτών των κριτηρίων είναι ο υπολογιστικός χρόνος όταν το μέγεθος δείγματος γίνεται μεγάλο. Σε αυτή την περίπτωση, γίνεται ακόμα πιο αργό το να υπολογιστεί από το διασταυρωμένα επικυρωμένο ποσοστό σφάλματος (CV error rate) και τα κριτήρια ελαχιστοποίησης γενικού κινδύνου (GRM criteria)

Το πείραμα θα επαναληφθεί δύο φορές, κάθε φορά με διαφορετικό πυρήνα για το SVM :

i) Γραμμικός πυρήνας: $K(x_1, x_2) = x_1' x_2$

ii) Τετραγωνικός πυρήνας: $K(x_1, x_2) = (\gamma x_1' x_2 + 1)^2$, με γ=$\frac{1}{p}$ (ο αντίστροφος του αριθμού των μεταβλητών)

Η παράμετρος ρύθμισης C σε κάθε SVM που εκπαιδεύεται, επιλέγεται να είναι C=1, όπως τυποποιούνται οι επεξηγηματικές μεταβλητές εκ των προτέρων. Τέλος, ελέγχεται η ακρίβεια των επιλεγμένων μοντέλων με την εκτίμηση των ποσοστών σφάλματος εκτός δείγματος, που βασίζεται σε ένα δείγμα δοκιμής 1000 παρατηρήσεων. Αυτές οι παρατηρήσεις παράγονται με τον ίδιο τρόπο όπως το δείγμα εκπαίδευσης.

Οι πίνακες (1) και (2) παραθέτουν τα εκτιμώμενα ποσοστά σφάλματος εκτός δείγματος. Για τον πρώτο πίνακα χρησιμοποιήθηκε ταξινόμηση μεταβλητών βασισμένη στην επίδραση των μεταβλητών εισόδου στο $||w||^2$. Ο δεύτερος πίνακας παρουσιάζει τα αποτελέσματα με τη χρήση των scores του Fisher ,S_j. Μια γενικότερη παρατήρηση είναι ότι τα κριτήρια επιλογής που βασίζονται στο ποσοστό σφάλματος (CV και GRM) έχουν τη χειρότερη απόδοση. Οι αποδόσεις των KRICs και των νέων SVMICs είναι συγκρίσιμες. Πιο συγκεκριμένα τα KRICs είναι καλύτερα σαν μέθοδος επιλογής μεταβλητών για μικρά μεγέθη δείγματος (n=25), ενώ τα SVMICs δίνουν καλύτερα αποτελέσματα για μεγαλύτερα μεγέθη δείγματος.

Η πρόσθετη υπολογιστική πολυπλοκότητα των κριτηρίων KRICs κατευθύνει την προτίμησή μας στα πολύ πιο απλά SVMICa και SVMICb. Για ένα μικρό αριθμό παρατηρήσεων σε σχέση με αριθμό μεταβλητών, το SVMICa υπερτερεί ελαφρώς του SVMICb όσον αφορά το ποσοστό σφάλματος εκτός δείγματος και το αντίστροφο ισχύει με πολλές παρατηρήσεις και λιγότερες μεταβλητές.

Οι διαφορές στα ποσοστά σφάλματος εκτός δείγματος γίνονται μικρότερες καθώς ο αριθμός των μεταβλητών αυξάνεται. Εντούτοις τα SVMICa και SVMICb έχουν το πλεονέκτημα να είναι πολύ ευκολότερο να υπολογιστούν (και λιγότερο χρονοβόρα) από τα άλλα κριτήρια, ειδικά τα KRICs καθώς ο υπολογιστικός τους χρόνος είναι τάξης $O(n^3)$. Επίσης καθώς το n μεγαλώνει, τα ποσοστά σφάλματος εκτός δείγματος των μοντέλων που λαμβάνονται από τα δύο προτεινόμενα κριτήρια συγκλίνουν προς το ελάχιστο θεωρητικά αποκτημένο ποσοστό σφάλματος εκτός δείγματος του 15,9%. Το ποια μοντέλα επιλέχτηκαν όντως από τα διάφορα κριτήρια και το γιατί το διασταυρωμένα επικυρωμένο ποσοστό σφάλματος αποδίδει τόσο σαν μοντέλο επιλογής κριτήριου, αναφέρονται στον πίνακα του ΣΧΗΜΑΤΟΣ 3 για λίγες συγκεκριμένες ρυθμίσεις.

Για κάθε ρύθμιση, ο Πίνακας στο ΣΧΗΜΑ 3 δείχνει πόσες φορές το σωστό υποσύνολο των μεταβλητών εισαγωγής επιλέχθηκε (C σωστή). Αυτός ο πίνακας δείχνει επίσης και πόσες φορές ένα πολύ αραιό σύνολο μεταβλητών επιλέχτηκε (U, υποπροσαρμογή) και πόσες φορές ένα πολύ πλούσιο σύνολο μεταβλητών επιλέχτηκε (O ,υπερπροσαρμογή). Η καλή απόδοση των SVMICa και SVMICb, μπορεί να οφείλεται στο γεγονός ότι αυτά τα κριτήρια φαίνεται να έχουν την τάση να επιλέγουν ένα σύνολο μεταβλητών που περιλαμβάνει όλες τις σημαντικές καθώς ο αριθμός των παρατηρήσεων μεγαλώνει. Τα αποτελέσματα της προσομοίωσης υποδεικνύουν ότι το SVMICa συμπεριφέρεται όπως το AIC με την τάση να υπερπροσαρμόζεται. Το SVIMCb φαίνεται να μοιράζεται την ιδιότητα του BIC, ότι επιλέγει το σωστό στοιχείο πιο αργά. Τα κριτήρια CV error rate και GRM φαίνεται να έχουν την τάση να αγνοούν τις μεταβλητές οι οποίες ωστόσο, είναι σημαντικές. Συνεπώς, τα μοντέλα που αυτά τα κριτήρια επιλέγουν είναι κακής ποιότητας πρόβλεψης.

Τα KRICs έχουν την ιδιότητα της υπερεπιλογής που παρουσίασε το SVMICa, αλλά τα KRICs επιλέγουν υπερβολικές μεταβλητές ακόμα πιο συχνά από το SVMIC. Αυτό έχει σαν αποτέλεσμα, τα αποτελέσματα για την περίπτωση των δύο πληθυσμών που προέρχονται από μια πανομοιότυπη κατανομή, διαφέροντας μόνο στο μέσο. Μία άλλη περίπτωση που εξετάστηκε είναι αυτή στην οποία οι διακυμάνσεις των δύο πληθυσμών διαφέρουν. Για παράδειγμα, όταν οι παρατηρήσεις στην κλάση +1, δοκιμάζονται από μία $N(\mu_1,\sigma_1)$ κατανομή και οι παρατηρήσεις στην κλάση -1, δοκιμάζονται από μία $N(\mu_2,\sigma_2)$ κατανομή. Διεξήχθη μία μελέτη προσομοίωσης με παρόμοιο τρόπο όπως η προηγούμενη, όπου τα δείγματα έχουν αναπτυχθεί από τη $N(\mu ,I_p)$ για την κλάση +1 και από τη $N(-2\mu, 4I_p)$, για την κλάση -1.

Τα αποτελέσματα αυτής της προσομοίωσης συνοψίζονται στους πίνακες των ΣΧΗΜΑΤΩΝ 4, 5 και 6. Σε αυτή τη ρύθμιση, παρατηρούνται παρόμοια αποτελέσματα όπως στη ρύθμιση όπου και οι δύο πληθυσμοί έχουν ίση διακύμανση. Η επιλογή που βασίζεται στα CV error rate και GRM αποδίδουν μάλλον άσχημα, ειδικά για τα μεγαλύτερα μεγέθη δειγμάτων. Όπως και πριν οι αποδόσεις των SVMICs και των KRICs είναι παρόμοιες. Πιο συγκεκριμένα, τα SVMICs έχουν μια βελτιωμένη απόδοση σε σχέση με τα KRICs όταν το μέγεθος δείγματος είναι μεγάλο (n≥ 50) και χρησιμοποιείται ο γραμμικός πυρήνας. Και τα KRICs λειτουργούν ελαφρά καλύτερα για μικρά μεγέθη δείγματος (n=25). Για τον τετραγωνικό πυρήνα παρατηρείται καλή απόδοση για τα KRICs το οποίο συνδυάζεται μόνο με το SVMICa για μεγαλύτερα μεγέθη δείγματος.

Ο πίνακας του ΣΧΗΜΑΤΟΣ 6 δείχνει τα μοντέλα που επιλέγονται από τα διάφορα κριτήρια. Από αυτόν τον πίνακα, γίνονται οι ίδιες παρατηρήσεις για τη συμπεριφορά των κριτηρίων, όπως στη ρύθμιση όπου οι πληθυσμοί έχουν ίσες διακυμάνσεις, όμως αυτή τη φορά μόνο ο γραμμικός πυρήνας χρησιμοποιείται για το SVM. Για τον πολυωνυμικό πυρήνα, κάνουμε τις ίδιες παρατηρήσεις για τη συμπεριφορά των κριτηρίων όπως στη ρύθμιση όπου οι πληθυσμοί έχουν ίσες διακυμάνσεις, όμως αυτή τη φορά μόνο όταν ο γραμμικός πυρήνας χρησιμοποιείται για το SVM. Για τον πολυωνυμικό πυρήνα, στη ρύθμιση με διαφορετικές πληθυσμιακές διακυμάνσεις, τα SVMICs έχουν περισσότερη δυσκολία επιλογής όλων των σχετικών μεταβλητών από τα KRICs, που εξηγεί γιατί αυτά τα κριτήρια έχουν βελτιωμένη απόδοση σε αυτή την περίπτωση.

Τελικά διεξήχθη ένα πείραμα προσομοίωσης όπου οι μεταβλητές εισόδου ήταν ισχυρά συσχετισμένες. Πρώτα οι παρατηρήσεις παρήχθησαν όπως στο πρώτο πείραμα προσομοίωσης. Έπειτα, εφαρμόστηκε ο μετασχηματισμός

$$x_{ij} = \rho x_{ik_j} + \varepsilon_{ij}$$

με $\varepsilon_{ij} \sim N(0, \rho^2)$ ανεξάρτητες και ισόνομες τυχαίες μεταβλητές, όπου i=1,…,n,k_j ,επιλέγονται αυθαίρετα μεταξύ 1 και 4 , και $4 < j \leq \frac{p}{2}$, έτσι ώστε περίπου οι μισές από τις μη σημαντικές μεταβλητές εισόδου να συσχετίζονται με τις 4 σημαντικές. Η παράμετρος $|\rho| < 1$ ελέγχει το βαθμό συσχέτισης. Επιλέχθηκε το ρ=0,8 και βρέθηκαν παρόμοια αποτελέσματα όπως στην περίπτωση που οι διακυμάνσεις και των δύο πληθυσμών κλάσεων διαφέρουν.

		linear				quadratic			
		$p = 25$	$p = 50$	$p = 100$	$p = 200$	$p = 25$	$p = 50$	$p = 100$	$p = 200$
$n = 25$	$SVMICa$	28.9%	33.9%	35.6%	36.5%	29.2%	35.1%	42.1%	50.1%
	$SVMICb$	30.1%	34.2%	35.7%	36.4%	31.8%	39.6%	48.2%	50.1%
	CV	30.4%	35.1%	36.0%	36.4%	31.8%	38.1%	42.2%	44.7%
	GRM	32.7%	35.3%	36.9%	36.6%	35.4%	42.8%	49.4%	50.1%
	KRIC	29.0%	32.7%	34.8%	36.4%	25.7%	30.5%	35.0%	38.9%
	KRICS	28.8%	32.5%	34.8%	36.1%	25.8%	31.3%	36.2%	40.4%
$n = 50$	$SVMICa$	23.3%	27.1%	28.3%		20.5%	23.1%	26.5%	
	$SVMICb$	23.9%	25.7%	27.4%		23.5%	26.1%	30.4%	
	CV	26.1%	27.7%	28.7%		25.9%	28.3%	34.5%	
	GRM	28.9%	29.1%	29.9%		30.6%	33.2%	40.5%	
	KRIC	24.2%	27.7%	28.4%		19.0%	23.8%	28.2%	
	KRICS	24.6%	27.6%	28.4%		19.5%	25.1%	30.1%	
$n = 100$	$SVMICa$	19.0%	21.8%			14.6%	17.9%		
	$SVMICb$	18.1%	19.3%			18.5%	18.4%		
	CV	22.7%	23.5%			20.8%	22.0%		
	GRM	27.6%	26.9%			27.8%	27.7%		
	KRIC	20.5%	24.8%			14.2%	18.1%		
	KRICS	21.0%	25.0%			14.5%	19.5%		
$n = 200$	$SVMICa$	17.0%				9.9%			
	$SVMICb$	15.9%				12.9%			
	CV	21.4%				19.6%			
	GRM	27.0%				29.3%			
	KRIC	17.9%				10.1%			
	KRICS	18.3%				9.7%			

ΣΧΗΜΑ 5.2.1 : Ποσοστά σφάλματος εκτός δείγματος για τις διαφορετικές ρυθμίσεις προσομοίωσης όπου η ταξινόμηση μεταβλητών έχει γίνει από την επιρροή μεταβλητής στο $\|w\|^2$.Ο αριστερός πίνακας δείχνει αυτά τα ποσοστά για τα SVMs με ένα γραμμικό πυρήνα και ο δεξιός πίνακας δείχνει αυτά τα ποσοστά για ένα τετραγωνικό πυρήνα. Η κορυφαία σειρά δείχνει τον αριθμό των μεταβλητών p , ενώ η αριστερότερη γραμμή δείχνει τον αριθμό των παρατηρήσεων n. Για κάθε ρύθμιση, 6 αριθμοί δίνονται, που αντιστοιχούν στα διαφορετικά κριτήρια επιλογής SVMICa , SVMICb , CV error rate, GRM και των KRICs που βασίζονται στο Μπεϋζιανό μοντέλο και Μπεϋζιανό μοντέλο του Sollich για τα SVMs.

		linear			quadratic		
		$p = 25$	$p = 50$	$p = 100$	$p = 25$	$p = 50$	$p = 100$
$n = 25$	$SVMICa$	29.4%	31.6%	33.9%	30.7%	35.3%	43.3%
	$SVMICb$	31.6%	32.6%	35.0%	33.8%	38.5%	48.4%
	CV	31.8%	33.5%	34.4%	32.9%	38.5%	42.7%
	GRM	34.5%	35.4%	35.7%	36.6%	42.6%	48.7%
	KRIC	29.0%	33.2%	34.9%	28.4%	33.0%	37.1%
	KRICS	29.9%	33.2%	34.9%	30.1%	34.1%	38.2%
$n = 50$	$SVMICa$	21.6%	23.3%	24.6%	21.3%	23.0%	25.6%
	$SVMICb$	23.2%	24.8%	25.0%	24.3%	26.8%	30.2%
	CV	25.5%	26.3%	28.0%	25.9%	28.1%	33.8%
	GRM	29.6%	30.5%	30.9%	31.7%	33.5%	40.1%
	KRIC	24.9%	28.7%	30.1%	22.5%	27.1%	30.9%
	KRICS	25.9%	29.7%	30.8%	25.1%	29.3%	32.8%
$n = 100$	$SVMICa$	18.5%	19.2%		18.5%	18.5%	
	$SVMICb$	18.9%	19.0%		19.1%	19.5%	
	CV	23.8%	25.4%		19.2%	22.0%	
	GRM	30.1%	29.6%		30.2%	30.7%	
	KRIC	20.6%	26.8%		20.0%	22.6%	
	KRICS	21.7%	27.8%		22.0%	25.1%	
$n = 200$	$SVMICa$	17.0%			20.3%		
	$SVMICb$	16.8%			16.8%		
	CV	21.5%			21.4%		
	GRM	29.3%			29.6%		
	KRIC	18.0%			18.1%		
	KRICS	18.9%			20.6%		

ΣΧΗΜΑ 5.2.2: πως στο ΣΧΗΜΑ 1 ,όμως αυτή τη φορά έχει γίνει με τη χρήση των scores του Fisher S_j.

Linear; $n = 25$; $p = 25$

	C	U	O	R
$SVMICa$	1	22	1	76
$SVMICb$	0	42	0	58
CV	0	38	4	58
GRM	0	77	0	23
KRIC	1	1	7	91
KRICS	0	0	9	91

Quadratic; $n = 25$; $p = 25$

	C	U	O	R
$SVMICa$	3	36	0	61
$SVMICb$	0	64	0	36
CV	1	40	5	54
GRM	0	75	0	25
KRIC	0	1	25	74
KRICS	0	0	49	51

Linear; $n = 200$; $p = 25$

	C	U	O	R
$SVMICa$	22	0	76	2
$SVMICb$	77	9	10	4
CV	7	48	43	2
GRM	1	98	1	0
KRIC	6	0	93	1
KRICS	1	0	99	0

Quadratic; $n = 200$; $p = 25$

	C	U	O	R
$SVMICa$	2	0	98	0
$SVMICb$	67	14	6	13
CV	4	43	49	4
GRM	1	99	0	0
KRIC	8	0	84	8
KRICS	0	0	100	0

Linear; $n = 25$; $p = 100$

	C	U	O	R
$SVMICa$	0	8	0	92
$SVMICb$	0	20	0	80
CV	0	23	6	71
GRM	0	56	0	44
KRIC	0	1	0	99
KRICS	0	0	1	99

Quadratic; $n = 25$; $p = 100$

	C	U	O	R
$SVMICa$	0	35	0	65
$SVMICb$	0	63	0	37
CV	0	33	10	57
GRM	0	64	0	36
KRIC	0	0	41	59
KRICS	0	0	56	44

ΣΧΗΜΑ 5.2.3 : Συχνότητες των οποίων μοντέλων έχουν επιλεχθεί από τα SVMICa , SVMICb ,το CV error rate, το GRM, το KRIC και τα KRICs για κάθε ρύθμιση προσομοίωσης. Οι πίνακες δίνουν τα αποτελέσματα όπου οι μεταβλητές ταξινομούνται από την επίδραση της μεταβλητής στο $\|w\|^2$. Η στήλη που ονομάζεται C (σωστό), δηλώνει τον αριθμό των φορών που οι τέσσερις σημαντικές μεταβλητές επιλέχθηκαν, χωρίς άλλες. Η στήλη που ονομάζεται U (υποπροσαρμογής) δίνει τον αριθμό των φορών, που όχι όλες οι σημαντικές μεταβλητές και όχι άλλες, επιλέχθηκαν, ενώ η στήλη O (υποπροσαρμογής) δίνει τον αριθμό των φορών που όλες οι σημαντικές μεταβλητές και τουλάχιστον μία μη σημαντική μεταβλητή επιλέχθηκε. Η τελευταία στήλη που ονομάζεται R, αναφέρει τον αριθμό των φορών που δε συνέβη καμιά από τις 3 προηγούμενες καταστάσεις.

		linear				quadratic			
		$p = 25$	$p = 50$	$p = 100$	$p = 200$	$p = 25$	$p = 50$	$p = 100$	$p = 200$
$n = 25$	$SVMICa$	28.9%	33.9%	35.6%	36.5%	29.2%	35.1%	42.1%	50.1%
	$SVMICb$	30.1%	34.2%	35.7%	36.4%	31.8%	39.6%	48.2%	50.1%
	CV	30.4%	35.1%	36.0%	36.4%	31.8%	38.1%	42.2%	44.7%
	GRM	32.7%	35.3%	36.9%	36.6%	35.4%	42.8%	49.4%	50.1%
	KRIC	29.0%	32.7%	34.8%	36.4%	25.7%	30.5%	35.0%	38.9%
	KRICS	28.8%	32.5%	34.8%	36.1%	25.8%	31.3%	36.2%	40.4%
$n = 50$	$SVMICa$	23.3%	27.1%	28.3%		20.5%	23.1%	26.5%	
	$SVMICb$	23.9%	25.7%	27.4%		23.5%	26.1%	30.4%	
	CV	26.1%	27.7%	28.7%		25.9%	28.3%	34.5%	
	GRM	28.9%	29.1%	29.9%		30.6%	33.2%	40.5%	
	KRIC	24.2%	27.7%	28.4%		19.0%	23.8%	28.2%	
	KRICS	24.6%	27.6%	28.4%		19.5%	25.1%	30.1%	
$n = 100$	$SVMICa$	19.0%	21.8%			14.6%	17.9%		
	$SVMICb$	18.1%	19.3%			18.5%	18.4%		
	CV	22.7%	23.5%			20.8%	22.0%		
	GRM	27.6%	26.9%			27.8%	27.7%		
	KRIC	20.5%	24.8%			14.2%	18.1%		
	KRICS	21.0%	25.0%			14.5%	19.5%		
$n = 200$	$SVMICa$	17.0%				9.9%			
	$SVMICb$	15.9%				12.9%			
	CV	21.4%				19.6%			
	GRM	27.0%				29.3%			
	KRIC	17.9%				10.1%			
	KRICS	18.3%				9.7%			

ΣΧΗΜΑ 5.2.4 : Τα ποσοστά σφάλματος εκτός δείγματος για τις διάφορες ρυθμίσεις προσομοίωσης, με τους άνισους πληθυσμούς διακύμανσης, όπου η κατάταξη μεταβλητών έχει γίνει από την επιρροή μεταβλητής στο $\|w\|^2$. Ο αριστερός πίνακας δείχνει αυτά τα ποσοστά για τα SVMs με ένα γραμμικό πυρήνα και ο δεξιός πίνακας δείχνει αυτά τα ποσοστά για ένα τετραγωνικό πυρήνα. Οι κορυφαίες γραμμές δείχνουν τον αριθμό των μεταβλητών p , ενώ οι αριστερές γραμμές δείχνουν τον αριθμό των παρατηρήσεων n. Για κάθε ρύθμιση, 6 αριθμοί δίνονται, που αντιστοιχούν στα διαφορετικά κριτήρια επιλογής SVMICa , SVMICb , CV error rate, GRM και των KRICs που βασίζονται στο Μπεϋζιανό μοντέλο και Μπεϋζιανό μοντέλο του Sollich για τα SVMs.

		linear				quadratic			
		$p = 25$	$p = 50$	$p = 100$	$p = 200$	$p = 25$	$p = 50$	$p = 100$	$p = 200$
$n = 25$	$SVMICa$	28.0%	30.2%	31.5%	33.2%	28.9%	35.8%	41.7%	50.1%
	$SVMICb$	29.2%	31.3%	32.3%	34.4%	31.8%	40.0%	48.1%	50.1%
	CV	28.4%	31.4%	32.6%	34.2%	28.7%	37.6%	42.3%	44.4%
	GRM	31.6%	33.1%	33.7%	35.6%	34.7%	42.4%	48.7%	50.1%
	KRIC	27.5%	30.7%	32.6%	33.5%	24.9%	30.8%	36.0%	40.0%
	KRICS	27.7%	30.5%	33.0%	33.7%	26.2%	32.3%	38.1%	41.8%
$n = 50$	$SVMICa$	20.5%	21.7%	23.1%		19.3%	22.2%	25.8%	
	$SVMICb$	21.9%	22.7%	23.7%		22.2%	26.2%	30.4%	
	CV	24.9%	25.2%	25.2%		24.5%	27.6%	33.7%	
	GRM	28.6%	28.4%	28.7%		30.2%	32.7%	40.4%	
	KRIC	23.6%	26.8%	26.7%		19.1%	23.9%	28.8%	
	KRICS	24.3%	27.1%	27.5%		19.9%	26.1%	32.3%	
$n = 100$	$SVMICa$	17.4%	17.8%			15.2%	17.0%		
	$SVMICb$	17.4%	18.0%			16.4%	17.8%		
	CV	21.5%	22.7%			19.9%	21.5%		
	GRM	27.6%	27.0%			27.1%	28.3%		
	KRIC	20.0%	25.0%			14.5%	18.5%		
	KRICS	20.9%	25.5%			14.9%	20.3%		
$n = 200$	$SVMICa$	16.1%				9.8%			
	$SVMICb$	15.6%				13.2%			
	CV	20.7%				17.6%			
	GRM	27.0%				26.8%			
	KRIC	17.0%				10.3%			
	KRICS	17.8%				9.8%			

ΣΧΗΜΑ 5.2.5 : Όπως στον πίνακα 5.2.4, όμως αυτή τη φορά, η κατάταξη των μεταβλητών έχει γίνει με τη χρήση των scores του Fisher S_j.

Linear; $n = 25$; $p = 25$

	C	U	O	R
$SVMICa$	0	22	1	77
$SVMICb$	0	47	0	53
CV	1	40	1	58
GRM	0	76	0	24
KRIC	0	0	6	94
KRICS	0	0	8	92

Quadratic; $n = 25$; $p = 25$

	C	U	O	R
$SVMICa$	1	36	0	63
$SVMICb$	1	57	0	42
CV	1	39	8	52
GRM	0	70	0	30
KRIC	0	0	25	75
KRICS	0	0	50	50

Linear; $n = 200$; $p = 25$

	C	U	O	R
$SVMICa$	11	0	85	4
$SVMICb$	69	10	16	5
CV	6	56	37	1
GRM	0	100	0	0
KRIC	5	0	93	2
KRICS	0	0	99	1

Quadratic; $n = 200$; $p = 25$

	C	U	O	R
$SVMICa$	0	20	0	80
$SVMICb$	0	45	0	55
CV	0	33	4	63
GRM	0	56	0	44
KRIC	0	0	40	60
KRICS	0	0	53	47

Linear; $n = 25$; $p = 200$

	C	U	O	R
$SVMICa$	0	1	0	99
$SVMICb$	0	8	0	92
CV	0	22	2	76
GRM	0	46	0	54
KRIC	0	1	0	99
KRICS	0	0	0	100

Quadratic; $n = 25$; $p = 200$

	C	U	O	R
$SVMICa$	0	52	0	48
$SVMICb$	0	54	0	46
CV	0	22	5	73
GRM	0	54	0	46
KRIC	0	0	46	54
KRICS	0	0	56	44

ΣΧΗΜΑ 5.2.6 : Οι ακολουθίες των οποίων τα μοντέλα έχουν επιλεχθεί από τα SVMICa , SVMICb ,το CV error rate, το GRM, το KRIC και τα KRICs για κάθε ρύθμιση προσομοίωσης με άνισες διακυμάνσεις πληθυσμών. Οι πίνακες δίνουν τα αποτελέσματα όταν οι μεταβλητές ταξινομούνται με τη μεταβλητή επίδραση στο $\|w\|^2$. Η στήλη που ονομάζεται C (σωστό), δηλώνει τον αριθμό των φορών που οι τέσσερις σημαντικές μεταβλητές επιλέχθηκαν, χωρίς άλλες. Η στήλη που ονομάζεται U (υποπροσαρμογής) δίνει τον αριθμό των φορών, που όχι όλες οι σημαντικές μεταβλητές και όχι άλλες, επιλέχθηκαν, ενώ η στήλη O (υποπροσαρμογής) δίνει τον αριθμό των φορών που όλες οι σημαντικές μεταβλητές και τουλάχιστον μία μη σημαντική μεταβλητή επιλέχθηκε. Η τελευταία στήλη που ονομάζεται R, αναφέρει τον αριθμό των φορών που δε συνέβη καμιά από τις 3 προηγούμενες καταστάσεις.

5.2.6 ΔΟΚΙΜΕΣ ΣΕ ΠΡΑΓΜΑΤΙΚΑ ΔΕΔΟΜΕΝΑ

Γίνεται σύγκριση της απόδοσης των νέων μεθόδων, με εκείνη των άλλων κριτηρίων σε αρκετά σύνολα δεδομένων του πραγματικού κόσμου.

Τα σύνολα δεδομένων που χρησιμοποιούνται είναι οι εξής βάσεις δεδομένων: η Pina Indians Diabetes (768 παρατηρήσεις, 8 μεταβλητές), η Statlog Cleveland Heart Disease (303 παρατηρήσεις, 14 μεταβλητές) ,η Leo Breiman riingnorm και twonorm σύνολα δεδομένων (και τα δύο 7400 παρατηρήσεων και 20 μεταβλητών). Διεξάγονται 100 τυχαίοι διαχωρισμοί

των δεδομένων σε ένα δείγμα εκπαίδευσης και σε ένα δείγμα δοκιμής, όπου το μέγεθος του δείγματος εκπαίδευσης επιλέγεται ως $\sqrt{2n}$, με n το συνολικό αριθμό παρατηρήσεων στο σύνολο δεδομένων. Επιλέχθηκε το μέγεθος του συνόλου εκπαίδευσης τέτοιο ώστε να υπάρχει μια επαρκής ποσότητα παρατηρήσεων στο δείγμα δοκιμής για την εκτίμηση του ποσοστού σφάλματος εκτός δείγματος. Αντίθετα επιλέχθηκε ένα λογικά μικρό δείγμα δοκιμής για να διατηρήσει το υπολογιστικό χρόνο λογικό, διότι η εκπαίδευση ενός SVM και ειδικά ο υπολογισμός του KRIC, στη συνέχεια παίρνει όλο και περισσότερο χρόνο για μεγαλύτερα μεγέθη δείγματος .

Για κάθε μία από αυτές τις διχοτομήσεις, εκτελούμε το σχήμα επιλογής μεταβλητών στο δείγμα εκπαίδευσης ακριβώς όπως και στη μελέτη προσομοίωσης. Πρώτα γίνεται ταξινόμηση των μεταβλητών, για τη διατήρηση των p σωρευμένων υποσυνόλων των μεταβλητών εισόδου και στη συνέχεια χρήση των κριτηρίων πληροφορίας για την επιλογή των μεταβλητών που εξηγούν καλύτερα τα δεδομένα εκπαίδευσης. Έπειτα, γίνεται πρόβλεψη των ετικετών κλάσης για το δείγμα δοκιμής και χρήση αυτών των προβλέψεων για την εκτίμηση του ποσοστού σφάλματος εκτός δείγματος.

Τα εκτιμώμενα ποσοστά σφάλματος εκτός δείγματος παρουσιάζονται στον πίνακα του ΣΧΗΜΑΤΟΣ 5.2.7 για κάθε σύνολο δεδομένων και ρύθμιση εκπαίδευσης. Παρατηρείται ότι τα KRICs είναι η προτιμώμενη επιλογή κριτηρίου επιλογής μεταβλητής όσον αφορά το ποσοστό σφάλματος εκτός δείγματος για τα ‘’two norm ‘’ και ‘’heart’’ σύνολα δεδομένων. Για τα ‘’ringnorm’’ και ‘’diabetes’’ σύνολα δεδομένων η διαφορά στην απόδοση μεταξύ των KRICs και των νεοπροταθέντων SVMICs είναι λιγότερο σαφής. Παρατηρείται επίσης ότι το CV error rate και ειδικά το GRM έχουν κακή απόδοση, το οποίο είναι σε συμφωνία με τα αποτελέσματα πού παίρνουμε από την προσομοίωση.

Από αυτά τα αποτελέσματα καθώς και από αυτά που επιτεύχθηκαν στην προηγούμενη ενότητα, προτείνεται η χρήση είτε του SVMICa είτε του SVMICb, αν μία προκαταρκτική ανάλυση των δεδομένων ή μια εκ των προτέρων γνώση δείχνει ότι η πραγματική συνάρτηση απόφασης είναι σχεδόν γραμμική. Όταν διαφέρει σημαντικά από μια γραμμική συνάρτηση, υπάρχει δυνατότητα επιλογής μεταξύ της ευκολίας υπολογισμού των κριτηρίων πληροφορίας SVMs ή της κάπως βελτιωμένης προγνωστικής απόδοσης του KRIC, αν και με υψηλότερο υπολογιστικό κόστος.

Linear; $n = 25$; $p = 25$

	C	U	O	R
$SVMICa$	0	22	1	77
$SVMICb$	0	47	0	53
CV	1	40	1	58
GRM	0	76	0	24
KRIC	0	0	6	94
KRICS	0	0	8	92

Quadratic; $n = 25$; $p = 25$

	C	U	O	R
$SVMICa$	1	36	0	63
$SVMICb$	1	57	0	42
CV	1	39	8	52
GRM	0	70	0	30
KRIC	0	0	25	75
KRICS	0	0	50	50

Linear; $n = 200$; $p = 25$

	C	U	O	R
$SVMICa$	11	0	85	4
$SVMICb$	69	10	16	5
CV	6	56	37	1
GRM	0	100	0	0
KRIC	5	0	93	2
KRICS	0	0	99	1

Quadratic; $n = 200$; $p = 25$

	C	U	O	R
$SVMICa$	0	20	0	80
$SVMICb$	0	45	0	55
CV	0	33	4	63
GRM	0	56	0	44
KRIC	0	0	40	60
KRICS	0	0	53	47

Linear; $n = 25$; $p = 200$

	C	U	O	R
$SVMICa$	0	1	0	99
$SVMICb$	0	8	0	92
CV	0	22	2	76
GRM	0	46	0	54
KRIC	0	1	0	99
KRICS	0	0	0	100

Quadratic; $n = 25$; $p = 200$

	C	U	O	R
$SVMICa$	0	52	0	48
$SVMICb$	0	54	0	46
CV	0	22	5	73
GRM	0	54	0	46
KRIC	0	0	46	54
KRICS	0	0	56	44

ΣΧΗΜΑ 5.2.7: Τα ποσοστά σφάλματος εκτός δείγματος που λαμβάνονται από τα διαφορετικά σύνολα δεδομένων. Το μεταβλητό σύστημα κατάταξης και ο πυρήνας που χρησιμοποιείται δηλώνεται στο άνω αριστερό κελί κάθε πίνακα. Οι κορυφαίες γραμμές δείχνουν το σύνολο δεδομένων που χρησιμοποιήθηκε, ενώ οι αριστερότερες γραμμές δείχνουν τα κριτήρια επιλογής μοντέλου: SVMICa , SVMICb , CV error rate, GRM και των KRICs που βασίζονται στο Μπεϋζιανό μοντέλο και Μπεϋζιανό μοντέλο του Sollich για τα SVMs.

6

ΚΡΙΤΗΡΙΑ ΠΛΗΡΟΦΟΡΙΑΣ ΓΙΑ ΕΠΙΛΟΓΗ ΜΟΝΤΕΛΟΥ ΣΤΑ MULTICLASS SVMS

6.1 ΕΙΣΑΓΩΓΗ

Όπως η επιλογή μοντέλου για τα δυαδικά SVMs έτσι και η ομόλογή της για τα multiclass χρησιμοποιείται για την επιλογή των παραμέτρων της συνάρτησης πυρήνα και της παραμέτρου ρύθμισης που εξισορροπεί το σφάλμα εκπαίδευσης και τη μηχανική πολυπλοκότητα. Πολύ συχνά, ένα μοναδικό σύνολο παραμέτρων του μοντέλου χρησιμοποιείται ομοιόμορφα σε όλους τους συμπεριλαμβανόμενους ταξινομητές. Αυτό προτιμάται εξαιτίας του ότι πολύ λιγότερες παράμετροι μοντέλου χρειάζεται να θεωρηθούν και του ότι ο κίνδυνος υπερπροσαρμογής μειώνεται με τη χρήση ενός απλούστερου μοντέλου. Ως εκ τούτου ο στόχος μας είναι η επιλογή μοντέλου για τα multiclass SVMs με την εύρεση του καλύτερου μοναδικού συνόλου παραμέτρων. Στο συγκεκριμένο κεφάλαιο, η επιλογή μοντέλου για τα multiclass SVMs χρησιμοποιεί μια εξαντλητική μέθοδο αναζήτησης πλέγματος (exhaustive grid-based search method). Το κριτήριο είναι το k-πλάσιο ή η leave-one-out ποσοστό σφάλματος της διασταυρωμένης επικύρωσης.

Η προτεινόμενη προσέγγιση είναι η ανάπτυξη νέων κριτηρίων βασισμένων στις αρχές των επιτυχημένων κριτηρίων στα δυαδικά SVMs για multiclass ρύθμιση. Στην επιλογή μοντέλου για τα δυαδικά SVMs, μια κλάση μεθόδων χρησιμοποιεί τεχνικές μη γραμμικής βελτιστοποίησης για την ελαχιστοποίηση ή τη μεγιστοποίηση ενός βέλτιστου συνόλου παραμέτρων μοντέλου. Μια σημαντική πρόοδος προς αυτή την κατεύθυνση είναι η μέθοδος ελαχιστοποίησης του φράγματος ακτινωτού περιθωρίου (radius-margin bound) για τον ταξινομητή δυαδικού SVM. Λεπτομερώς, το φράγμα ακτινωτού περιθωρίου για τα δυαδικά SVMs χρησιμοποιείται με του εξής δύο τρόπους: 1) συνδέοντας τα ποσοστά σφάλματος δοκιμής για τους δυαδικούς και τους multiclass SVMs ταξινομητές (το πρώτο κριτήριο αναπτύσσεται βασισμένο στο ζευγαρωτό συνδυασμό μεταξύ των φραγμάτων ακτινωτού περιθωρίου και ενός συνόλου δυαδικών SVMs για την επιλογή μοντέλου) και 2) εμπνευσμένο από τη σχέση μεταξύ του φράγματος ακτινωτού περιθωρίου και του μέτρου διαχωριστικότητας κλάσεων, το δεύτερο κριτήριο ορίζει μια νέα ακτίνα και περιθώριο για να συμβιβάσουν κλάσεις. Τα δύο κριτήρια επιτρέπουν για αποτελεσματική βελτιστοποίηση για μερικές εκατοντάδες παραμέτρων μοντέλου ταυτόχρονα.

6.2 ΑΠΑΙΤΟΥΜΕΝΕΣ ΕΝΝΟΙΕΣ

6.2.1 ΠΡΟΣΕΓΓΙΣΗ ΜΕΜΟΝΩΜΕΝΗΣ ΜΗΧΑΝΗΣ (SINGLE-MACHINE APPROACH)

Οι προσεγγίσεις μεμονωμένης μηχανής προσπαθούν να επιλύσουν ένα μεμονωμένο πρόβλημα βελτιστοποίησης που εκπαιδεύει ταυτόχρονα πολλούς δυαδικούς ταξινομητές. Στη συνέχεια θα παρουσιαστούν τρεις προσεγγίσεις μεμονωμένης μηχανής.

1. Weston, Watkins and Vapnik

$$\min_{f_1,\dots,f_N \in H, \xi \in R^{l(N-1)}} \sum_{i=1}^{N} \|f_i\|_K^2 + C \sum_{i=1}^{l} \sum_{j \neq y_i} \xi_{ij} \qquad (6.2.1.1)$$

Υπόκειται σε: $f_{y_i}(x_i) + b_{y_i} \geq f_j(x_i) + b_j + 2 - \xi_{ij}$ (6.2.1.2)

$$\xi_{ij} \geq 0$$

Υποθέτουμε ότι το σημείο i είναι στην κλάση y_i. Τότε για j≠ y_i θέλουμε:
$f_{y_i}(x_i) - f_j(x_i) \geq 2$

ή δίνουμε μια γραμμική ποινή ξ_{ij}.

2. Lee, Lin and Wahba

$$\min_{\xi_1,\dots,\xi_N \in H_K} \frac{1}{l} \sum_{i=1}^{l} \sum_{j=1, j \neq y_i}^{N} \left(\left(f_j(x_i) + \frac{1}{N-1} \right) \right)_+ + \lambda \sum_{j=1}^{N} \left\| f_j \right\|_K^2 \qquad (6.2.1.3)$$

υπόκειται σε $\sum_{j=i}^{N} f_j(x)$ ∀x (6.2.1.4)

3. Crammer and Singer

$$\min_{\xi_1,\dots,\xi_N \in R^l} \sum_{i=1}^{N} \|\xi_\iota\|_K^2 + C \sum_{i=1}^{N} \sum_{j \neq \neq y_i} \xi_\iota \qquad (6.2.1.5)$$

υπόκειται σε $f_{y_i}(x_i) \geq f_j(x_i) + 1 - \xi_\iota$ (6.2.1.6)

Αυτή η διατύπωση είναι μια τροποποιημένη εκδοχή της (1) και η διαφορά τους είναι ότι σε αυτή, υπάρχει μία μόνο χαλαρή μεταβλητή ξ_ι ανά σημείο δεδομένων, αντί για N-1 χαλαρές ανά σημείο δεδομένων όπως στην (1). Αντί να δίνουμε μια γραμμική ποινή για κάθε κλάση $j \neq i$, για την οποία $f_i(x) < f_j(x) + 1$, δίνουμε μόνο για τη μεγαλύτερη $f_j(x)$.

6.2.2 ΜΕΘΟΔΟΙ ΕΡΕΥΝΑΣ ΒΑΣΙΣΜΕΝΕΣ ΣΤΟ ΠΛΕΓΜΑ (GRID-BASED SEARCH METHODS)

Όταν το γενικό πρόβλημα βελτιστοποίησης περιλαμβάνει συνεχείς μεταβλητές, υπάρχει ένας άπειρος αριθμός σημείων στο πεδίο και η πλήρης απαρίθμηση είναι αδύνατη. Μια κοινή προσέγγιση είναι να διεξάγουμε μία έρευνα πλέγματος, ουσιαστικά

διακριτοποιώντας το πεδίο. Μια ανάλυση πλέγματος δημιουργεί ισαπέχοντα σημεία πλέγματος στην εφικτή περιοχή και αξιολογεί την αντικειμενική συνάρτηση σε κάθε σημείο. Εάν η αντικειμενική συνάρτηση ικανοποιεί τη συνθήκη Lipschitz με σταθερά K και το πεδίο είναι ένα υπερ-ορθογώνιο παραλληλόγραμμο μέγιστου μήκους D σε κάθε πλευρά, τότε η απόσταση του πλέγματος μπορεί να προσδιοριστεί έτσι ώστε να επιτευχθεί μια επιθυμητή ακρίβεια, που να έχει την εκτίμηση εντός ϵ από το ολικό βέλτιστο. Αν η απόσταση μεταξύ των σημείων πλέγματος είναι ϵ/K για κάθε συντεταγμένη, τότε υπάρχουν προσεγγιστικά $(KD/\epsilon)^n$ σημεία πλέγματος και το πιθανό σφάλμα μεταξύ γειτονικών σημείων φράσσεται από το ϵ. Ως εκ τούτου, ο αριθμός των αξιολογήσεων που χρειάζεται να έχουν ακρίβεια ϵ είναι αναλογικός του $(KD/\epsilon)^n$. Ως εκ τούτου, ο αριθμός των αξιολογήσεων συνάρτησης για να πετύχουμε ακρίβεια ϵ είναι εκθετικός στη n-οστή διάσταση.

6.2.3 ΜΕΘΟΔΟΙ ΒΕΛΤΙΣΤΟΠΟΙΗΣΗΣ ΒΑΣΙΣΜΕΝΕΣ ΣΤΗΝ ΚΛΙΣΗ (GRADIENT-BASED OPTIMIZATION TECHNIQUES)

Η γενική μέθοδος για αυτή την κατηγορία μεθόδων είναι η εξής:

- Δεδομένης μιας $y = f(x)$, ορίζουμε ένα σημείο x_0 και υπολογίζουμε την κλίση $\nabla f(x_0)$
- Κινούμαστε κατά μήκος της κλίσης για να πάρουμε $x_1 = x_0 + a\nabla f(x_0)$
- Επαναλαμβάνουμε, έως ότου ληφθεί το ακρότατο

Η συγκεκριμένη μέθοδος απαιτεί την ύπαρξη παραγώγων και εν γένει ''κολλά'' εύκολα σε τοπικά ακρότατα.

6.2.4 BROYDEN-FLETCHER-GOLDFARB-SHANNO (BFGS) QUASI-NEWTON ΜΕΘΟΔΟΣ

Η BFGS ενημέρωση είναι η εξής

$$H_k = H_{k-1} + \frac{yy^T}{y^T s} - \frac{H_{k-1} s s^T H_{k-1}}{s^T H_{k-1} s} \qquad (6.2.4.1)$$

όπου $s = x^{(k)} - x^{(k-1)}$ και $y = \nabla f\left(x^{(k)}\right) - \nabla f\left(x^{(k-1)}\right)$

Η αντίστροφη ενημέρωση είναι η εξής:

$$H_k^{-1} = \left(I - \frac{sy^T}{y^T s}\right) H_{k-1}^{-1} \left(I - \frac{ys^T}{y^T s}\right) + \frac{ss^T}{y^T s} \qquad (6.2.4.2)$$

Να σημειωθεί ότι $y^T s > 0$ για αυστηρή κυρτή f. Το κόστος της ενημέρωσης ή της αντίστροφης ενημέρωσης είναι $O(n^2)$ λειτουργίες.

Στη συνέχεια θα παρουσιαστεί ο αλγόριθμος της BFGS quasi-Newton μεθόδου, του οποίου το κόστος ανά επανάληψη είναι $O(n^2)$ αριθμητικές λειτουργίες, αξιολόγηση συναρτήσεων και κλίσης, ενώ το ποσοστό σύγκλισης είναι υπερ-γραμμικό.

Given starting point $\bar{\mathbf{x}}_0$, convergence tolerance $\epsilon > 0$, and initial inverse Hessian approximation H_0:

```
k = 0
while( ||∇f(x̄_k)|| > ε )
```

$$\bar{\mathbf{p}}_k = -H_k \nabla f(\bar{\mathbf{x}}_k)$$
$$\bar{\mathbf{x}}_{k+1} = \mathtt{linesearch}(\bar{\mathbf{p}}_k, \ldots)$$
$$\bar{\mathbf{s}}_k = \bar{\mathbf{x}}_{k+1} - \bar{\mathbf{x}}_k$$
$$\bar{\mathbf{y}}_k = \nabla f(\bar{\mathbf{x}}_{k+1}) - \nabla f(\bar{\mathbf{x}}_k)$$
$$\rho_k = \frac{1}{\bar{\mathbf{y}}_k^T \bar{\mathbf{s}}_k}$$
$$H_{k+1} = \left(I - \rho_k \bar{\mathbf{s}}_k \bar{\mathbf{y}}_k^T\right) H_k \left(I - \rho_k \bar{\mathbf{y}}_k \bar{\mathbf{s}}_k^T\right) + \rho_k \bar{\mathbf{s}}_k \bar{\mathbf{s}}_k^T$$

```
  k = k + 1
end-while
```

Σχήμα 6.2.1: Αλγόριθμος της BFGS quasi-Newton μεθόδου

6.3 ΦΡΑΓΜΑ ΑΚΤΙΝΩΤΟΥ ΠΕΡΙΘΩΡΙΟΥ ΓΙΑ ΔΥΑΔΙΚΑ SVMS.

Έστω $L(D)$ ο αριθμός των σφαλμάτων σε μια leave-one-out διαδικασία που διεξάγεται στο D. Το φράγμα ακτινωτού περιθωρίου είναι ένα άνω φράγμα του $L(D)$. Για ένα δυαδικό SVM σκληρού περιθωρίου, δείχνεται ότι:

$$L(D) \leq \frac{4R^2}{\gamma^2} = 4R^2\|w\|^2 \qquad (6.3.1)$$

όπου R είναι η ακτίνα της μικρότερης σφαίρας που περικλείει l δείγματα εκπαίδευσης στο F, το γ είναι το περιθώριο, το w είναι το τυπικό διάνυσμα του βέλτιστου διαχωριστικού επιπέδου και το $\gamma = 1/\|w\|$. Για τα μη διαχωρίσιμα δεδομένα, ένα L2-norm SVM χαλαρού περιθωρίου χρησιμοποιείται. Η τετραγωνική ακτίνα εκφράζεται $R^2 = \min_{\|\varphi(x_i)-\hat{c}\|^2 \leq \hat{R}^2}(\hat{R}^2)$, όπου $\varphi(x_i)$ $(i = 1, \ldots, l)$ είναι η εικόνα του x_i $\sigma\tau o$ $F, \hat{R}$ είναι η ακτίνα της σφαίρας που περικλείει όλα τα $\varphi(x_i)$ και $\hat{c}$ το κέντρο της σφαίρας. Αυτό οδηγεί σε ένα τετραγωνικό πρόβλημα βελτιστοποίησης , και μπορεί να βρεθεί ότι:

$$R^2 = \max_{\beta \in R^l}\left[\sum_{i=1}^{l} \beta_i k(x_i, x_j) - \sum_{i,j=1}^{l} \beta_i \beta_j k(x_i, x_j)\right]$$

υπό τους περιορισμούς : $\sum_{i=1}^{l} \beta_i = 1; \quad \beta_i \geq 0 \ (i = 1,2, \ldots, l)$ (6.3.2)

όπου β_i είναι ο i-οστός πολλαπλασιαστής Lagrange και το κέντρο της σφαίρας αναπαρίσταται ως $\hat{c} = \sum_{i=1}^{l} \beta_i \varphi(x_i)$. Όσο για το $\|w\|^2$, μπορεί να βρεθεί από τη στιγμή που λύνεται το πρόβλημα βελτιστοποίησης SVM. Λεπτομερώς έχουμε ότι

$$\frac{1}{2}\|w\|^2 = \max_{a \in R^l}\left[\sum_{i=1}^{l} a_i - \frac{1}{2}\sum_{i,j=1}^{l} a_i a_j y_i y_j k(x_i, x_j)\right]$$

υπό τους περιορισμούς: $\sum_{i=1}^{l} \alpha_i y_i = 0; \quad \alpha_i \geq 0 \ (i = 1,2, \ldots, l)$ (6.3.3)

όπου a_i είναι ο i-οστός πολλαπλασιαστής Lagrange. Έστω $\theta_t(\theta_t \in \theta)$, η t-οστή παράμετρος μοντέλου

$$\frac{\partial R^2}{\partial \theta_t} = \sum_{i=1}^{l} \frac{\partial k(x_i,x_i)}{\partial \theta_t} - \sum_{i,j=1}^{l} \beta_i^* \beta_j^* \frac{\partial k(x_i,x_j)}{\partial \theta_t} \quad (6.3.4)$$

όπου β_i^* $(i = 1, \dots, l)$ είναι η λύση της (6.3.2). Η παράγωγος του $\|w\|^2$ όσον αφορά τη θ_t είναι ίση με

$$\frac{\partial \|w\|^2}{\partial \theta_t} = (-1) \cdot \sum_{i,j=1}^{l} a_i^* a_j^* y_i y_j \frac{\partial k(x_i,x_j)}{\partial \theta_t} \quad (6.3.5)$$

όπου $a_i^*(i = 1, \dots, l)$ είναι η λύση της (6.3.3). Με αυτό τον τρόπο η παράγωγος του φράγματος ακτινωτού περιθωρίου, όσον αφορά τη θ_t είναι ίση με

$$\frac{\partial (R^2\|w\|^2)}{\partial \theta_t} = \|w\|^2 \frac{\partial R^2}{\partial \theta_t} + R^2 \frac{\partial \|w\|^2}{\partial \theta^t} \quad (6.3.6)$$

Το μοντέλο επιλογής με το φράγμα ακτινωτού περιθωρίου περιγράφεται εν ολίγοις ως εξής:

1. Θέσε τη θ_r σε μια αρχική τιμή θ_0.
2. Βάσει του πρόσφατου θ_r, βελτιστοποίησε τα α και β βάσει των (6.3.3) και (6.3.2) αντίστοιχα, και δήλωσε τις βέλτιστες λύσεις από τα α_r^* και β_r^*.
3. Από τη στιγμή που τα α_r^* και β_r^* βρεθούν, η παράγωγος της (2.3.6) μπορεί να υπολογιστεί ρητά από για μια δοθείσα θ_r. Ως εκ τούτου, μια μέθοδος έρευνας βασισμένη στην κλίση μπορεί να χρησιμοποιηθεί για να ελαχιστοποιήσει το $R^2\|w\|^2$ όσον αφορά τη θ_r. Ο ελαχιστοποιητής δηλώνεται από θ_{r+1}.
4. Σταμάτα αν ένα δοθέν κριτήριο διακοπής ικανοποιείται και θ_{r+1} είναι το επιλεγμένο μοντέλο. Διαφορετικά, έστω $\theta_r \leftarrow \theta_{r+1}$ και πήγαινε στο βήμα 2.

6.4 ΚΡΙΤΗΡΙΟ ΕΠΙΛΟΓΗΣ ΜΟΝΤΕΛΟΥ Ι

Έστω τα D και D_t δηλώνουν τα σύνολα εκπαίδευσης και δοκιμής, αντιστοίχως. Το $E(D_t)$, δηλώνει τον αριθμό των εσφαλμένα ταξινομημένων δειγμάτων που παίρνουμε εφαρμόζοντας το multiclass SVM ταξινομητή στο D_t. Ο ταξινομητής και το σύνολο δοκιμής υποτίθεται ότι είναι σταθερά και άγνωστα. Για ένα πρόβλημα c-κλάσεων:

$$E(D_t) = \sum_{i=1}^{c} \sum_{j=1, j\neq i}^{c} E_{ij}(D_t) \quad (6.4.1)$$

Το $E_{ij}(D_t)$ δηλώνει τον αριθμό των εσφαλμένα ταξινομημένων δειγμάτων, από την κλάση i στη j, και εκφράζεται ως:

$$E_{ij}(D_t) = |\{x | x \in D_t, y^0(x) = i, y^m(x) = j\}| \quad (6.4.2)$$

όπου $|\cdot|$ δηλώνει το μέγεθος του συνόλου. Ένα δείγμα x στο D_t θα υπολογιστεί στο $E_{ij}(D_t)$, εάν και μόνο εάν, η πραγματική του ετικέτα y^0 είναι i, ενώ η ετικέτα y^m προβλέπεται από ένα multiclass SVM ταξινομητή ότι είναι j. Θεωρώντας ότι και η πραγματική και η

προβλεπόμενη ετικέτα είναι μοναδικές για κάθε δείγμα, ένα εσφαλμένα ταξινομημένο δείγμα δε θα υπολογιστεί σε δύο διαφορετικά E_{ij}. Ως εκ τούτου, δεν υπάρχει επικάλυψη μεταξύ αυτών των E_{ij}.

Επικεντρωνόμαστε στη one-versus-all στρατηγική με τον κανόνα ταξινόμησης max wins , η οποία χρησιμοποιείται ευρέως για την επίλυση multiclass SVM προβλημάτων. Με αυτή τη στρατηγική, ένα σύνολο $c(c-1)/2$, ζευγαρωτών δυαδικών SVM ταξινομητών κατασκευάζονται. Έστω ότι το SVM_{ij} δηλώνει το δυαδικό SVM ταξινομητή που εκπαιδεύεται με τα δείγματα από τις κλάσεις i και j. Το $E'_{ij}(D_t)$ είναι ο αριθμός των δειγμάτων δοκιμής που ανήκουν στην κλάση i αλλά είναι εσφαλμένα ταξινομημένα στην κλάση j όταν τα SVM_{ij} εφαρμόζονται στις κλάσεις i και j. Χάριν ευκολίας συμβολισμού, οι ετικέτες που προβλέπονται από τα SVM_{ij} γράφονται ως i ή j παρόλο που είναι +1 ή -1 εν γένει. Το $E'_{ij}(D_t)$ επισήμως εκφράζεται ως:

$$E'_{ij}(D_t) = \left|\{x|x \in D_t, y^0(x) = i, y^b_{ij}(x) = j\}\right| \qquad (6.4.3)$$

όπου το $y^b_{ij}(x)$ αντικαθιστά την ετικέτα που προβλέπεται από το δυαδικό SVM ταξινομητή, SVM_{ij}. Ο ολικός αριθμός των σφαλμάτων που γίνονται από τους $c(c-1)/2$ δυαδικούς SVM ταξινομητές είναι:

$$E'(D_t) = \sum_{1\le i,j\le c, i\ne j} E'_{ij}(D_t) \qquad (6.4.4)$$

Το ακόλουθο αποδεικνύει ότι το $E(D_t)$ είναι άνω φραγμένο από το $E'(D_t)$. Υπό τον κανόνα max wins, η ετικέτα του δείγματος δοκιμής x αποφασίζεται από:

$$y^m(x) = arg\max_{i=1,\dots,c} S_i(x) = arg\max_{i=1,\dots,c}\left(\sum_{j=1, j\ne i}^{c} sign\left[\langle w_{ij}, \varphi(x)\rangle + b_{ij}\right]\right)$$
(6.4.5)

όπου το $\langle w_{ij}, \varphi(x)\rangle + b_{ij}$ είναι θετικό εάν το x είναι ταξινομημένο στην κλάση i. Το sign(α) δηλώνει τη συνάρτηση προσήμου, και είναι +1 για α>0 , 0 για α=0 και -1 διαφορετικά. Το άθροισμα των (c-1) συναρτήσεων προσήμου είναι ένα score, και δηλώνεται από το $S_i(x)$, για την κλάση i. Το δείγμα x αποδίδεται στην κλάση που έχει το υψηλότερο score. Ο κανόνας αμέσως οδηγεί στα ακόλουθα τρία αποτελέσματα:

1) $\forall x \in D_t$, πρέπει να υπάρχουν $S_i(x) \le (c-1)(i = 1, \dots, c)$, και η ισότητα πετυχαίνεται αν και μόνο εάν όλοι οι δυαδικοί SVM ταξινομητές SVM_{ij} $(j = 1, \dots, c, j \ne i)$ ταξινομεί το x στην κλάση i.

2) Εάν $S_i(x) \le S_j(x)$, πρέπει να υπάρχουν $S_i(x) \le (c-1)$. Αναφερόμενοι στο αποτέλεσμα 1), αυτό υποδεικνύει ότι τουλάχιστον ένας από τους δυαδικούς SVM ταξινομητές δεν ταξινομεί το x στην κλάση i.

3) Εάν $S_i(x) = S_j(x)$, τότε και τα δύο πρέπει να είναι μικρότερα του $(c-1)$. Αυτό είναι επειδή τα δυαδικά SVM_{ij} δεν μπορούν να ταξινομήσουν το x στις i και j ταυτόχρονα.

Ας υποθέσουμε ότι ένα multiclass SVM ταξινομεί ένα δείγμα δοκιμής $x_t (t \in D_t)$. Δηλαδή, η πραγματική ετικέτα της $y^0(x_t)$ *είναι* i, ενώ η προβλεπόμενη ετικέτα της $y^m(x_t)$ είναι j. Αυτό συμβάλλει σε μια μέτρηση του $E_{ij}(D_t)$, βασισμένη στην (6.4.2). Αναφερόμενοι στην (6.4.5) , αυτό σημαίνει ότι $S_j(x_t)$ είναι το υψηλότερο score, και ως εκ τούτου, $S_i(x_t) = S_j(x_t)$. Εφαρμόζοντας τα αποτελέσματα 2) και 3), παίρνουμε $S_i(x_t) \le (c-1)$, υποδεικνύοντας ότι τουλάχιστον ένα από τα δυαδικά SVM έχουν ταξινομήσει το σφάλμα x_t. Αυτό συμβάλλει σε μια μέτρηση του $E'_{ik}(D_t)$; Επίσης να σημειωθεί το k δεν είναι αναγκαίο

να είναι ακριβώς το j στο $E_{ij}(D_t)$. Ως εκ τούτου, για κάθε δείγμα δοκιμής που ταξινομείται εσφαλμένα από ένα multiclass SVM, θα πρέπει να έχει ταξινομηθεί εσφαλμένα από τουλάχιστον δυαδικό SVM ταξινομητή. Αθροίζοντας τα E_{ij} και E'_{ik} για i και j (ή k) δημιουργείται η σχέση

$$\sum_{1\le i,j\le c, i\ne j} E_{ij} \le \sum_{1\le i,k\le c, i\ne k} E'_{ik} \Leftrightarrow E(D_t) \le E'(D_t) \qquad (6.4.6)$$

Αυτό αποδεικνύει ότι το $E(D_t)$ είναι άνω φραγμένο από το $E'(D_t)$. Εντωμεταξύ, αξίζει να αναφερθεί ότι η $E_{ij}(D_t) \le E'_{ij}(D_t)$ δεν είναι αναγκαίο να ισχύει.

Το προαναφερθέν αποτέλεσμα υποδεικνύει ότι για να μειώσουμε την αξία του $E(D_t)$, πρέπει να ελαχιστοποιήσουμε το άνω φράγμα $E'(D_t)$. Αυτό οδηγεί σε ένα κριτήριο επιλογής ως ακολούθως. Όπως είναι γνωστό το σφάλμα δοκιμής $\left(E'_{ij} + E'_{ji}\right)$ μπορεί να εκτιμηθεί μέσω του leave-one-out σφάλματος του SVM_{ij} , που δηλώνεται από L_{ij}, ικανοποιεί τη σχέση

$$L_{ij} \le 4R_{ij}^2 \left\|w_{ij}\right\|^2 \qquad (6.4.7)$$

Έτσι, το $E'(D_t)$ μπορεί να εκτιμηθεί από $\sum_{1\le i<j\le c} L_{ij}$, και ικανοποιεί τη:

$$\sum_{1\le i<j\le c} L_{ij} \le \sum_{1\le i<j\le c} 4R_{ij}^2 \left\|w_{ij}\right\|^2 \qquad (6.4.8)$$

Για να ελαχιστοποιήσουμε το $E'(D_t)$ ή πιο συγκεκριμένα την εκτίμησή του, το δεξί μέλος πρέπει να ελαχιστοποιηθεί.

Βασιζόμενοι στην προαναφερθείσα ανάλυση, το $\sum_{1\le i<j\le c} 4R_{ij}^2 \left\|w_{ij}\right\|^2$ ορίζεται ως ένα μοντέλο επιλογής κριτηρίου για τα multiclass SVMs. Είναι ένας ζευγαρωτός συνδυασμός των φραγμάτων ακτινωτού περιθωρίου των ταξινομητών δυαδικών SVMs. Το βέλτιστο σύνολο παραμέτρων μοντέλου βρίσκεται από:

$$\theta^* = arg\min_{\theta\in\Theta} \left(\sum_{1\le i<j\le c} R_{ij}^2 \left\|w_{ij}\right\|^2 \right) \qquad (6.4.9)$$

Η παράγωγος του κριτηρίου σε σχέση με τη t-οστή παράμετρο μοντέλου θ_t είναι:

$$\frac{\partial}{\partial\theta_t}\left(\sum_{1\le i<j\le c} R_{ij}^2 \left\|w_{ij}\right\|^2 \right) = \sum_{1\le i<j\le c}\left(\left\|w_{ij}\right\|^2 \frac{\partial R_{ij}^2}{\partial\theta_t} + R_{ij}^2 \frac{\partial\left\|w_{ij}\right\|^2}{\partial\theta^t}\right) \qquad (6.4.10)$$

Ο υπολογισμός του $\frac{\partial R_{ij}^2}{\partial\theta_t}$ και του $\frac{\partial\left\|w_{ij}\right\|^2}{\partial\theta^t}$ ακολουθεί τις (6.4.4) και (6.4.5). Όπως και στη δυαδική ταξινόμηση, το βέλτιστο σύνολο παραμέτρων μοντέλου θ^*μπορεί να βρεθεί με τη χρήση τεχνικών βελτιστοποίησης βασισμένων στην κλίση.

Πριν τη λήξη αυτής της ενότητας θα ασχοληθούμε με τη σχέση μεταξύ των προτεινόμενων κριτηρίων επιλογής μοντέλου και του φράγματος ακτινωτού περιθωρίου, γενικευμένου για τα multiclass SVMs, το οποίο μπορεί να εκφραστεί ως

$$L(D) \le (4K/c)\left(R^2 \sum_{1\le i<j\le c}\left\|w_i - w_j\right\|^2\right) \triangleq (4K/c)\left(R^2 \sum_{1\le i<j\le c}\left\|\widetilde{w_{ij}}\right\|^2\right) \qquad (6.4.11)$$

όπου K είναι μια σταθερά και c, ο αριθμός των κλάσεων. Το R περικλείει όλα τα δείγματα εκπαίδευσης. Για το προτεινόμενο Κριτήριο I στην (6.4.9) , το R_{ij} είναι η ακτίνα της μικρότερης σφαίρας που περικλείει τα δείγματα εκπαίδευσης από τις κλάσεις i και j, και ως εκ τούτου $R_{ij}^2 \leq R^2$. Αντικαθιστώντας όλα τα R_{ij}^2 στην (6.4.9) με R^2και βγάζοντας το R^2 εκτός του αθροίσματος μετατρέπει το Κριτήριο I σε $\left(R^2 \sum_{1\leq i<j\leq c}\|w_{ij}\|^2\right)$.

6.5 ΚΡΙΤΗΡΙΟ ΕΠΙΛΟΓΗΣ ΜΟΝΤΕΛΟΥ II

Το μέτρο που βασίζεται στη διασπορά πίνακα (scatter-matrix-based measure) συχνά προτιμάται, χάρις στην απλότητα και την πρακτικότητα των δυαδικών και των multiclass προβλημάτων. Ορίζονται ως εξής:

$$S_W = \sum_{i=1}^{c}\left[\sum_{x\in D_i}(x-m_i)(x-m_i)^T\right]$$
$$S_B = \sum_{i=1}^{c} n_i(m_i-m)(m_i-m)^T$$
$$S_T = \sum_{i=1}^{c}\left[\sum_{x\in D_i}(x-m)(x-m)^T\right] = S_W + S_B \qquad (6.5.1)$$

C είναι ο αριθμός των κλάσεων, D_i είναι το σύνολο των δειγμάτων εκπαίδευσης από την κλάση i, και n_i είναι το μέγεθος του D_i. Οι m_i και m είναι οι μέσοι της κλάσης και ολικοί μέσοι, αντίστοιχα. Πολλοί συνδυασμοί των δύο S_W, S_B και S_T μπορούν να χρησιμοποιηθούν σαν μέτρο της διαχωρισιμότητας της κλάσης. Τα ευρέως χρησιμοποιημένα περιλαμβάνουν τα $tr(S_B)/tr(S_W)$ και $|S_B|/|S_W|$, όπου τα $tr(A)$ και $|A|$ δηλώνουν το ίχνος και την ορίζουσα του τετραγωνικού πίνακα A, αντίστοιχα.

Στη συνέχεια θα εξετάσουμε τη σχέση μεταξύ του μέτρου διαχωρισιμότητας κλάσεων που βασίζεται στους πίνακες διασποράς και του φράγματος ακτινωτού περιθωρίου στη multiclass περίπτωση και αυτή η σύγκριση θα χρησιμοποιηθεί για να αναπτύξουμε το δεύτερο κριτήριο επιλογής μοντέλου. Για να γίνει αυτό, το ακόλουθο πρώτα επεκτείνει τη διαχωρισιμότητα κλάσεων σε ένα χώρο χαρακτηριστικών F που προκαλείται από τους πυρήνες. Θεωρώντας ότι η υψηλή διάσταση του F μπορεί εύκολα να κάνει τους πίνακες διακύμανσης ιδιάζοντες και τους ορίζουσες μηδέν, αντ' αυτού το μέτρο που βασίζεται στο ίχνος χρησιμοποιείται. Στα ακόλουθα, ο εκθέτης φ χρησιμοποιείται για να διακρίνει τις μεταβλητές στο F από αυτές στο $\mathbb{R}^d$. Υπενθυμίζεται ότι τα D_i, δηλώνουν τα δείγματα εκπαίδευσης από την i-οστή κλάση. Το D ορίζεται ως η ένωση των D_i (i=1,…,c), που εκφράζεται ως $D = \bigcup_{i=1}^{c} D_i$. Ο $K_{A,B}$ είναι ένας πίνακας πυρήνα όπου $\{K_{A,B}\}_{i,j} = k(x_i, x_j)$ με τους περιορισμούς τα $x_i \in A$ και $x_j \in B$. Το $Sum(\cdot)$ δηλώνει το άθροισμα όλων των στοιχείων σε έναν πίνακα. Τα ίχνη βρίσκονται ως εξής:

$$tr\left(S_B^{\varphi}\right) = \sum_{i=1}^{c}\frac{Sum\left(K_{D_i,D_i}\right)}{n_i} - \frac{Sum(K_D,D)}{n} \qquad (6.5.2)$$

$$tr\left(S_W^{\varphi}\right) = tr(K_D,D) - \sum_{i=1}^{c}\frac{Sum\left(K_{D_i,D_i}\right)}{n_i} \qquad (6.5.3)$$

$$tr\left(S_T^{\varphi}\right) = tr(K_D,D) - \frac{Sum(K_D,D)}{n} \qquad (6.5.4)$$

Για να διευκολύνουμε την ανάλυση, το μέτρο διαχωρισιμότητας κλάσεων στο F ορίζεται ως $tr(S_B^\varphi)/tr(S_T^\varphi)$ αντί του $tr(S_B^\varphi)/tr(S_W^\varphi)$. Να σημειωθεί ότι αυτά είναι ουσιαστικά ταυτόσημα διότι, $tr(S_T^\varphi) = tr(S_B^\varphi) + tr(S_W^\varphi)$.

Υπενθυμίζεται ότι τα n_1 και n_2 είναι τα μεγέθη των D_1 και D_2 , αντιστοίχως. Η σχέση μεταξύ του $tr(S_B^\varphi)$ και του τετραγωνικού περιθωρίου γ^2 μπορεί να αποδειχθεί ως:

$$\gamma^2 \leq \frac{1}{4-\left(\frac{n_1+n_2}{n_1 n_2}\right)tr(S_B^\Phi)} = \frac{1}{4-\left\|m_1^\varphi - m_2^\varphi\right\|^2} \qquad (6.5.5)$$

Αυτό το αποτέλεσμα υποδεικνύει ότι το $1/\left(4-\left\|m_1^\varphi - m_2^\varphi\right\|^2\right)$ είναι ένα άνω φράγμα του γ^2. Η ισότητα στο'' $\leq$ '' πετυχαίνεται αν και μόνο εάν η λύση του προβλήματος στο (6.3.3), που δηλώνεται από το α_i^*, είναι $1/n_1$ για $x_i \in D_1$και $1/n_2$ για $x_i \in D_2$. Θεωρώντας ότι αυτή η λύση, σπάνια συμβαίνει στην πράξη, το $1/\left(4-\left\|m_1^\varphi - m_2^\varphi\right\|^2\right)$ είναι ένα αυστηρό άνω φράγμα, εν γένει. Υπενθυμίζουμε, ότι όταν ελαχιστοποιούμε το φράγμα ακτινωτού περιθωρίου για την επιλογή μοντέλου, τότε ελαχιστοποιείται και το γ^2. Βασιζόμενοι στην (6.5.5) , για να επιτρέψουμε να μεγιστοποιηθεί το γ^2, το άνω φράγμα του χρειάζεται να είναι επαρκώς μεγάλο και θα εμποδίσει το γ^2 από το να αυξηθεί διαφορετικά. Αυτό, με τη σειρά του, απαιτεί το $\left\|m_1^\varphi - m_2^\varphi\right\|^2$να είναι επαρκώς μεγάλο. Εντωμεταξύ, μειώνοντας το $\left\|m_1^\varphi - m_2^\varphi\right\|^2$μειώνεται η τιμή του άνω φράγματος, εξαναγκάζοντας το γ^2, να διατηρηθεί μικρό. Να σημειωθεί επίσης ότι ένα μεγαλύτερο (ή μικρότερο) $\left\|m_1^\varphi - m_2^\varphi\right\|^2$δεν οδηγεί απαραιτήτως σε ένα μεγαλύτερο (ή μικρότερο) γ^2, οι τιμές τους είναι συχνά θετικά, ισχυρά συσχετισμένες στην πράξη.

Ένα παρόμοιο αποτέλεσμα μπορεί να αποδειχτεί για την τετραγωνική ακτίνα R^2 ως:

$$R^2 \geq \frac{1}{(n_1+n_2)} tr(S_T^\varphi) \qquad (6.5.6)$$

Αυτό δείχνει ότι το $tr(S_T^\varphi)/(n_1+n_2)$, είναι ένα κάτω φράγμα του R^2. Η ισότητα στο $\geq$ πετυχαίνεται αν και μόνο εάν, η λύση του προβλήματος στη (6.3.2), που δηλώνεται από τα β_i^*, είναι $1/(n_1+n_2)$, για όλα τα δείγματα εκπαίδευσης. Άλλη μια φορά, μια τέτοια λύση είναι σπάνια στην πράξη, και αυτό είναι ένα αυστηρό κάτω φράγμα, εν γένει. Όταν ελαχιστοποιούμε το φράγμα ακτινωτού περιθωρίου για την επιλογή μοντέλου, τότε ελαχιστοποιείται το R^2. Βασιζόμενοι στην (6.5.6), το $tr(S_T^\varphi)$ χρειάζεται να είναι επαρκώς μικρό για να αποφευχθεί η παρεμπόδιση της μείωσης του R^2. Επιπροσθέτως, οι τιμές των $tr(S_T^\varphi)$ και R^2 είναι ισχυρά θετικά συσχετισμένες.

Εννοιολογικά μιλώντας, τα $\left\|m_1^\varphi - m_2^\varphi\right\|^2$και γ^2 αντανακλούν την ίδια ιδιότητα της διαχωρισιμότητας δεδομένων, ενώ τα $tr(S_T^\varphi)$ και R^2 μετρούν την ίδια ιδιότητα της διακύμανσης δεδομένων. Εμπνεόμενοι από τα προαναφερθέντα αποτελέσματα, μετατοπίζουμε το φράγμα ακτινωτού περιθωρίου σε ένα multiclass σενάριο, μιμούμενοι το μέτρο διαχωρισιμότητας κλάσεων. Συγχρόνως, να σημειωθεί ότι, αυτό το νέο κριτήριο επιλογής μοντέλου θα βασίζεται ακόμα στα R και $\|w\|$ αντί στα ίχνη των πινάκων διασποράς.

Σε μια multiclass περίπτωση, το $tr(S_T^\varphi)/(n_1+n_2)$ μετρά το μέσο της τετραγωνικής ακτίνας διακύμανσης των δειγμάτων εκπαίδευσης στο F. Θεωρώντας την αναλογία μεταξύ του $tr(S_T^\varphi)/(n_1+n_2)$ και R^2 σε μια δυαδική ταξινόμηση, το νέο κριτήριο ορίζει ξανά το R^2 ως την ακτίνα της μικρότερης σφαίρας που περιλαμβάνει όλα τα δείγματα εκπαίδευσης από τις c κλάσεις:

$R_c^2 = \min_{\|\varphi(x)-\hat{c}\|^2 \le \hat{R}^2}\left(\hat{R}^2\right) \ \forall x \in D.$
(6.5.7)

Για το $tr\left(S_B^{\varphi}\right)$, μπορεί να δειχτεί ότι στην περίπτωση των c κλάσεων:

$$tr = \frac{\sum_{1\le i<j\le c} n_i n_j \left\|m_i^{\varphi} - m_j^{\varphi}\right\|^2}{n^2} \tag{6.5.8}$$

Σημειώνοντας την αναλογία μεταξύ του $\left\|m_i^{\varphi} - m_j^{\varphi}\right\|^2$ και του γ^2 σε μια δυαδική ταξινόμηση, το περιθώριο στο νέο κριτήριο ορίζεται ξανά ως εξής:

$$\bar{\gamma}^2 = \frac{\sum_{1\le i<j\le c} n_i n_j \gamma_{ij}^2}{n^2} = \sum_{1\le i<j\le c} P_i P_J \left\|w_{ij}\right\|^{-2} \tag{6.5.9}$$

Όπου το γ_{ij}, είναι το περιθώριο του δυαδικού SVM ταξινομητή που εκπαιδεύεται με τα δείγματα εκπαίδευσης των κλάσεων i και j, και $P_i = n_i/n$, που είναι η εκ των προτέρων πιθανότητα της κλάσης I εκτιμημένης από τα δείγματα εκπαίδευσης. Το ξαναορισμένο περιθώριο είναι ένα σταθμισμένος μέσος αυτών από τους ζευγαρωτούς SVM ταξινομητές, και ο συντελεστής βαρύτητας είναι το προϊόν των εκ των προτέρων πιθανοτήτων, των δύο συμπεριλαμβανομένων κλάσεων. Αυτό συνεπάγεται ότι τα περιθώρια μεταξύ των κλάσεων που κυριαρχούν στα σύνολα εκπαίδευσης και δοκιμής πρέπει να τονιστούν. Διαφορετικά, ο αριθμός των εσφαλμένα ταξινομημένων δειγμάτων θα είναι υψηλός. Με αυτό τον τρόπο, βρίσκεται το δεύτερο κριτήριο επιλογής μοντέλου, και το βέλτιστο σύνολο παραμέτρων μοντέλου δίνεται από τη σχέση

$$\theta^* = arg \min_{\theta\in\Theta}\left(\frac{R_c^2}{\bar{\gamma}^2}\right) \tag{6.5.10}$$

Η παράγωγος του κριτηρίου όσον αφορά την t-οστή παράμετρο μοντέλου θ_t είναι ίση με

$$\frac{\partial}{\partial\theta_t}\left(\frac{R_c^2}{\bar{\gamma}^2}\right) = \frac{1}{\bar{\gamma}^4}\left(\bar{\gamma}^2 \frac{\partial R_c^2}{\partial\theta_t} - R_c^2 \frac{\partial\bar{\gamma}^2}{\partial\theta_t}\right) \tag{6.5.11}$$

όπου

$$\frac{\partial\bar{\gamma}^2}{\partial\theta_t} = -\left(\sum_{1\le i<j\le c} P_i P_J \left\|w_{ij}\right\|^{-4} \frac{\partial\left\|w_{ij}\right\|^2}{\partial\theta_t}\right) \tag{6.5.12}$$

Ξανά, η ελαχιστοποίηση αυτού του φράγματος μπορεί να επιτευχθεί με τη χρήση τεχνικών βελτιστοποίησης βασισμένων στην κλίση. Σε σύγκριση με το Κριτήριο I, αυτό το κριτήριο είναι περισσότερο πειραματικό και απέχει πολύ από το να ερμηνευθεί σαν ένα φράγμα σφάλματος γενίκευσης.

Τελικά, να σημειωθεί επίσης ότι αυτά τα δύο Κριτήρια μπορούν πρακτικά να επεκταθούν στο να διαχειρίζονται την περίπτωση όπου τα κόστη εσφαλμένης ταξινόμησης μεταξύ των διαφόρων κλάσεων διαφέρουν. Το πρώτο Κριτήριο είναι, επί του παρόντος, ένας ζευγαρωτός συνδυασμός των φραγμάτων ακτινωτού περιθωρίου με τα ίσα βάρη. Όταν διαφορετικά κόστη εσφαλμένης ταξινόμησης ορίζονται, αντ'αυτού ένας σταθμισμένος συνδυασμός μπορεί να εφαρμοστεί. Ένα μεγαλύτερο βάρος μπορεί να ανατεθεί αν τα

κόστη εσφαλμένης ταξινόμησης μεταξύ ενός συγκεκριμένου ζεύγους κλάσεων είναι υψηλότερο. Για το δεύτερο Κριτήριο, η στάθμιση μπορεί να εφαρμοστεί στο ζευγαρωτό συνδυασμό περιθωρίων. Στην ορολογία της διαχωρισιμότητας κλάσεων, αυτό σημαίνει ότι οι δύο κλάσεις με τα υψηλότερα κόστη εσφαλμένης ταξινόμησης θα ωθηθούν η μία μακριά από την άλλη για να μειώσουν τις δυνατές ευκαιρίες εσφαλμένης ταξινόμησης.

6.6 ΥΠΟΛΟΓΙΣΤΙΚΟ ΘΕΜΑ

Και τα δύο κριτήρια έχουν συνεχείς τις παραγώγους δεύτερης και πρώτης τάξης σε σχέση με τις παραμέτρους μοντέλου για όσο διάστημα έχουν οι εφαρμοσμένες συναρτήσεις πυρήνα. Η ελαχιστοποίηση αυτών μπορεί να επιλυθεί επαρκώς με τη χρήση τεχνικών βελτιστοποίησης βασισμένων στην κλίση, όπως για παράδειγμα η μέθοδος Broyden-Fletcher-Goldfarb-Shanno (BFGS) quasi-Newton η οποία προτιμάται συχνά διότι χρειάζεται ευρέως ένα μικρό αριθμό επαναλήψεων πριν την κλίση. Σε κάθε επανάληψη, το υπολογιστικό κόστος είναι μεγάλο εξαιτίας της αξιολόγησης της αντικειμενικής συνάρτησης. Έστω το QP(n) δηλώνει ένα πρόβλημα τετραγωνικού προγραμματισμού με n δείγματα εκπαίδευσης. Έστω s ο αριθμός των τιμών που δοκιμάζονται για κάθε παράμετρο μοντέλου σε μια εξαντλητική μέθοδο έρευνας βασισμένη στο πλέγμα. |θ| είναι ο αριθμός των προς βελτιστοποίηση παραμέτρων μοντέλου, και k είναι ο αριθμός των φορών της διασταυρωμένης επικύρωσης. Εκτός αυτών, το e συμβολίζει τον αριθμό των αξιολογήσεων συναρτήσεων σε μια διαδικασία βελτιστοποίησης. Τα υπολογιστικά φορτία των μεθόδων επιλογής μοντέλου παρατίθενται στον πίνακα του ΣΧΗΜΑΤΟΣ (6.6.1). Όπως φαίνεται στο ΣΧΗΜΑ (6.6.1), για ένα multiclass SVM ταξινομητή που χρησιμοποιεί τη one-versus-all στρατηγική, το να εκπαιδευτεί αυτός ο ταξινομητής έχει ως αποτέλεσμα ένα υπολογιστικό κόστος ίσο με $c \cdot QP(n)$. Για αυτόν που χρησιμοποιεί τη one-versus-one στρατηγική, αυτό το αποτέλεσμα γίνεται $\sum_{1\le i<j\le c} QP(n_i + n_j)$. Ο υπολογισμός των R_c^2 και R_{ij}^2 (ή του $\|w_{ij}\|^2$) στα προτεινόμενα Κριτήρια περιλαμβάνουν ένα $QP(n)$ και ένα $QP(n_i + n_j)$, αντιστοίχως. Από αυτό το ΣΧΗΜΑ, βρίσκεται ότι το υπολογιστικό φορτίο των μεθόδων έρευνας που βασίζονται στο πλέγμα αυξάνεται ραγδαία με την αυξανόμενη τιμή των s, |θ| και k. Γρήγορα γίνονται δισεπίλυτα όταν το |θ| γίνεται μεγαλύτερο του τρία. Σε αντίθεση, τα προτεινόμενα Κριτήρια, έχουν πολύ χαμηλότερο υπολογιστικό φορτίο, χάρις στην εφαρμογή των τεχνικών βελτιστοποίησης που βασίζονται στην κλίση. Όπως φαίνεται στην πειραματική μελέτη, η ελαχιστοποίησή τους μπορεί να επιτευχθεί σε μερικές επαναλήψεις με έναν αριθμό αξιολογήσεων συναρτήσεων, ακόμα και αν το |θ| είναι τόσο μεγάλο όσο μερικές εκατοντάδες. Σε σύγκριση με τις υπάρχουσες μεθόδους, τα προτεινόμενα Κριτήρια μπορούν να εξοικονομήσουν σημαντικό υπολογιστικό κόστος, τόσο πιο πολλές αποταμιεύσεις θα έχουμε.

Method	Computational load
CV/LOO (One-vs-all) [1]	$s^{\lvert\theta\rvert} k \cdot c \cdot \mathrm{QP}(n)$
CV/LOO (One-vs-one) [9]	$s^{\lvert\theta\rvert} k \cdot \sum_{1\le i<j\le c} \mathrm{QP}(n_i + n_j)$
LOO-bound (ECOC dense encoding) [13]	$s^{\lvert\theta\rvert} \cdot \lceil 10 \log_2 c \rceil \cdot \mathrm{QP}(n)$
LOO-bound (ECOC sparse encoding) [13]	$s^{\lvert\theta\rvert} \cdot \lceil 15 \log_2 c \rceil \cdot \mathrm{QP}(0.5n)$
GACV (Single-machine) [12]	$s^{\lvert\theta\rvert} \cdot \mathrm{QP}(cn)$
Criterion I	$e \cdot \left[2 \sum_{1\le i<j\le c} \mathrm{QP}(n_i + n_j)\right]$
Criterion II	$e \cdot \left[\mathrm{QP}(n) + \sum_{1\le i<j\le c} \mathrm{QP}(n_i + n_j)\right]$

* CV : Cross-validation, LOO : Leave-one-out.

ΣΧΗΜΑ 6.6.1: Σύγκριση υπολογιστικών φορτίων

6.7 ΠΕΙΡΑΜΑΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

Αυτό το πείραμα αξιολογεί την αποδοτικότητα των προτεινόμενων κριτηρίων για την επιλογή μοντέλου των multiclass SVMs. Δύο μορφές της Γκαουσιανής ακτινωτής βάσης συνάρτησης πυρήνα (GRBF) χρησιμοποιούνται. Η μία είναι ο σφαιρικός (GRBF) πυρήνας, σε αυτή την περίπτωση το σύνολο παραμέτρων πυρήνα είναι θ={C,σ}, και C είναι η παράμετρος συστηματοποίησης. Η άλλη είναι ο ελλειψοειδής (GRBF) πυρήνας που αποδίδει διαφορετικά σ σε κάθε διάσταση χαρακτηριστικού, σε αυτή την περίπτωση το θ επεκτείνεται στο {C,$\sigma_1, \sigma_2, \dots, \sigma_d$}. Σε αυτό το πείραμα, οι δύο πυρήνες χρησιμοποιούνται για να αξιολογήσουν την απόδοση των προτεινόμενων Κριτηρίων στην αξιολόγηση μικρού και μεγάλου μεγέθους συνόλων παραμέτρων μοντέλου αντιστοίχως.

Η BFGS quasi-Newton μέθοδος εφαρμόζεται στην ελαχιστοποίηση των δύο κριτηρίων για να βρούμε το βέλτιστο σύνολο παραμέτρων μοντέλου. Για να αποφύγουμε τους περιορισμούς των C>0 και σ>0, οι μετασχηματισμοί των $\mu = -\ln(C)\ \kappa\alpha\iota\ \nu = \ln(g) = \ln 12\sigma 2$ εφαρμόζονται, όπου ο $\ln\cdot$ δηλώνει το φυσικό λογάριθμο. Τα μ και ν βελτιστοποιούνται αντ'αυτού. Ως εκ τούτου η ελαχιστοποίηση των δύο Κριτηρίων γίνεται ένα πρόβλημα βελτιστοποίησης χωρίς περιορισμούς. Οι αρχικές τιμές των μ και ν θέτονται ως $\mu_0 = 0\ (\text{ή}, \iota\sigma o\delta\acute{\upsilon}\nu\alpha\mu\alpha, C = 1)\ \kappa\alpha\iota\ \nu_0 = -\ln(2d)\ \left(\text{ή}, \iota\sigma o\delta\acute{\upsilon}\nu\alpha\mu\alpha, \sigma = \sqrt{d}\right)$, όπου d είναι η διάσταση του διανύσματος χαρακτηριστικών. Να σημειωθεί ότι για ένα σύνολο δεδομένων, οι συνιστώσες χαρακτηριστικών κατά μήκος κάθε διάστασης, έχουν φτιαχτεί γραμμικά σε κλίμακα στο [-1,1] με τη χρήση των δειγμάτων εκπαίδευσης. Οι συνιστώσες χαρακτηριστικών των δειγμάτων δοκιμής, θα φτιαχτούν σε κλίμακα με τις ίδιες παραμέτρους κλίμακας όταν κάνουμε ταξινόμηση.

Δύο κριτήρια τερματισμού θα χρησιμοποιηθούν, και η βελτιστοποίηση θα τερματιστεί όταν οποιοδήποτε από αυτά ικανοποιηθεί. Έστω ότι τα θ_t και θ_{t+1} δηλώνουν τις παραμέτρους μοντέλου που βρίσκονται στην t-οστή και στην (t+1)-οστή επαναλήψεις, αντίστοιχα. Οι $f(\theta_t)$ και $f(\theta_{t+1})$ είναι οι αντίστοιχες τιμές της αντικειμενικής συνάρτησης. Το πρώτο Κριτήριο τερματισμού είναι $|f(\theta_{t+1}) - f(\theta_t)| \leq \epsilon f(\theta_t)$, όπου ϵ, ένας μικρός θετικός αριθμός που θέτουμε ως 10^{-5}, σε αυτό το πείραμα. Με αυτό το κριτήριο τερματισμού, η βελτιστοποίηση θα τερματιστεί εάν η διαφορά των συναρτησιακών τιμών, σε δύο διαδοχικές επαναλήψεις, είναι μικρότερη από μια προκαθορισμένη ανεκτικότητα. Το δεύτερο κριτήριο τερματισμού είναι ειδικά σχεδιασμένο για την ελαχιστοποίηση φράγματος ακτινωτού περιθωρίου. Σε κάθε επανάληψη του BFGS quasi-Newton αλγορίθμου, μια έρευνα γραμμής διεξάγεται για να βρεθεί το σημείο έναρξης και η διεύθυνση για κάθε επανάληψη. Εντούτοις, υπερβολικά πολλές έρευνες γραμμής μπορεί να διεξαχθούν σε μία και μόνο επανάληψη όταν το ελάχιστο του φράγματος προσεγγίζεται. Ο λόγος είναι ότι στην πράξη, οι παράγωγοι που υπολογίζονται βασισμένες στις (6.3.4) και (6.3.5) μπορεί να αποκλίνουν ελαφρώς από τις θεωρητικές τους τιμές. Όταν το ελάχιστο του φράγματος προσεγγίζεται, οι παράγωγοι θα είναι μικρές, και ο αντίκτυπος της απόκλισης θα γίνει σημαντικός. Εξαιτίας της ανακριβούς πληροφορίας παραγώγου, μια επανάληψη μπορεί να πάρει πολλές έρευνες γραμμής για να βρει μια λύση. Για να αντιμετωπιστεί αυτό, το δεύτερο κριτήριο τερματισμού, τερματίζει τη βελτιστοποίηση εάν ο αριθμός των ερευνών γραμμής σε μια επανάληψη ξεπερνά μια προκαθορισμένη τιμή. Τα σύνολα δεδομένων αναφοράς που χρησιμοποιούνται καταγράφονται στον πίνακα του ΣΧΗΜΑΤΟΣ (6.7.1). Το d δηλώνει τη διάσταση , ''#Class'' είναι ο αριθμός των κλάσεων, και n είναι το μέγεθος του συνόλου δεδομένων. Για τα ''DNA''και ''Satimage'' n και n_{test} είναι τα μεγέθη των συνόλων εκπαίδευσης και δοκιμής, αντιστοίχως. Η τελευταία στήλη του πίνακα του ΣΧΗΜΑΤΟΣ (6.7.1) καταγράφει τους μέγιστους, ελάχιστους και μέσους αριθμούς δειγμάτων

σε κάθε κλάση. Για ένα σύνολο δεδομένων χωρίς προκαθορισμένα σύνολα εκπαίδευσης/δοκιμής, το συνολικό σύνολο δεδομένων χωρίζεται τυχαία σε 100 ζεύγη υποσυνόλων εκπαίδευσης/δοκιμής (50%:50%), και τα πέντε πρώτα υποσύνολα εκπαίδευσης χρησιμοποιούνται για επιλογή μοντέλου.

Data	d	#Class	n	n_{test}	max/min/avg number of samples per class
Iris	4	3	150	–	50/50/50
Wine	13	3	178	–	71/48/59.3
Glass	13	6	214	–	76/0/35.7
Car	6	4	1728	–	1210/65/432
Dermatology	34	6	366	–	112/20/61
E.coli	7	8	336	–	143/2/42
Yeast	8	10	1484	–	463/5/148.4
Zoo	16	7	101	–	41/4/14.4
Vowel	10	11	528	–	48/48/48
Vehicle	18	4	846	–	218/199/211.5
Segment	19	7	2310	–	330/330/330
DNA	180	3	2000	1186	1654/765/1062
Satimage	36	6	4435	2000	1533/626/1072.5

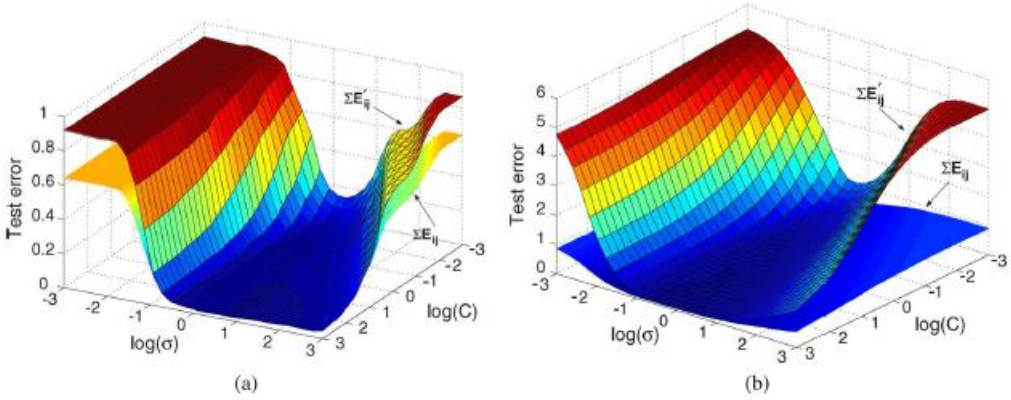
ΣΧΗΜΑ 6.7.1: Αποδόσεις των multiclass συνόλων δεδομένων αναφοράς

6.8 ΙΔΙΟΤΗΤΕΣ ΤΩΝ ΔΥΟ ΚΡΙΤΗΡΙΩΝ ΕΠΙΛΟΓΗΣ ΜΟΝΤΕΛΟΥ

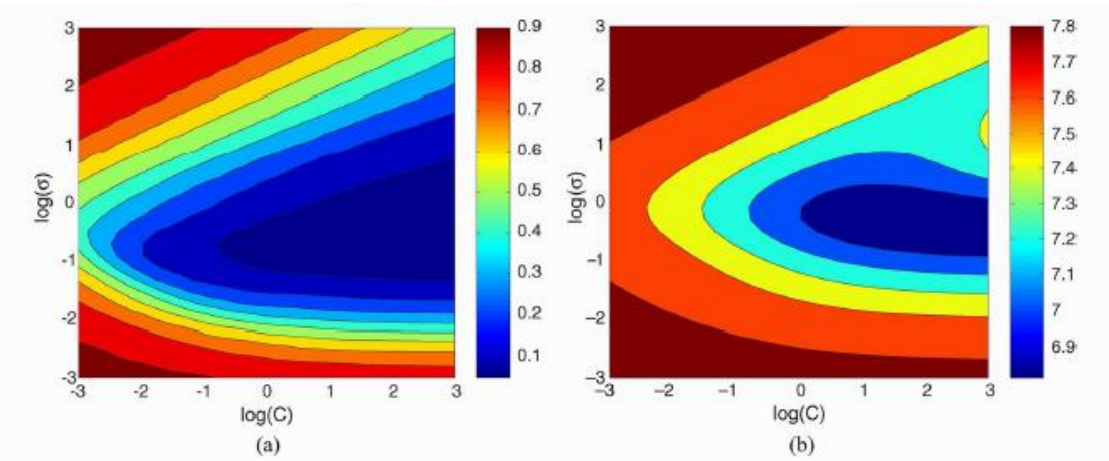
1) Σχέση μεταξύ των $\sum E_{ij}, \sum E'_{ij}$και του κριτηρίου I: όπως ορίστηκε και προηγουμένως, το E_{ij} είναι ο αριθμός των εσφαλμένα ταξινομημένων δειγμάτων δοκιμής που από την κλάση I στην κλάση j, όταν εφαρμόζεται ένας multiclass SVM ταξινομητής, ενώ το E'_{ij} είναι ένας τέτοιος αριθμός όταν ένας δυαδικός multiclass SVM ταξινομητής, SVM_{ij} χρησιμοποιείται. Σε αυτό το πείραμα, με διαφορετικά ζεύγη C και σ, οι τιμές των $\sum E_{ij}, \sum E'_{ij}$και του Κριτηρίου I, υπολογίζονται και συγκρίνονται. Τα αποτελέσματα των ''Wine'' και ''Vowel'' συνόλων δεδομένων φαίνονται στο ΣΧΗΜΑ 6.8.1 Οι άξονες είναι $\ln(C), \ln(\sigma)$, και σφάλμα δοκιμής αντιστοίχως. Οι κυρτές επιφάνειες των $\sum E_{ij}, και \sum E'_{ij}$ σημειώνονται από τα βέλη. Μπορούμε να δούμε ότι και από τα δύο επιμέρους σχήματα ότι δύο επιφάνειες δείχνουν παρόμοια προφίλ όσον αφορά τα $\ln(C)$ $και$ $\ln(\sigma)$, παρόλο που τα μεγέθη τους είναι διαφορετικά όταν το σφάλμα δοκιμής είναι μεγάλο. Για να μετρηθεί ποσοτικά η συσχέτιση μεταξύ τους, $\sum E_{ij}, και \sum E'_{ij}$ αντιμετωπίζονται ως δύο τυχαίες μεταβλητές. Ο συντελεστής συσχέτισης ρ υπολογίζεται και καταγράφεται κάτω από κάθε επιμέρους σχήμα. Οι δύο μεταβλητές βρίσκονται να είναι ισχυρά θετικά συσχετισμένες, υποδηλώνοντας ότι ένα μικρότερο $\sum E'_{ij}$ αντιστοιχεί, εν γένει, σε ένα μικρότερο $\sum E_{ij}$. Παρόμοια αποτελέσματα παρατηρούνται επίσης και από άλλα σύνολα δεδομένων. Αυτά τα αποτελέσματα δείχνουν ότι είναι λογικό να χρησιμοποιούμε το $\sum E'_{ij}$, για να εκτιμήσουμε το $\sum E_{ij}$ για την επιλογή μοντέλου.

Τώρα, η αντιστοιχία μεταξύ του σφάλματος δοκιμής $\sum E_{ij}$ και του Κριτηρίου I έχει ελεγχθεί περαιτέρω. Το $\sum E_{ij}$ και οι τιμές του Κριτηρίου συχνά δεν έχουν τόση ισχυρή συσχέτιση όπως έχουν τα $\sum E_{ij}, και \sum E'_{ij}$. Αυτό συμβαίνει διότι, το φράγμα ακτινωτού περιθωρίου δεν είναι ένα αυστηρό φράγμα του σφάλματος δοκιμής στη δυαδική ταξινόμηση. Το ΣΧΗΜΑ (6.8.2) δείχνει την αντιστοιχία στο σύνολο δεδομένων ''Vowel''. Τα ΣΧΗΜΑΤΑ (6.8.2) (a) και (b) δείχνουν τις τιμές των $\sum E_{ij}$ και του κριτηρίου I, αντιστοίχως. Αυτή είναι μια κάτοψη και η ράβδος στη δεξιά πλευρά υποδηλώνει το μέγεθος. Το $\sum E_{ij}$ και το κριτήριο I δείχνουν παρόμοια περιγράμματα, και η περιοχή των μικρότερων τιμών των κριτηρίων συμμορφώνεται καλά με αυτή των χαμηλότερων ποσοστών σφαλμάτων δοκιμής. Παρόμοια

αποτελέσματα βρίσκονται επίσης και από άλλα σύνολα δεδομένων. Αυτές οι παρατηρήσεις υποδεικνύουν ότι είναι πιθανό να εντοπιστεί η περιοχή των χαμηλότερων ποσοστών σφαλμάτων δοκιμής με την ελαχιστοποίηση του Κριτηρίου Ι. Εντούτοις, μία εξαιρετική περίπτωση βρίσκεται επίσης στο σύνολο δεδομένων ''Vehicle''. Εκεί, η περιοχή των χαμηλότερων τιμών του κριτηρίου δε συμμορφώνεται με την περιοχή των χαμηλότερων ποσοστών σφαλμάτων δοκιμής. Περαιτέρω έρευνα, βρίσκει ότι το φράγμα ακτινωτού περιθωρίου στις κλάσεις 1 και 2 είναι πολύ χαλαρό για να αντανακλά λογικά την αξία των $(E'_{12} + E'_{21})$.



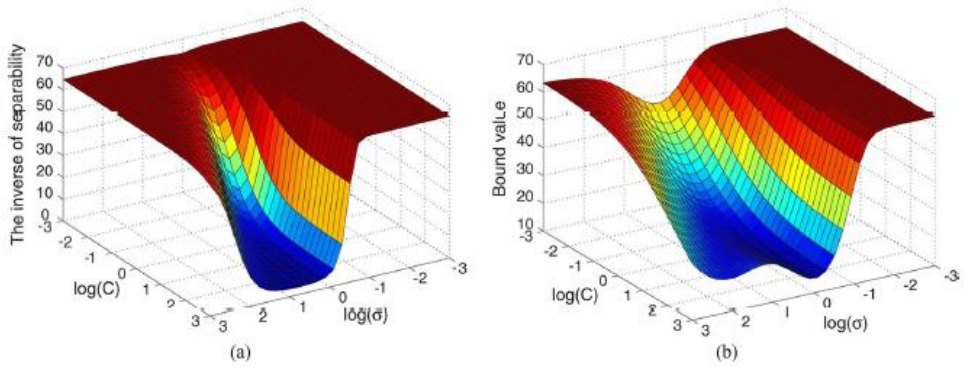
ΣΧΗΜΑ 6.8.1: Αντιστοιχία μεταξύ των $\sum E'_{ij}$ και $\sum E_{ij}$ (ρ είναι ο συντελεστής συσχέτισης μεταξύ τους) (a) Wine, ρ=0.998 (b) Vowel, ρ=0.942.



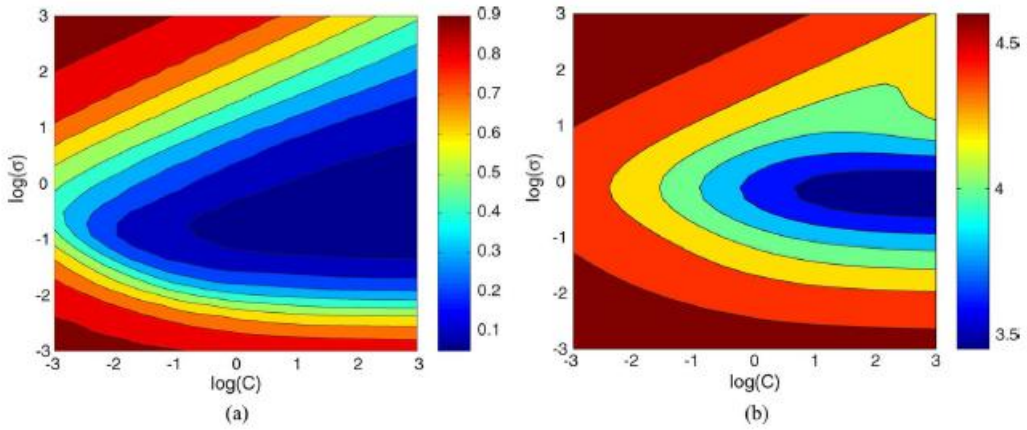
ΣΧΗΜΑ 6.8.2: Αντιστοιχία μεταξύ των $\sum E_{ij}$ και του Κριτηρίου Ι. (a) Vowel, $\sum E_{ij}$ (b) Vowel, Κριτήριο Ι.

2) Σχέση μεταξύ του μέτρου διαχωρισιμότητας κλάσεων και του κριτήριου ΙΙ: Όπως έχει προαναφερθεί, υπάρχει μια κάποια σχέση μεταξύ του μέτρου διαχωρισιμότητας κλάσεων βασισμένου στον πίνακα διακύμανσης και του φράγματος ακτινωτού περιθωρίου. Εμπνευσμένοι από αυτό, αναπτύσσεται το κριτήριο ΙΙ. Παρόλο που αυτή η σχέση δε σχετίζεται άμεσα με την αποδοτικότητα του κριτηρίου, εξακολουθεί να αποδεικνύεται σε αυτό το πείραμα για χάρη της ακεραιότητας. Μετά από αυτό, η αντιστοιχία μεταξύ των $\sum E_{ij}$ και του Κριτηρίου ΙΙ, θα δειχθεί. Το ΣΧΗΜΑ (6.8.3) δείχνει τη διαχωρισιμότητα κλάσεων και το φράγμα ακτινωτού περιθωρίου υπολογισμένα με τη χρήση των δύο πρώτων κλάσεων του συνόλου δεδομένων ''Wine''. Αυτά τα αποτελέσματα φαίνονται στο ΣΧΗΜΑ (6.8.3) (a) και (b), αντιστοίχως. Κατά μήκος του άξονα που δείχνει το φυσικό

λογάριθμο του σ, οι δύο επιφάνειες φτάνουν χαμηλότερες τιμές μέσα στη γειτονική τους τοποθεσία. Τελικά, η αντιστοιχία μεταξύ των $\sum E_{ij}$ και του Κριτηρίου ΙΙ φαίνεται στο ΣΧΗΜΑ (6.8.4).



ΣΧΗΜΑ 6.8.3: Συσχέτιση μεταξύ της διαχωρισιμότητας κλάσεων και του φράγματος ακτινωτού περιθωρίου (a) Wine, $\sum E_{ij}$ (b) Vowel, Κριτήριο ΙΙ.



ΣΧΗΜΑ 6.8.4: Αντιστοιχία μεταξύ των $\sum E_{ij}$ και του Κριτηρίου ΙΙ (a) Vowel, $\sum E_{ij}$ (b) Vowel, Κριτήριο ΙΙ.

6.9 ΠΕΙΡΑΜΑΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ ΓΙΑ ΤΑ ΣΥΝΟΛΑ ΔΕΔΟΜΕΝΩΝ ΑΝΑΦΟΡΑΣ

Τα σύνολα δεδομένων αναφοράς του πίνακα του ΣΧΗΜΑΤΟΣ (6.7.1) χρησιμοποιούνται στο πείραμα. Έχουν διαφορετικές διαστάσεις, άγνωστες πραγματικές κατανομές, και διαφορετικά μεγέθη δείγματος. Κάποια από αυτά έχουν μη σταθμισμένες κλάσεις, όπως τα ''Car'', ''E.Coli'',και το ''Yeast''. Αυτά τα σύνολα δεδομένων σχηματίζουν μια καλή βάση για την αξιολόγηση των δύο κριτηρίων επιλογής μοντέλου. Σε αυτό το πείραμα, η επιλογή μοντέλου που προκύπτει από τα προτεινόμενα Κριτήρια συγκρίνεται με αυτή που χρησιμοποιεί η προσέγγιση της πενταπλής διασταυρωμένης επικύρωσης, η οποία θεωρείται ως σημείο αναφοράς. Σε αυτή την προσέγγιση, διαφορετικά ζεύγη των {σ,C} αξιολογούνται μέσω μιας 30×30 έρευνας πλέγματος στα πέντε ανώτερα υποσύνολα εκπαίδευσης. Για κάθε υποσύνολο εκπαίδευσης, το ζεύγος που αυξάνει το ελάχιστο σφάλμα διασταυρωμένης επικύρωσης επιλέγεται, και πέντε ζεύγη {σ,C}, βρίσκονται

συνολικά. Ο μέσος των πέντε τιμών σ, επιλέγεται ως βέλτιστο σ, και ο μέσος των πέντε C τιμών επιλέγεται ως το βέλτιστο C.

Για τα κριτήρια I και II, τα βέλτιστα σ και C βρίσκονται με τη χρήση BFGS quasi-Newton μεθόδου βελτιστοποίησης. Για να εκφράσουν το υπολογιστικό τους φορτίο, ο αριθμός των επαναλήψεων και αυτός των αξιολογήσεων συναρτήσεων καταγράφονται. Έπειτα, ένας multiclass SVM ταξινομητής με τις βελτιστοποιημένες παραμέτρους μοντέλου δημιουργείται και αξιολογείται με τη χρήση καθενός από τα από τα 100 ζεύγη των υποσυνόλων εκπαίδευσης και δοκιμής, και βρίσκεται το μέσο ποσοστό σφάλματος δοκιμής. Έξι διαφορετικές multiclass SVM μέθοδοι που καταγράφονται στον πίνακα του ΣΧΗΜΑΤΟΣ (6.9.1) ερευνώνται. Και αυτό γίνεται για να αξιολογηθεί η αποδοτικότητα των προτεινόμενων κριτηρίων επιλογής μοντέλου για τις multiclass SVM μεθόδους με τη χρήση διαφορετικών στρατηγικών και κανόνων ταξινόμησης. Ο πίνακας του ΣΧΗΜΑΤΟΣ (6.9.2) καταγράφει τα αποτελέσματα για τις μεθόδους με τη χρήση των one-versus-one και των one-versus-all στρατηγικών. Ο σφαιρικός GRBF πυρήνας εφαρμόζεται. Οι στήλες χωρίζονται σε τρεις ομάδες, δείχνοντας τα αποτελέσματα για τα δύο κριτήρια και την πενταπλή διασταυρωμένη επικύρωση, αντιστοίχως. Στις πρώτες δύο ομάδες, τα t και e δηλώνουν τον αριθμό των επαναλήψεων και τον αριθμό των αξιολογήσεων συναρτήσεων, αντιστοίχως. Τα ''1vs1'' και ''1vsall'' αντιπροσωπεύουν τα μέσα ποσοστά σφάλματος δοκιμής από τις μεθόδους one-versus-one και των one-versus-all, αντιστοίχως. Οι αριθμοί στις παρενθέσεις είναι οι τυπικές αποκλίσεις επισημαίνονται με έντονους χαρακτήρες.

Συγκρίνοντας τα ποσοστά σφάλματος δοκιμής, βλέπουμε και για τις δύο multiclass SVM μεθόδους, τα προτεινόμενα κριτήρια αυξάνουν την απόδοση ταξινόμησης σε σύγκριση με αυτή που προκύπτει με τη χρήση της πενταπλάσιας διασταυρωμένης επικύρωσης. Για το Κριτήριο I, πετυχαίνει το ελάχιστο ποσοστό σφάλματος δοκιμής στα ''Wine'', ''Zoo'', και ''DNA''. Στα ''Iris'',''Car'', ''Dermatology'', ''E.Coli'' και ''Yeast'', τα ποσοστά σφάλματος δοκιμής είναι παρόμοια με αυτά από τη διασταυρωμένη επικύρωση. Το Κριτήριο I παράγει ελαφρά υψηλότερα σφάλματα δοκιμής στα ''Glass'', ''Vowel'', ''Segment'', και ''Satimage''. Όσο για το κριτήριο II είναι λίγο υψηλότερα. Η σύγκριση μεταξύ των δύο Κριτηρίων δείχνει ότι το κριτήριο I οδηγεί σε ελαφρά καλύτερη συνολική απόδοση ταξινόμησης. Αυτό μπορεί να διαπιστωθεί από τα σύνολα δεδομένων ''Dermatology'', ''E.Coli'' και ''Yeast'' και ''DNA''. Και για τα δύο κριτήρια, τα σφάλματα δοκιμής των μεθόδων των ''1vs1'' και ''1vsall'' είναι συγκρίσιμα στα περισσότερα σύνολα δεδομένων; Εντούτοις, στο ''Vowel'' και στο , ''Segment''τα ποσοστά σφάλματος δοκιμής για τη ''1vsall''μέθοδο είναι λίγο υψηλότερα. Από την άλλη πλευρά, μια αποτυχία του μοντέλου επιλογής παρατηρείται στο ''Vehicle'', όπου τα ποσοστά σφάλματος δοκιμής που βρίσκονται είναι σημαντικά υψηλότερα από αυτά της πενταπλάσιας διασταυρωμένης επικύρωσης.

Από τις τιμές των t και e, είναι γνωστό ότι η διαδικασία ελαχιστοποίησης συχνά επιτυγχάνεται σε μερικές επαναλήψεις με ένα μικρό αριθμό αξιολογήσεων συναρτήσεων. Για παράδειγμα, στο ''Iris'', όταν το κριτήριο I χρησιμοποιείται, η επιλογή μοντέλου σε ένα υποσύνολο εκπαίδευσης ολοκληρώνεται σε περίπου εφτά επαναλήψεις, συμπεριλαμβανομένων συνολικά 26 αξιολογήσεων συναρτήσεων. Αναφερόμενοι στον πίνακα του ΣΧΗΜΑΤΟΣ (6.6.1) , αυτό σημαίνει ότι, $156(26 \times 2 \times 3)$ QP προβλήματα λύνονται. Εντούτοις, για τη μέθοδο της πενταπλάσιας διασταυρωμένης επικύρωσης, πρέπει να λύσει $2700(30 \times 30 \times 3)$ QP προβλήματα, όταν χρησιμοποιείται η στρατηγική one-versus-one. Ακόμα και εάν ο αριθμός πλέγματος μειώνεται σε, για παράδειγμα, 10×10, κάνοντας μεγαλύτερα βήματα ή χρησιμοποιώντας την ''coarse-to-fine'' έρευνα, αυτός ο αριθμός μένει τόσο υψηλός όσο το 300.

Τα αποτελέσματα για τις μεθόδους ταξινόμησης των ''dag'', ''dense'',''sparse'' και''single''παρουσιάζονται στον πίνακα του ΣΧΗΜΑΤΟΣ (6.9.3). Τα προτεινόμενα Κριτήρια ακόμα παρέχουν καλή απόδοση ταξινόμησης στα περισσότερα σύνολα δεδομένων εκτός του ''Vehicle''. Λεπτομερώς, το κριτήριο I έχει ελάχιστα ποσοστά σφάλματος δοκιμής στα

''Wine'', ''Zoo'', ενώ τα ποσοστά σφάλματος δοκιμής στα ''Glass'', ''Segment'', και ''Satimage'' είναι λίγο υψηλότερα. Για το κριτήριο ΙΙ, τα ποσοστά σφάλματος δοκιμής είναι, εν γένει, συγκρίσιμα με αυτά της πενταπλής διασταυρωμένης επικύρωσης, αν εξαιρέσουμε ότι κάποιες αυξήσεις παρατηρούνται στα ''Dermatology'', ''E.Coli'' , ''Yeast'', ''Segment'', και ''Satimage''. Όσον αφορά το ποσοστό σφάλματος δοκιμής που παίρνουμε από την πενταπλάσια διασταυρωμένη επικύρωση , η μέγιστη αύξηση που προκαλείται από τη χρήση του κριτηρίου Ι είναι 4.35% στο ''Satimage'' και αυτή που προκαλείται από τη χρήση του κριτηρίου ΙΙ που είναι 3.89% στο ''Yeast''. Το Κριτήριο Ι ακόμα δείχνει περιθωριακά καλύτερη απόδοση ταξινόμησης από το Κριτήριο ΙΙ. Επιπροσθέτως, τα ποσοστά σφάλματος δοκιμής που καταγράφονται στους πίνακες των ΣΧΗΜΑΤΩΝ (6.9.2) και (6.9.3) συγκρίνονται μεταξύ των έξι μεθόδων ταξινόμησης. Καμιά ισχυρή ένδειξη δε δείχνει ότι τα κριτήρια Ι και ΙΙ δεν αποδίδουν με συνέπεια καλύτερα ή χειρότερα σε μια συγκεκριμένη μέθοδο.

Εν συντομία, παρατηρείται από τους πίνακες των ΣΧΗΜΑΤΩΝ (6.9.2) και (6.9.3), ότι και τα δύο Κριτήρια δουλεύουν καλά για τις έξι μεθόδους ταξινόμησης. Παρόλο, που οι διαφορές που παρατηρούνται μεταξύ των ποσοστών σφαλμάτων δοκιμής που λαμβάνονται από τα προτεινόμενα κριτήρια και αυτών της πενταπλάσιας διασταυρωμένης επικύρωσης, δεν είναι σημαντικές, ειδικά όταν οι τυπικές αποκλίσεις λαμβάνονται υπόψη. Τελικά, για να απεικονίσουμε τις λεπτομέρειες σε μια πρακτική διαδικασία βελτιστοποίησης, η εξέλιξη των τιμών του κριτηρίου Ι και του αντίστοιχου ποσοστού σφάλματος δοκιμής, γρήγορα πέφτει με την φθίνουσα τιμή του κριτηρίου . Το ακόλουθο μέρος παρουσιάζει τα πειραματικά αποτελέσματα όταν εφαρμόζεται ο ελλειψοειδής GRBF πυρήνας. Θεωρώντας ότι διαφορετικές τιμές των τιμών σ αποδίδονται σε κάθε διάσταση, ο αριθμός των παραμέτρων του μοντέλου αυξάνει με τη διάσταση των χαρακτηριστικών συν μία. Σε αυτή την περίπτωση, ακόμα και για τα δεδομένα που αναπαρίστανται από τα διανύσματα χαρακτηριστικών χαμηλής διάστασης, η επιλογή μοντέλου με εξαντλητικές μεθόδους έρευνας που βασίζονται στο πλέγμα γίνεται δύσχρηστη (θεωρώντας ότι η μέθοδος της πενταπλής διασταυρωμένης επικύρωσης είναι δύσχρηστη για την περίπτωση των πολλαπλών παραμέτρων μοντέλου, τα σφάλματα δοκιμής που λαμβάνονται με τη χρήση του σφαιρικού πυρήνα πρέπει να χρησιμοποιηθούν στον πίνακα του ΣΧΗΜΑΤΟΣ (6.9.5)). Και τα δύο προτεινόμενα κριτήρια μπορούν ακόμα να δουλεύουν καλά. Όπως φαίνεται στον πίνακα του ΣΧΗΜΑΤΟΣ (6.9.5), η διαδικασία ελαχιστοποίησης ακόμα τελείωνει σε έναν αριθμό επαναλήψεων, παρόλο που ο αριθμός γίνεται μεγαλύτερος εξαιτίας του ότι περισσότερες παράμετροι μοντέλου πρέπει να βελτιστοποιηθούν. Σε σύγκριση με την περίπτωση που χρησιμοποιείται ένας σφαιρικός πυρήνας, χαμηλότερα ποσοστά σφάλματος δοκιμής παρατηρούνται στα 'Iris'',''Car'', ''Wine'', ''Zoo''και ''Segment''. Αυτό μπορεί να είναι το όφελος της χρήσης ενός ελλειψοειδούς πυρήνα όπου οι συνιστώσες χαρακτηριστικών συνδυάζονται σε μια σταθμισμένη τάση. Παρόμοια ποσοστά σφάλματος δοκιμής λαμβάνονται στα ''E.Coli'', ''Vowel'', ''DNA'', όπου υψηλότερα ποσοστά σφάλματος δοκιμής παρατηρούνται στο ''Satimage''. Η απόδοση της επιλογής μοντέλου του κριτηρίου Ι είναι ακόμα καλύτερη από αυτή για το κριτήριο ΙΙ, εν γένει.

Από την άλλη πλευρά, μια υποβαθμισμένη απόδοση παρατηρείται στο ''Glass'' , η οποία δεν είναι τόσο καλή όσο αυτή που λαμβάνεται όταν χρησιμοποιείται ο σφαιρικός πυρήνας. Μέσω της ανάλυσης, πιστεύουμε ότι αυτό οφείλεται στο ότι ο αριθμός των δειγμάτων εκπαίδευσης του ''Glass'' σε κάποιες κλάσεις είναι τόσο μικρός (π.χ. υπάρχουν 2-6 δείγματα εκπαίδευσης μόνο) που η διαδικασία επιλογής μοντέλου υποφέρει από υπερπροσαρμογή, δηλαδή όταν τα δείγματα εκπαίδευσης είναι σπάνια, η ελαχιστοποίηση των κριτηρίων μπορεί να ταιριάξει σε θόρυβο δείγματος και να αποτύχει να αιχμαλωτίσει το πραγματικό μοτίβο εκεί. Σε αυτή την περίπτωση, παρόλο που το ελάχιστο έχει επιτευχθεί, το επιλεγμένο μοντέλο μπορεί να μην είναι καλό, και ένας SVM ταξινομητής που χρησιμοποιεί αυτό το μοντέλο δε θα επιτύχει ικανοποιητική απόδοση ταξινόμησης στα δεδομένα δοκιμής. Επιπροσθέτως, η απόδοση της επιλογής μοντέλου συχνά γίνεται

ευαίσθητη στη ρύθμιση βελτιστοποίησης (με μια άλλη ρύθμιση βελτιστοποίησης, παίρνουμε ένα χαμηλότερο ποσοστό σφάλματος δοκιμής ($33.96\% \pm 3.48\%$) στο σύνολο δεδομένων ''Glass'' , το οποίο είναι συγκρίσιμο με αυτό της πενταπλάσιας διασταυρωμένης επικύρωσης. Εντούτοις, το πραγματικό αποτέλεσμα αναφέρεται χάριν συνέπειας των ρυθμίσεων βελτιστοποίησης για όλα τα σύνολα δεδομένων). Τελικά, ο πίνακας του ΣΧΗΜΑΤΟΣ (6.9.6) παρουσιάζει τα αποτελέσματα από τις μεθόδους ''dag'', ''dense'', ''sparse''και ''single''. Ομοίως, στα σύνολα δεδομένων όπως τα ''Iris'', ''Wine'', ''Car'' και ''Segment'' κάποιες μειώσεις στα ποσοστά σφάλματος δοκιμής παρατηρούνται, ενώ στα άλλα σύνολα δοκιμής όπως στα ''Dermatology'', ''Vowel'' και ''Satimage'' τα ποσοστά σφάλματος δοκιμής αυξάνονται λίγο. Γενικά μιλώντας, για την περίπτωση της χρήσης του ελλειψοειδούς GRBF πυρήνα, τα δύο προτεινόμενα Κριτήρια ακόμα παρουσιάζουν καλή απόδοση για την επιλογή μοντέλου στα multiclass SVMs. Δεν υπάρχει καμιά σημαντική αποικοδόμηση απόδοσης κλίμακας όταν ο αριθμός των παραμέτρων μοντέλου αυξάνεται σημαντικά, για παράδειγμα , τόσο ψηλά όσο 180 για το ''DNA''.

Στη συνέχεια, ο χρόνος επιλογής μοντέλου που χρειάζεται από τα προτεινόμενα Κριτήρια συγκρίνεται με αυτόν που χρειάζεται από την πενταπλάσια διασταυρωμένη επικύρωση. Εντούτοις, όπως φαίνεται στον πίνακα του ΣΧΗΜΑΤΟΣ (6.9.7), τα δύο Κριτήρια είναι ικανά να επιτύχουν γρηγορότερη επιλογή μοντέλου από την προσέγγιση της διασταυρωμένης επικύρωσης. Ο χρόνος επιλογής μοντέλου για τον ελλειψοειδή GRBF πυρήνα είναι περισσότερος διότι περισσότερες παράμετροι πυρήνα πρέπει να βελτιστοποιηθούν.

Αυτό είναι αληθές ιδιαίτερα για σύνολα δεδομένων με χώρους εισόδου υψηλής διάστασης. Επιπροσθέτως, το κριτήριο II κοστίζει εν γένει λιγότερο χρόνο επιλογής μοντέλου από το κριτήριο I. Χρησιμοποιούν διαφορετικούς τρόπους για να αξιολογήσουν την ακτίνα R (λύνοντας ένα και μόνο μεγαλύτερο τετραγωνικό πρόβλημα, σε σχέση με το να λύνουν πολλαπλά μικρότερα). Εντούτοις, πιστεύεται ότι σε πρακτικές εφαρμογές, το να θεωρήσουμε ποιο κριτήριο είναι γρηγορότερο εξαρτάται από το σχεδιασμό του κώδικα. Η εξέλιξη του Κριτηρίου I και του αντίστοιχου ποσοστού σφάλματος δοκιμής φαίνεται επίσης στο ΣΧΗΜΑ (6.9.8). Σε ένα αρχικό στάδιο, η τιμή του κριτηρίου μειώνεται, αλλά το ποσοστό σφάλματος δοκιμής αυξάνεται. Πιστεύεται ότι αυτό οφείλεται στο ότι το κριτήριο δεν είναι μια αυστηρή εκτίμηση του ποσοστού σφάλματος δοκιμής. Όταν η τιμή του κριτηρίου είναι σχετικά μεγάλη, η μείωση της μπορεί να μην οδηγεί σε άμεση μείωση του ποσοστού σφάλματος δοκιμής. Εντούτοις, όπως φαίνεται και στο ΣΧΗΜΑ (6.9.8), όταν η τιμή του κριτηρίου συγκλίνει στο ελάχιστό της, ένα αρκετά χαμηλότερο ποσοστό σφάλματος δοκιμής επιτυγχάνεται.

Name in short	Classification Strategy	Final classification rule
1vsall	One-vs-all	Max wins
1vs1	One-vs-one	Max wins
dag	One-vs-one	Directed acyclic graph
dense	ECOC (dense encoding)	Exp. loss-based decoding
sparse	ECOC (sparse encoding)	Exp. loss-based decoding
single	One single optimization	Max wins

ΣΧΗΜΑ 6.9.1: Έξι SVM μέθοδοι ελέγχονται.

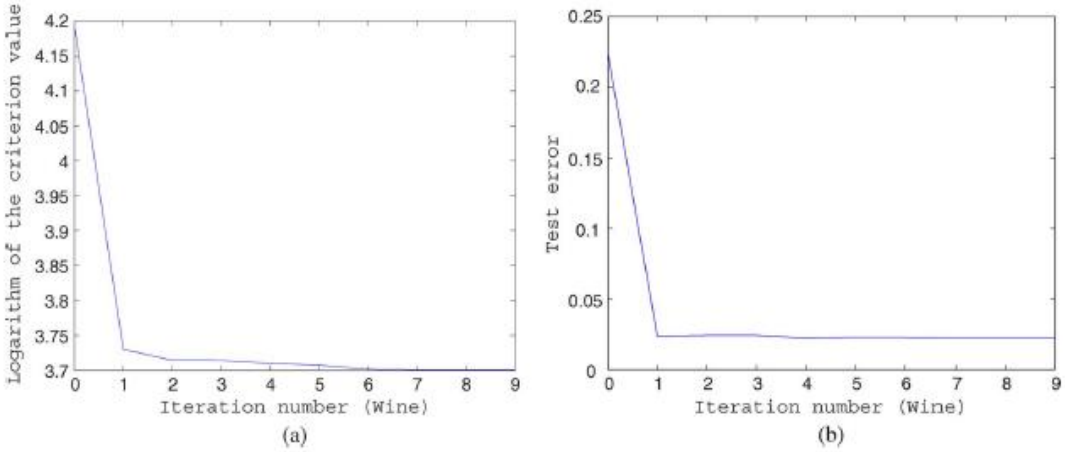
Data set	Criterion I			Criterion II			5-fold CV	
	(t, e)	1vs1	1vsall	(t, e)	1vs1	1vsall	1vs1	1vsall
Iris	(6.4, 25.6)	4.09 (±1.71)	4.17 (±1.72)	(8.8, 29.6)	4.15 (±1.93)	4.17 (±1.83)	**4.33** (±**1.99**)	5.03 (±2.04)
Wine	(7.0, 31.6)	2.22 (±1.40)	**2.15** (±**1.44**)	(6.2, 25.6)	2.43 (±1.51)	2.33 (±1.51)	2.42 (±1.44)	2.40 (±1.49)
Glass	(7.8, 34.4)	35.82 (±4.89)	35.18 (±4.97)	(7.0, 30.0)	33.99 (±4.04)	33.94 (±3.89)	**33.23** (±**3.75**)	34.20 (±3.66)
Car	(6.2, 23.0)	12.28 (±0.67)	12.22 (±0.66)	(6.6, 29.8)	12.27 (±0.67)	12.29 (±0.67)	**11.93** (±**0.62**)	12.03 (±0.66)
Dermat.	(7.2, 27.6)	2.72 (±1.04)	2.48 (±1.10)	(10.0, 41.2)	3.23 (±1.23)	3.86 (±1.19)	2.55 (±1.02)	**2.42** (±**0.97**)
E.coli	(7.4, 34.8)	13.99 (±2.03)	13.55 (±1.84)	(6.8, 30.0)	14.44 (±1.92)	14.08 (±1.86)	13.99 (±1.93)	**13.18** (±**1.84**)
Zoo	(8.8, 33.8)	**6.61** (±**3.91**)	6.80 (±3.88)	(7.0, 31.2)	6.65 (±3.93)	6.80 (±3.94)	7.69 (±4.33)	7.75 (±4.42)
Yeast	(5.8, 24.2)	41.06 (±1.35)	40.44 (±1.43)	(7.6, 31.2)	43.11 (±1.34)	41.88 (±1.58)	40.23 (±1.33)	**40.14** (±**1.39**)
Vowel	(8.2, 34.8)	4.92 (±2.03)	6.23 (±2.20)	(7.8, 37.4)	**4.78** (±**2.01**)	6.34 (±2.21)	4.84 (±2.08)	5.77 (±2.18)
Vehicle$^{\triangle}$	(7.8, 35.2)	28.89 (±1.92)	27.32 (±1.98)	(6.6, 27.8)	32.08 (±2.20)	29.27 (±1.93)	16.62 (±1.88)	**16.44** (±**1.92**)
Segment	(5.8, 24.4)	4.84 (±0.57)	5.29 (±0.60)	(4.2, 17.4)	3.84 (±0.45)	4.69 (±0.56)	3.39 (±0.47)	**3.34** (±**0.49**)
DNA	(9.4, 39.8)	4.55	**4.22**	(5.4, 23.4)	4.81	4.72	4.55	4.30
Satimage	(4.6, 24.2)	10.15	10.10	(5.4, 32.0)	9.20	9.50	**7.90**	7.95

* The error rates of "DNA" and "Satimage" reported in Statlog are 4.1% and 9.4%, respectively.

ΣΧΗΜΑ 6.9.2: Σφάλματα δοκιμής (σφαιρικός GRBF πυρήνας, one-versus-one, one-versus-all)

Data set	Criterion I				Criterion II				5-fold CV			
	dag	dense	sparse	single	dag	dense	sparse	single	dag	dense	sparse	single
Iris	4.31 (±1.87)	4.29 (±1.83)	4.33 (±1.78)	4.15 (±1.76)	4.03 (±1.92)	**4.09 (±1.88)**	4.23 (±1.94)	4.49 (±2.10)	4.33 (±1.99)	5.03 (±2.04)	4.81 (±2.55)	4.55 (±1.91)
Wine	2.27 (±1.44)	2.18 (±1.43)	**2.13 (±1.29)**	2.19 (±1.43)	2.43 (±1.50)	2.40 (±1.45)	2.27 (±1.33)	2.45 (±1.52)	2.42 (±1.44)	2.40 (±1.49)	2.29 (±1.39)	2.34 (±1.46)
Glass	35.24 (±4.57)	34.51 (±4.50)	34.83 (±4.52)	34.57 (±4.42)	34.10 (±3.94)	34.11 (±3.60)	33.60 (±3.56)	33.94 (±3.79)	**32.79 (±4.38)**	33.22 (±3.76)	33.55 (±3.80)	33.43 (±3.98)
Car	12.27 (±0.67)	12.38 (±0.93)	12.38 (±0.93)	12.30 (±0.67)	12.27 (±0.67)	12.01 (±0.87)	12.01 (±0.87)	12.37 (±0.66)	11.91 (±0.71)	**11.68 (±0.81)**	11.77 (±0.85)	11.99 (±0.68)
Dermat.	2.72 (±1.04)	2.52 (±1.07)	2.90 (±1.38)	2.50 (±1.00)	3.24 (±1.21)	3.94 (±1.20)	3.58 (±1.17)	3.30 (±1.03)	2.55 (±1.02)	**2.39 (±0.98)**	2.66 (±1.17)	2.42 (±0.98)
E.coli	13.98 (±2.04)	13.57 (±1.80)	13.95 (±2.15)	13.29 (±1.93)	14.44 (±1.92)	14.07 (±1.90)	14.38 (±1.87)	13.61 (±1.91)	14.00 (±1.94)	**13.30 (±1.99)**	13.53 (±2.01)	13.27 (±1.98)
Zoo	**6.61 (±3.91)**	6.96 (±3.97)	6.78 (±4.05)	6.69 (±3.93)	6.67 (±3.92)	6.98 (±3.97)	6.92 (±4.17)	6.69 (±3.99)	7.65 (±4.30)	7.45 (±4.24)	6.94 (±3.94)	6.88 (±4.02)
Yeast	41.06 (±1.34)	40.53 (±1.42)	43.35 (±1.90)	40.45 (±1.36)	43.09 (±1.31)	42.51 (±1.45)	44.12 (±1.78)	40.27 (±1.32)	**40.23 (±1.33)**	40.54 (±1.37)	42.56 (±1.86)	40.35 (±1.42)
Vowel	5.52 (±2.08)	5.70 (±2.15)	6.00 (±2.16)	6.31 (±2.23)	**4.78 (±2.02)**	6.01 (±2.29)	5.98 (±2.34)	6.22 (±2.25)	5.02 (±1.85)	5.94 (±2.34)	5.94 (±2.35)	5.67 (±2.10)
Vehicle△	28.89 (±1.92)	28.84 (±2.08)	25.54 (±2.05)	28.84 (±2.08)	32.06 (±1.27)	32.82 (±2.08)	32.82 (±2.08)	24.27 (±1.95)	**16.72 (±1.82)**	17.76 (±1.76)	17.38 (±1.66)	16.77 (±1.85)
Segment	4.66 (±0.55)	4.94 (±0.52)	5.33 (±0.70)	5.40 (±0.59)	4.70 (±0.62)	7.08 (±0.67)	6.36 (±0.51)	4.15 (±0.45)	3.41 (±0.49)	3.71 (±0.48)	**3.40 (±0.52)**	3.69 (±0.52)
DNA	4.89	4.64	4.64	4.38	4.81	4.72	4.55	4.38	4.72	**4.22**	4.30	5.56
Satimage	10.75	11.00	12.15	10.00	9.95	10.50	11.25	9.40	8.10	7.95	**7.80**	8.10

ΣΧΗΜΑ 6.9.3: Σφάλματα δοκιμής (σφαιρικός GRBF πυρήνας, DAG, DENSE, SPARSE, SINGLE)



ΣΧΗΜΑ 6.9.4: Εξέλιξη του Κριτηρίου I και του ποσοστού σφάλματος δοκιμής (Wine, σφαιρικός GRBF πυρήνας).(a) Τιμή Κριτηρίου (b) Ποσοστό σφάλματος δοκιμής.

Data set	Criterion I			Criterion II			5-fold CV	
	(t, e)	1vs1	1vsall	(t, e)	1vs1	1vsall	1vs1	1vsall
Iris	(11.0, 50.6)	**3.99** (±**1.71**)	4.01 (±1.70)	(11.68, 50.0)	4.07 (±2.10)	4.03 (±2.06)	4.33 (±1.99)	5.03 (±2.04)
Wine	(33.0, 147.2)	**1.53** (±**1.68**)	1.55 (±1.76)	(15.0, 68.8)	2.03 (±1.97)	1.90 (±1.95)	2.42 (±1.44)	2.40 (±1.49)
Glass$^{\triangle}$	(13.0, 61.6)	40.02 (±4.99)	39.45 (±4.89)	(11.40, 46.4)	40.97 (±5.54)	40.65 (±5.27)	**33.23** (±**3.75**)	34.20 (±3.66)
Car	(15.0, 66.2)	12.10 (±0.71)	12.17 (±0.70)	(14.6, 62.4)	12.06 (±0.71)	12.14 (±0.70)	**11.93** (±**0.62**)	12.03 (±0.66)
Dermat.	(19.0, 86.4)	2.44 (±0.98)	2.48 (±0.99)	(17.2, 79.2)	3.49 (±1.14)	4.15 (±1.17)	2.55 (±1.02)	**2.42** (±**0.97**)
E.coli	(15.0, 69.4)	14.10 (±2.10)	13.45 (±1.88)	(10.6, 47.8)	14.76 (±2.30)	14.30 (±2.23)	13.99 (±1.93)	**13.18** (±**1.84**)
Zoo	(5.2, 23.0)	**5.57** (±**3.78**)	5.57 (±3.78)	(3.0, 13.8)	5.69 (±3.98)	5.84 (±3.95)	7.69 (±4.33)	7.75 (±4.42)
Yeast	(15.0, 62.8)	40.87 (±1.31)	40.34 (±1.36)	(13.0, 59.2)	42.87 (±1.62)	41.88 (±1.58)	40.23 (±1.33)	**40.14** (±**1.39**)
Vowel	(32.0, 146.4)	5.16 (±1.88)	6.16 (±2.04)	(17.8, 83.2)	5.01 (±1.87)	5.74 (±2.15)	**4.84** (±**2.08**)	5.77 (±2.18)
Vehicle$^{\triangle}$	(17.0, 79.2)	27.54 (±1.82)	27.57 (±1.89)	(10.0, 45.6)	28.78 (±1.79)	28.91 (±1.97)	16.62 (±1.88)	**16.44** (±**1.92**)
Segment	(20.0, 88.2)	4.02 (±0.48)	4.06 (±0.49)	(10.2, 48.8)	3.43 (±0.42)	3.92 (±0.49)	3.39 (±0.47)	**3.34** (±**0.49**)
Dna	(11.0, 48.6)	4.72	4.47	(5.4, 23.4)	4.72	4.55	4.55	**4.30**
Satimage	(15.2, 70.8)	11.50	12.55	(5.4, 32.0)	10.95	11.40	**7.90**	7.95

ΣΧΗΜΑ 6.9.5: Σφάλματα δοκιμής (ελλειψοειδής GRBF πυρήνας, one-versus-one,one-versus-all)

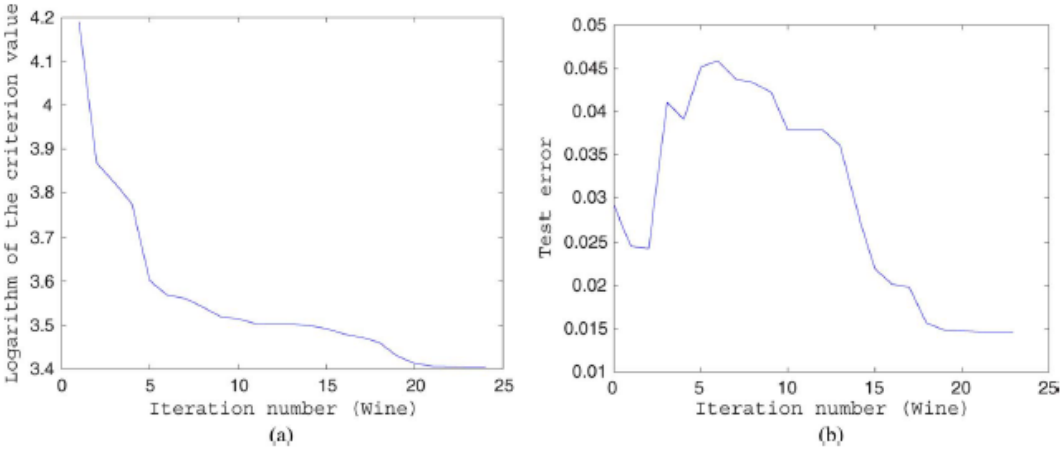
Data set	Criterion I				Criterion II				5-fold CV			
	dag	dense	sparse	single	dag	dense	sparse	single	dag	dense	sparse	single
Iris	3.99 (±1.71)	4.01 (±1.70)	4.00 (±1.71)	4.01 (±1.70)	4.00 (±2.05)	4.03 (±2.06)	**3.95 (±2.08)**	4.27 (±2.19)	4.33 (±1.99)	5.03 (±2.04)	4.81 (±2.55)	4.55 (±1.91)
Wine	1.53 (±1.68)	1.55 (±1.76)	**1.36 (±1.49)**	1.56 (±1.56)	2.03 (±1.97)	1.90 (±1.95)	1.65 (±1.81)	1.56 (±1.56)	2.42 (±1.44)	2.40 (±1.49)	2.29 (±1.39)	2.34 (±1.46)
Glass△	40.03 (±4.99)	39.44 (±4.89)	41.18 (±5.36)	34.90 (±3.84)	40.97 (±5.57)	40.69 (±5.31)	44.03 (±6.06)	34.70 (±3.45)	**32.79 (±4.38)**	33.22 (±3.76)	33.55 (±3.80)	33.43 (±3.98)
Car	12.10 (±0.71)	11.70 (±0.72)	11.70 (±0.72)	12.17 (±0.71)	12.06 (±0.71)	**11.67 (±0.74)**	11.67 (±0.74)	12.14 (±0.70)	11.91 (±0.71)	11.68 (±0.81)	11.77 (±0.85)	11.99 (±0.68)
Dermat.	2.44 (±0.98)	2.53 (±0.99)	2.73 (±1.12)	2.89 (±1.05)	3.53 (±1.13)	4.32 (±1.22)	3.84 (±1.21)	4.27 (±1.25)	2.55 (±1.02)	**2.39 (±0.98)**	2.66 (±1.17)	2.42 (±0.98)
E.coli	14.09 (±2.11)	13.31 (±1.87)	13.80 (±1.99)	**13.05 (±1.93)**	14.74 (±2.31)	14.25 (±2.31)	14.97 (±2.59)	13.57 (±2.00)	14.00 (±1.94)	13.30 (±1.99)	13.53 (±2.01)	13.27 (±1.98)
Zoo	**5.57 (±3.78)**	5.57 (±3.78)	5.59 (±3.75)	8.10 (±3.83)	5.69 (±3.98)	5.78 (±3.92)	5.88 (±3.93)	8.16 (±4.58)	7.65 (±4.30)	7.45 (±4.24)	6.94 (±3.94)	6.88 (±4.02)
Yeast	40.85 (±1.31)	40.50 (±1.29)	44.06 (±2.24)	**39.73 (±1.38)**	42.87 (±1.63)	42.15 (±1.66)	46.49 (±2.30)	39.82 (±1.33)	40.23 (±1.33)	40.54 (±1.37)	42.56 (±1.86)	40.35 (±1.42)
Vowel	5.20 (±1.89)	6.74 (±2.10)	6.46 (±2.36)	6.08 (±2.05)	5.10 (±1.89)	6.47 (±2.19)	6.11 (±2.35)	5.92 (±2.21)	**5.02 (±1.85)**	5.94 (±2.34)	5.94 (±2.35)	5.67 (±2.10)
Vehicle△	27.52 (±1.82)	27.86 (±2.03)	27.86 (±2.03)	24.49 (±1.72)	28.77 (±1.80)	28.30 (±1.86)	28.30 (±1.86)	24.11 (±1.71)	**16.72 (±1.82)**	17.76 (±1.76)	17.38 (±1.66)	16.77 (±1.85)
Segment	4.00 (±0.48)	4.10 (±0.50)	4.26 (±0.53)	3.63 (±0.45)	3.44 (±0.42)	3.88 (±0.46)	3.86 (±0.55)	4.33 (±0.44)	3.41 (±0.45)	3.71 (±0.48)	**3.40 (±0.52)**	3.69 (±0.52)
DNA	4.72	4.47	4.38	4.30	4.72	4.55	4.47	4.64	4.72	**4.22**	4.30	5.56
Satimage	11.45	12.60	13.35	11.25	10.90	11.50	11.95	9.85	8.10	7.95	**7.80**	8.10

ΣΧΗΜΑ 6.9.6: Σφάλματα δοκιμής (ελλειψοειδής GRBF πυρήνας, DAG, DENSE, SPARSE, SINGLE)

Data set	Criterion I		Criterion II		5-fold CV
	Spatial GRBF	Ellipsoidal GRBF	Spatial GRBF	Spatial GRBF	Spatial GRBF
Iris	4.30	7.46	2.77	5.55	14.40
Wine	7.79	11.58	3.07	7.88	18.00
Glass	19.37	56.00	5.62	12.74	39.60
Car	37.83	46.35	32.07	20.79	367.20
Dermatology	19.45	64.35	6.48	44.56	82.80
E.coli	30.56	32.08	8.66	7.94	57.60
Yeast	170.31	225.35	97.40	74.67	918.00
Zoo	25.44	36.77	3.59	8.82	28.80
Vowel	66.75	101.37	14.96	17.11	151.20
Vehicle	34.89	107.47	26.41	95.71	248.40
Segment	88.11	198.72	44.74	66.49	1386.00
Dna	322.12	1219.5	300.62	1853.25	5346.00
Satimage	403.14	1020.60	232.29	962.87	4932.00

* In this experiment, the proposed criteria are computed and optimized by using the mixture codes of Matlab and C. Faster model selection can be expected when entirely implemented in C.

ΣΧΗΜΑ 6.9.7: Σύγκριση του μοντέλου επιλογής χρόνου (μονάδα:second)



ΣΧΗΜΑ 6.9.8: Εξέλιξη του Κριτηρίου I και του ποσοστού σφάλματος δοκιμής (Wine, ελλειψοειδής GRBF πυρήνας).(a) Τιμή Κριτηρίου (b) Ποσοστό σφάλματος δοκιμής

6.10 ΣΥΜΠΕΡΑΣΜΑΤΑ

Βασισμένοι στα προαναφερθέντα πειραματικά αποτελέσματα μπορούμε να εξάγουμε τα ακόλουθα συμπεράσματα:

1) Και τα δύο κριτήρια παρουσιάζουν καλή απόδοση επιλογής μοντέλου στα περισσότερα σύνολα δεδομένων, αυξάνοντας την ακρίβεια ταξινόμησης, σε σύγκριση με αυτή που λαμβάνεται από την προσέγγιση της πενταπλάσιας διασταυρωμένης επικύρωσης.

2)Τα δύο κριτήρια δουλεύουν καλά και με το σφαιρικό και με τον ελλειψοειδή GRBF πυρήνα, και το τελευταίο επαληθεύει την ικανότητά τους στη διαχείριση ενός μεγάλου αριθμού παραμέτρων μοντέλου. Συγχρόνως, παρατηρείται ότι ικανοποιητική επιλογή μοντέλου μπορεί να επιτευχθεί αν δείγματα εκπαίδευσης είναι σπάνια.

3)Παρόλο που τα προτεινόμενα κριτήρια αναπτύσσονται στα multiclass SVM με τη χρήση της one-versus-one στρατηγικής, βοηθούν όλους τους multiclass SVM ταξινομητές να επιτύχουν αρκετά καλή απόδοση ταξινόμησης

4)Η σύγκριση των δύο κριτηρίων βρίσκει το κριτήριο I να δείχνει περιθωριακά καλύτερη απόδοση. Εντωμεταξύ, υπάρχει μια διαφορά μεταξύ των υπολογιστικών φορτίων. Για τον υπολογισμό της ακτίνας $R^{2,}$ το κριτήριο I λύνει πολλαπλά μικρότερης κλίμακας QP προβλήματα, ενώ το κριτήριο II λύνει ένα μεγαλύτερης κλίμακας QP πρόβλημα. Όταν ο αριθμός των κλάσεων είναι μεγάλος, η επιλογή μοντέλου με το δεύτερο κριτήριο μπορεί να γίνει γρηγορότερα.

7

ΑΞΙΟΛΟΓΗΣΗ ΜΟΝΤΕΛΟΥ ΚΑΙ ΕΦΑΡΜΟΓΕΣ

7.1 ΕΙΣΑΓΩΓΗ

Αρκετά παγκόσμια δεδομένα που σχετίζονται με περισσότερα σχέδια γνώσης είναι πολύ μεγάλα για να θεωρηθούν ως σύνολα κατάρτισης για μηχανές διανυσματικής υποστήριξης. Ως εκ τούτου, είναι αναγκαίο να χρησιμοποιούμε υποσύνολα των δεδομένων αυτών ως σύνολα εκπαίδευσης. Από τη στιγμή που χρησιμοποιούμε μόνο ένα υποσύνολο των παγκόσμιων δεδομένων ως σύνολο εκπαίδευσης, βρισκόμαστε αντιμέτωποι με το ερώτημα: Πόσο καλά αποδίδει το μοντέλο, στις περιπτώσεις όπου τα δεδομένα δεν είναι μέρος του συνόλου της κατάρτισης;

Στη συνέχεια θα αναλυθούν θεωρητικά ο Έλεγχος Στατιστικών Υποθέσεων καθώς επίσης και τα Κριτήρια Απόδοσης του Μοντέλου-Confusion Matrix και έπειτα θα αξιολογηθεί πειραματικά η εφαρμογή των μεθόδων Διανυσματικής Υποστήριξης σε προβλήματα ταξινόμησης δύο και τριών κλάσεων. Για να παρέχουμε μία αμερόληπτη εκτίμηση για την ποιότητα ταξινόμησης του κάθε μοντέλου χρησιμοποιώντας τη μέθοδο της διάκρισης (discrimination), οι τιμές των κριτηρίων απόδοσης υπολογίζονται από ένα σύνολο δεδομένων που δεν χρησιμοποιήθηκε στη διαδικασία μοντελοποίησης. Για το σκοπό αυτό χρησιμοποιήσαμε από το πραγματικό σύνολο δεδομένων, ένα μέρος (το σύνολο δοκιμής) το οποίο αφήσαμε στην άκρη για αυτό το σκοπό.

Για τα προβλήματα με Μηχανές Διανυσματικής Ταξινόμησης (γραμμικά ή μη), ένας ταξινομητής θα πρέπει να παρέχει υψηλές τιμές των Κριτηρίων Απόδοσης του Confusion Matrix, δηλαδή της ευαισθησίας (sensitivity), της ειδικότητας (specificity) και της ακρίβειας (accuracy) και η γενικευμένη απόδοση συχνά εκτιμάται με holdout επικύρωση (εκπαίδευση/ δοκιμή).

7.1.1 ΣΤΑΤΙΣΤΙΚΟΣ ΕΛΕΓΧΟΣ ΥΠΟΘΕΣΕΩΝ

Σε πολλά προβλήματα δεν ενδιαφερόμαστε να εκτιμήσουμε με κάποια ακρίβεια την τιμή μίας παραμέτρου αλλά να διαπιστώσουμε αν η παράμετρος είναι μικρότερη ή μεγαλύτερη από μία δεδομένη τιμή που έχει σημασία για το πρόβλημά μας. Δηλαδή επιθυμούμε να ελέγξουμε αν μία ή περισσότερες παράμετροι ενός πληθυσμού ικανοποιούν μία βασική υπόθεση έναντι μίας εναλλακτικής υπόθεσης. Σχεδόν πάντα δεν είμαστε σε θέση να καταγράψουμε όλο τον πληθυσμό οπότε αρκούμαστε σε ένα τυχαίο δείγμα από αυτόν. Με βάση αυτό το τυχαίο δείγμα θέλουμε να πάρουμε μία απόφαση: να απορρίψουμε ή όχι αν ισχύει η βασική υπόθεση.

Ο έλεγχος υπόθεσης (hypothesis testing) επεξεργάζεται στατιστικά εργαλεία (τον εκτιμητή και την κατανομή του) σε μία διαδικασία λήψης απόφασης. Για τη διαδικασία ελέγχου μίας στατιστικής υπόθεσης πρώτα ορίζουμε τη στατιστική υπόθεση, και μετά υπολογίζουμε το στατιστικό ελέγχου και την περιοχή απόρριψης και τέλος αποφασίζουμε για την υπόθεση με βάση την ένδειξη που έχουμε από το δείγμα.

Η στατιστική υπόθεση (statistical hypothesis) μπορεί να είναι μία οποιαδήποτε στατιστική δήλωση ή πρόταση που θέτουμε υπό έλεγχο με βάση τις παρατηρήσεις. Η μηδενική υπόθεση (null hypothesis) την οποία θέτουμε υπό έλεγχο συμβολίζεται με H_0 ενώ η εναλλακτική υπόθεση (alternative hypothesis) την οποία δεχόμαστε αν απορρίψουμε την H_0 συμβολίζεται με H_1. Οι δυνατές αποφάσεις του ελέγχου είναι:

1. Σωστή απόφαση: Αποδεχόμαστε την H_0 όταν η H_0 είναι σωστή. Η πιθανότητα αυτής της απόφασης είναι $P(\alpha\pi o\delta o\chi\acute{\eta}\ \tau\eta\varsigma\ H_0|H_0\ \sigma\omega\sigma\tau\acute{\eta}) = 1-\alpha$.
2. **Σφάλμα τύπου ΙΙ** (type II error): Αποδεχόμαστε την H_0 όταν η H_0 είναι λανθασμένη. Η πιθανότητα αυτού του σφάλματος είναι $P(\alpha\pi o\delta o\chi\acute{\eta}\ \tau\eta\varsigma\ H_0|H_0\ \lambda\alpha\nu\theta\alpha\sigma\mu\acute{\varepsilon}\nu\eta) = \beta$.
3. **Σφάλμα τύπου Ι** (type I error): Απορρίπτουμε την H_0 όταν η H_0 είναι σωστή. Η πιθανότητα αυτού του σφάλματος είναι το επίπεδο σημαντικότητας $P(\alpha\pi o\rho\rho\iota\psi\eta\ \tau\eta\varsigma\ H_0|H_0\ \sigma\omega\sigma\tau\acute{\eta}) = \alpha$.
4. Σωστή απόφαση: Απορρίπτουμε την H_0 και η H_0 είναι λανθασμένη. Η πιθανότητα αυτής της απόφασης είναι $P(\alpha\pi\acute{o}\rho\rho\iota\psi\eta\ \tau\eta\varsigma\ H_0|H_0\ \lambda\alpha\nu\theta\alpha\sigma\mu\acute{\varepsilon}\nu\eta) = \beta$, και δηλώνει την ισχύ του ελέγχου (power of the test).

Οι 4 δυνατές περιπτώσεις στην απόφαση του ελέγχου δίνονται στον πίνακα του **ΣΧΗΜΑΤΟΣ 7.1.1**

	Αποδοχή της H_0	Απόρριψη της H_0
H_0 σωστή	ορθή απόφαση $(1-\alpha)$	σφάλμα τύπου Ι (α)
H_0 λανθασμένη	σφάλμα τύπου ΙΙ (β)	ορθή απόφαση $(1-\beta)$

ΣΧΗΜΑ 7.1.1: Οι 4 περιπτώσεις στην απόφαση ελέγχου με την αντίστοιχη πιθανότητα σε παρένθεση.

Για να είναι ένας έλεγχος ακριβής θα πρέπει το πραγματικό σφάλμα τύπου Ι να είναι στο επίπεδο σημαντικότητας α στο οποίο γίνεται ο έλεγχος. Στην πράξη αυτό δεν είναι βέβαια εφικτό.

Αλλά μπορούμε να το διαπιστώσουμε αν έχουμε τη δυνατότητα να κάνουμε προσομοιώσεις. Για να υπολογίσουμε το πραγματικό σφάλμα τύπου Ι θα πρέπει να γνωρίζουμε ότι η H_0 είναι σωστή και να επαναλάβουμε τον έλεγχο σε Μ διαφορετικά δείγματα ίδιου τύπου και στο ίδιο επίπεδο σημαντικότητας α. Αν η H_0 απορρίπτεται m φορές σε επίπεδο σημαντικότητας α θα πρέπει για να είναι ο έλεγχος ακριβής (να έχει σωστή σημαντικότητα) η αναλογία m/M να είναι κοντά στο α. Επίσης μας ενδιαφέρει ο

έλεγχος να έχει μεγάλη ισχύ. Την ισχύ αυτού του ελέγχου μπορούμε υπολογιστικά να την μετρήσουμε και πάλι με προσομοιώσεις όπου τώρα θα πρέπει να γνωρίζουμε ότι η H_0 δεν είναι σωστή και επιπλέον ότι ο έλεγχος έχει σωστή σημαντικότητα. Οι παραπάνω προσομοιώσεις συνήθως ακολουθούνται για να αξιολογήσουμε το στατιστικό που χρησιμοποιείται στον έλεγχο.

<u>p-value (ή significance value)</u>

Αν η περιοχή απόρριψης της H_0 είναι της μορφής T(x)>c (όπου T κατάλληλη στατιστική συνάρτηση T(X)=T(X_1,X_2,…X_n), του τυχαίου δείγματος (X_1,X_2,…X_n), τότε το p-value των τιμών x του δείγματος είναι η πιθανότητα:

$$p - value = P(T(X) > T(x)|H_0) = 1 - F_{T|H_0(T(x))}$$

η οποία μπορεί να θεωρηθεί ότι εκφράζει την πιθανότητα να εμφανιστεί ένα τόσο ή ακόμα πιο ακραίο δείγμα από αυτό που εμφανίστηκε, δεδομένου ότι ισχύει η H_0. Διαισθητικά αν το p-value είναι κοντά στο μηδέν τότε συμπεραίνουμε ότι είναι απίθανο, δεδομένης, της H_0 να εμφανιστεί αυτό το δείγμα και όπως είναι φυσικό φτάνουμε στο συμπέρασμα ότι μάλλον δεν ισχύει η H_0. Πράγματι, αν p-value<α απορρίπτουμε την H_0. Επομένως αντί να εξετάζουμε αν T(x)>c, ισοδύναμα εξετάζουμε αν:

- $p - value < a$: $απορ\ την\ H_0, ενώ$
- $p - value\ \geq a$: $δεν\ απορ\ την\ H_0$

Αν το p-value είναι πάρα πολύ μικρό (π.χ. 0,0001) τότε απορρίπτουμε την H_0 χωρίς δεύτερη σκέψη ενώ αν το p-value είναι σχετικά μικρό (π.χ. κοντά στο 0,05) τότε μπορεί μεν να απορρίψουμε την H_0 αλλά με κάποια επιφυλακτικότητα.

Το πλεονέκτημα από τη χρήση του p-value είναι ότι δεν απορρίπτουμε ή δεχόμαστε απλώς την H_0, αλλά μπορούμε να δούμε και πόσο πιθανή ήταν η εμφάνιση του δείγματος x που πήραμε (υπό την H_0) ενώ επίσης μπορούμε να την συγκρίνουμε άμεσα με όποιο α και αν επιλέξουμε.

7.1.2 ΚΡΙΤΗΡΙΑ ΑΠΟΔΟΣΗΣ ΤΟΥ ΜΟΝΤΕΛΟΥ-CONFUSION MATRIX

Όταν έχουμε να κάνουμε με ένα δυαδικό πρόβλημα ταξινόμησης, υπάρχουν τέσσερις πιθανές εκβάσεις, όταν ένα μοντέλο εφαρμόζεται σε μια παρατήρηση. Ας είναι $(\bar{x}, y) \in R^n \times \{+1, -1\}$ μια παρατήρηση και έστω $f: R^n \times \{+1, -1\}$ να είναι ένα μοντέλο. Στη συνέχεια, έχουμε τις ακόλουθα τέσσερις δυνατότητες όταν το μοντέλο εφαρμόζεται στην παρατήρηση:

$$\hat{f}(\bar{x}) = \begin{cases} +1\ αν\ y{=}+1\ true\ positive \\ -1\ αν\ y{=}+1\ false\ negative \\ +1\ αν\ y{=}-1\ false\ positive \\ -1\ αν\ y{=}-1\ true\ negative \end{cases} \qquad (7.1.2.1)$$

Εάν η τιμή εξόδου $\hat{f}(\bar{x})$ του μοντέλου ταιριάζει με την ετικέτα της παρατήρησης, έχουμε είτε μια αληθώς θετική ή μια αληθώς αρνητική έκβαση. Εάν η τιμή εξόδου του μοντέλου δεν ταιριάζει με την παρατηρούμενη ετικέτα, έχουμε είτε ένα ψευδώς θετικό ή ψευδώς αρνητικό αποτελέσματα, και τα δύο εκ των οποίων είναι λάθος αποτελέσματα.

Ένας πίνακας σύγχυσης για ένα δυαδικό μοντέλο ταξινόμησης είναι ένας 2 × 2 πίνακας, στον οποίο εμφανίζονται οι παρατηρούμενες ετικέτες έναντι των προβλεπόμενων ετικετών

για ένα σύνολο δεδομένων. Ένας τρόπος για να απεικονίσουμε έναν πίνακα σύγχυσης είναι να θεωρήσουμε ότι η εφαρμογή ενός μοντέλου $\hat{f}$ σε μια παρατήρηση $(\bar{x}, y)$ θα μας δώσει δύο ετικέτες. Η πρώτη ετικέτα, y, οφείλεται στην παρατήρηση, και η δεύτερη ετικέτα, $\hat{y} = \hat{f}(\bar{x})$, οφείλεται στην πρόβλεψη του μοντέλου. Δηλαδή, για κάθε παρατήρηση έχουμε ένα ζεύγος ετικετών $(y, \hat{y})$. Αυτό το ζεύγος των ετικετών προσδιορίζει τις συντεταγμένες κάθε παρατήρησης μέσα στον πίνακα σύγχυσης: Η πρώτη ετικέτα διευκρινίζει τη γραμμή του πίνακα και η δεύτερη ετικέτα διευκρινίζει τη στήλη του πίνακα. Ως εκ τούτου, μια παρατήρηση με το ζεύγος ετικέτας $(y, \hat{y})$ θα χαρτογραφηθεί επάνω σε έναν πίνακα, όπως στον παρακάτω πίνακα του ΣΧΗΜΑΤΟΣ (7.1.2):

	Predicted ($\hat{y}$)	
Observed (y)	$+1$	-1
$+1$	True positive (TP)	False negative (FN)
-1	False positive (FP)	True negative (TP)

ΣΧΗΜΑ 7.1.2: Διάταξη ενός πίνακα σύγχυσης

Οι αληθώς θετικές προβλέψεις χαρτογραφούνται στην επάνω αριστερή γωνία του πίνακα σύγχυσης, και οι αληθώς αρνητικές προβλέψεις χαρτογραφούνται στην κάτω δεξιά γωνία του πίνακα. Οι ψευδώς θετικές και ψευδώς αρνητικές αντιστοιχίζονται στην κάτω αριστερά και πάνω δεξιά γωνία του πίνακα, αντίστοιχα. Για ένα μοντέλο το οποίο δεν διαπράττει τυχόν λάθη, όλες οι προβλέψεις θα πρέπει να χαρτογραφηθούν στα πάνω αριστερά και κάτω δεξιά πεδία. Για μοντέλα που διαπράττουν λάθη, βλέπουμε ότι τα λάθη θα αντιστοιχηθούν στον πίνακα σύμφωνα με τον τύπο του σφάλματος που το μοντέλο διαπράττει.

Δεδομένου ενός πίνακα σύγχυσης για ένα μοντέλο όπως στον πίνακα του ΣΧΗΜΑΤΟΣ (7.1.2), μπορούμε να υπολογίσουμε το σφάλμα του μοντέλου μας και την ακρίβεια (accuracy και precision) απευθείας από τον πίνακα, ως εξής:

$$err = \frac{FP+FN}{TP+TN+FP+FN} \qquad (7.1.2.2)$$

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (7.1.2.3)$$

Και

$$precision = \frac{TP}{TP+FP} \qquad (7.1.2.4)$$

αντιστοίχως. Εκτός από αυτές τις μετρικές έχουμε δύο άλλες μετρήσεις που συνήθως χρησιμοποιούνται για να χαρακτηρίσουμε την επίδοση του μοντέλου: ευαισθησία (sensitivity) και ειδικότητα (specificity). Η *ευαισθησία* του μοντέλου ορίζεται ως το πηλίκο των αληθώς θετικών προβλέψεων δια του αθροίσματος όλων των θετικών παρατηρήσεων,

$$sensivity = TRP = \frac{TP}{P} = \frac{TP}{TP+FN} \qquad (7.1.2.5)$$

Η *ειδικότητα* ενός μοντέλου ορίζεται ως οι αληθώς αρνητικές προβλέψεις διαιρούμενες με το άθροισμα όλων των αρνητικών παρατηρήσεων,

$$specificity = TNR = \frac{TN}{N} = \frac{TN}{TN+FP} \qquad (7.1.2.6)$$

Μια ευαισθησία του 1,0 για ένα μοντέλο σημαίνει ότι το μοντέλο προβλέπει όλες τις θετικές παρατηρήσεις σωστά: με άλλα λόγια, το μοντέλο δεν διαπράττει καμία ψευδώς αρνητική πρόβλεψη.
Μία εξειδίκευσή του 1,0 για ένα μοντέλο σημαίνει ότι το μοντέλο προβλέπει όλες τις αρνητικές παρατηρήσεις σωστά, με άλλα λόγια, το μοντέλο δεν διαπράττει καμία ψευδώς θετική πρόβλεψη.

7.2 ΕΙΣΑΓΩΓΗ ΣΤΗΝ R ΚΑΙ ΓΕΝΙΚΗ ΜΕΘΟΔΟΛΟΓΙΑ

7.2.1 ΤΟ ΠΑΚΕΤΟ R

- Η R μπορεί να θεωρηθεί σαν μια εφαρμογή της γλώσσας S, η οποία αναπτύχθηκε στα Bell Laboratories από τους Rick Becker, John Chambers και Allan Wilks, και επίσης αποτελεί τη βάση των S-Plus συστημάτων.
- Υπάρχει μια σημαντική διαφορά στη φιλοσοφία, μεταξύ της S (και ως εκ τούτου της R) και των υπολοίπων κύριων στατιστικών συστημάτων. Στην S, μια στατιστική ανάλυση γίνεται κανονικά ως μια σειρά από βήματα, με τα ενδιάμεσα αποτελέσματα να αποθηκεύονται σε αντικείμενα. Η R θα δώσει ελάχιστα αποτελέσματα (output) και θα τα αποθηκεύσει σε ένα κατάλληλο αντικείμενο

- Η R είναι μια γλώσσα προγραμματισμού που χρησιμεύει κυρίως για την εφαρμογή διαφόρων ‘’κλασικών’’ και σύγχρονων στατιστικών τεχνικών, την ανάλυση και το χειρισμό δεδομένων καθώς επίσης τον υπολογισμό και τη γραφική έκθεση. Μεταξύ άλλων, έχει:
    1. Αποδοτική διαχείριση δεδομένων και ικανότητα αποθήκευσης
    2. Μια συλλογή τελεστών για υπολογισμούς σε διατάξεις, ιδίως σε πίνακες.
    3. Μια μεγάλη, συνεκτική, ολοκληρωμένη συλλογή των ενδιάμεσων εργαλείων για την ανάλυση των δεδομένων.
    4. Γραφικές ικανότητες για την ανάλυση των δεδομένων και την έκθεση είτε απευθείας στον υπολογιστή ή σε έντυπη μορφή, και
    5. Μια καλά αναπτυγμένη, απλή και αποδοτική γλώσσα προγραμματισμού (που ονομάζεται «S»), η οποία περιλαμβάνει υποθετικά (conditionals), βρόχους (loops), οριζόμενες από το χρήστη αναδρομικές συναρτήσεις και ικανότητες εισόδου και εξόδου.

- Μπορεί να χρησιμοποιηθεί είτε με κατευθείαν εντολές που υπάρχουν, είτε με προγράμματα που ο χρήστης μπορεί να προγραμματίσει για επίλυση πιο πολύπλοκων στατιστικών προβλημάτων. Επίσης ο χρήστης μπορεί να χρησιμοποιήσει και έτοιμα προγράμματα τα οποία είναι ενσωματωμένα μέσα σε

πακέτα τα οποία διατίθενται πάλι ελεύθερα. Η ποικιλία τέτοιων προγραμμάτων είναι τεράστια.

- Στο συγκεκριμένο Κεφάλαιο χρησιμοποιούμε την έκδοση 3.1.1

7.2.2 ΒΑΣΙΚΑ ΒΗΜΑΤΑ

Πριν πραγματοποιηθεί η ανάλυση των δεδομένων με τη χρηση των ταξινομητών που αναλύσαμε προηγουμένως στην Εισαγωγή, θα πρέπει να εφαρμόσουμε κάποια βασικά βήματα:

1. Αρχικά εισάγουμε τα δεδομένα στο πρόγραμμα μέσω ενός txt αρχείου.
2. Καθορίζουμε τον τύπο των δεδομένων (factor, data frame).
3. Χωρίζουμε τα δεδομένα σε:

A. **δεδομένα εκπαίδευσης (training set):** με βάση αυτά, γνωρίζοντας την τιμή του αποτελέσματος προσπαθούμε να κατασκευάσουμε ένα μοντέλο πρόβλεψης και χρησιμοποιούνται για την προσαρμογή του μοντέλου.
B. **δεδομένα ελέγχου-εξέτασης (test dataset):** Το μοντέλο που δημιουργήσαμε θα το χρησιμοποιήσουμε στη συνέχεια για να προβλέψουμε το αποτέλεσμα νέων συνόλων δεδομένων εξέτασης (test set), στα οποία σύνολα είναι γνωστές οι τιμές των χαρακτηριστικών αλλά δεν είναι γνωστή η τιμή του αποτελέσματος, δηλαδή η τιμή της τάξης και τα οποία χρησιμοποιούνται για τον υπολογισμό της γενικευμένης τιμής σφάλματος του τελικά επιλεγμένου μοντέλου.

Στα προβλήματα που ακολουθούν, χωρίσαμε, τα δεδομένα σε σύνολο εκπαίδευσης που αποτελείται από το 75% των περιπτώσεων και σύνολο δοκιμής που αποτελείται από το 25% των περιπτώσεων.

7.3 ΠΡΩΤΗ ΕΦΑΡΜΟΓΗ

7.3.1 ΟΡΙΣΜΟΣ ΤΟΥ CREDIT SCORING ΚΑΙ ΤΟ ΠΡΟΒΛΗΜΑ ΠΟΥ ΚΑΛΕΙΤΑΙ ΝΑ ΛΥΣΕΙ

Είναι ένα εξελιγμένο στατιστικά εργαλείο ή σύστημα που χρησιμοποιείται ευρέως από χρηματοοικονομικά ιδρύματα (κυρίως από Τράπεζες) στην προσπάθεια τους να κρίνουν αν θα εγκρίνουν ή να απορρίψουν ένα προσωπικό ή εταιρικό δάνειο. Καθορίζεται έτσι το ρίσκο που υπάρχει κατά τη διάρκεια της χορήγησης ενός δανείου.

Για να κατασκευαστεί ένα τέτοιο μοντέλο, θα πρέπει να αναλυθούν εις βάθος ιστορικά δεδομένα από πιο παλιά δάνεια που τυχόν να έκανε ο πελάτης αλλά και άλλες χρήσιμες πληροφορίες για τον πελάτη-δανειζόμενο οι οποίες μπορούν να ληφθούν από την αίτηση δανείου την οποία θα συμπληρώσει. Στοιχεία όπως το μηνιαίο εισόδημα του πελάτη, τα συνολικά έξοδα του, πόσο καιρό ο πελάτης εργάζεται στην ίδια δουλειά, αν ο πελάτης είναι ιδιοκτήτης σπιτιού ή ενοικιάζει, αν κατέχει δικό του αυτοκίνητο, αν οι λογαριασμοί που διατηρεί σε Τράπεζες δεν είναι χρεωστικοί είναι μερικά από τα στοιχεία που οι

περισσότερες Τράπεζες ζητούν να μάθουν έτσι ώστε να δημιουργήσουν ένα «προφίλ» για το δανειζόμενο.

Στα περισσότερα μοντέλα (scoring systems), ένα ψηλό σκορ συνεπάγεται χαμηλό ρίσκο ενώ η Τράπεζα θέτει ένα όριο (το λεγόμενο cut off) βασιζόμενη στο βαθμό ρίσκου που είναι διαθετημένη να ανεχτεί. Είναι βέβαια στη δική της αρμοδιότητα και θέληση να αποδεχτεί πελάτες με σκορ πιο ψηλό από το όριο που έθεσε και να απορρίψει πελάτες με σκορ μικρότερο από το όριο.

Εδώ και αρκετά χρόνια διάφοροι οργανισμοί και εταιρείες, στην πλειοψηφία τους τράπεζες, προσπαθούν να υιοθετήσουν διάφορες αυτοματοποιημένες τεχνικές, οι οποίες θα τους βοηθήσουν τόσο στις αποφάσεις τους σε θέματα έγκρισης δανείων όσο και στην εξοικονόμηση χρόνου και κόστους. Το μεγαλύτερο πρόβλημα που υπάρχει εδώ και χρόνια, είναι τα χρέη που δημιουργούνται στους χρηματοοικονομικούς οργανισμούς, λόγω της αδυναμίας διάφορων πελατών να ξεπληρώσουν τα δάνεια που τους παραχωρήθηκαν. Για το σκοπό αυτό έχουν αναπτυχθεί διάφορα συστήματα τα οποία διαχωρίζουν τους πελάτες ως κακοπληρωτές ή καλοπληρωτές, βασισμένα σε παλαιότερες συναλλαγές τους. Με τον τρόπο αυτό, τα συστήματα παρέχουν περισσότερες πληροφορίες στους αρμόδιους για τα θέματα δανεισμού, βελτιώνοντας έτσι την ορθότητα των αποφάσεών τους.

7.3.2 ΠΕΡΙΓΡΑΦΗ ΤΗΣ ΕΦΑΡΜΟΓΗΣ ΚΑΙ ΤΟΥ ΣΥΝΟΛΟΥ ΔΕΔΟΜΕΝΩΝ (GERMAN CREDIT DATA)

Τα δεδομένα που κατεβάσαμε (German Data Set) προέρχονται από την τοποθεσία:

UCI Machine Learning Repository, Professor Dr. Hans Hofmann, Institut f"ur Statistik und "Okonometrie, Universit"at Hamburg, FB Wirtschaftswissenschaften, Von-Melle-Park 5, 2000 Hamburg 13

Προτείνεται μια προσέγγιση εξόρυξης δεδομένων (Data Mining) για τον πιστωτικό έλεγχο για δάνειο αιτητών σε μια γερμανική τράπεζα. Το σύνολο δεδομένων που χρησιμοποιήσαμε έχει 1000 καταγραφές (instances), με 20 επεξηγηματικές μεταβλητές (7 αριθμητικές και 13 κατηγορηματικές) και 1 μεταβλητή απόκρισης (response) με ένα δυαδικό (binary) αποτέλεσμα (0 ή 1). Από τις 1.000 παρατηρήσεις, 700 (ή 70,0%) έχουν μικρό πιστωτικό κίνδυνο (καλοπληρωτές=1) και 300 (ή 30,0%) έχουν μεγάλο πιστωτικό κίνδυνο (κακοπληρωτές=0). Τα 20 χαρακτηριστικά (επεξηγηματικές μεταβλητές) που είναι διαθέσιμα για την κατασκευή μοντέλων βαθμολόγησης πιστοληπτικής ικανότητας, περιλαμβάνουν δημογραφικά χαρακτηριστικά όπως π.χ. το φύλο και η ηλικία καθώς και μερικά λογιστικά στοιχεία (π.χ. ιστορικό , το ποσό πίστωσης που ζητά ο αιτητής). Σκοπός της γερμανικής τράπεζας είναι να αναπτύξει ένα μοντέλο βαθμολόγησης πιστοληπτικής ικανότητας ώστε να προβλέψει τον πιστωτικό κίνδυνο των αιτούντων δανείου. Προφανώς, λοιπόν, έχουμε ένα πρόβλημα ταξινόμησης δύο κλάσεων. Η τράπεζα προτίθεται να αναπτύξει το μοντέλο κατά το χρόνο επεξεργασίας των αιτήσεων για παραχώρηση δανείων. Η κατασκευή του μοντέλου βαθμολόγησης πιστοληπτικής ικανότητας απαιτεί την κατασκευή μοντέλων πρόβλεψης. Για το σκοπό αυτό χρησιμοποιείται μια τεχνική εξόρυξης δεδομένων, αυτή των SVMs. Έχουμε, λοιπόν, τον παρακάτω πίνακα δεδομένων:

ΠΙΝΑΚΑΣ ΔΕΔΟΜΕΝΩΝ

X1	**Ισοζύγιο τρεχουσών συναλλαγών** σε γερμανικά μάρκα (DM)
X2	**Διάρκεια σε μήνες**
X3	**Ιστορικό πελάτη**
X4	**Σκοπός δανείου**
X5	**Ποσό πίστωσης**
X6	**Λογαριασμοί καταθέσεων / ταμιευτηρίου**
X7	**Διάρκεια εργασίας** σε χρόνια
X8	**Δόση δανείου σε ποσοστό του διαθέσιμου εισοδήματος**
X9	**Συζυγική κατάσταση**
X10	**Άλλοι πιστωτές / εγγυητές**
X11	**Διάρκεια ζωής σε μόνιμη κατοικία** σε χρόνια
X12	**Περιουσιακά στοιχεία που διαθέτει ο αιτητής**
X13	**Ηλικία**
X14	**Σχέδια αποπληρωμής δανείου**
X15	**Στέγαση**
X16	**Αριθμός προηγούμενων δανείων σε αυτή την τράπεζα**
X17	**Απασχόληση**
X18	**Πλήθος ατόμων που καλείται να συντηρήσει**
X19	**Τηλέφωνο**
X20	**Αλλοδαπός εργαζόμενος**
Y	**Χαρακτηρισμός του πελάτη ως ''καλοπληρωτή''(=1) ή ''κακοπληρωτή''(=0)**

Όπως προαναφέρθηκε, πρόκειται για ένα σύνολο δεδομένων που αποτελείται από 1000 εγγραφές, που χωρίζονται με τυχαίο τρόπο σε σύνολο εκπαίδευσης που αποτελείται από το 75% των περιπτώσεων (training set : $\approx 75\% \times 1000 = 750$) και σύνολο δοκιμής που

αποτελείται από το 25% των περιπτώσεων ($test\ set$: $\approx 25\%1000 = 250$) για να εκτιμήσουμε την απόδοση των ταξινομητών σε νέα δεδομένα.

7.3.3 ΧΕΙΡΙΣΜΟΣ ΤΟΥ ΣΥΝΟΛΟΥ ΔΕΔΟΜΕΝΩΝ (GERMAN CREDIT DATA) ΣΤΗΝ R

Κάνοντας χρήση των πακέτων e1071, caret, θα χρησιμοποιήσουμε τους έτοιμους αλγόριθμούς που υπάρχουν για ταξινόμηση με Μηχανές Διανυσματικής Υποστήριξης (SVMs). Πρώτα, θα εκπαιδεύσουμε τον ταξινομητή μας και έπειτα θα τον αξιολογήσουμε. Για το σκοπό αυτό, θα εξετάσουμε τη συμπεριφορά του ταξινομητή μας με καθέναν από τους τέσσερις πυρήνες: γραμμικό (linear), πολυωνυμικό (polynomial), ακτινωτό (radial), και σιγμοειδή (sigmoid) τόσο στο σύνολο εκπαίδευσης όσο και στο σύνολο δοκιμής. Στη συνέχεια, με κατάλληλα μέτρα (Confusion Matrix) θα εκτιμήσουμε την απόδοσή του.

Κάνοντας χρήση της C-ταξινόμησης (C-classification) και διατηρώντας την παράμετρο του κόστους σταθερή και ίση με τη μονάδα (cost=1) παίρνουμε τα αποτελέσματα του πίνακα του ΣΧΗΜΑΤΟΣ (7.3.1):

	Accuracy		Sensitivity		Specificity	
	Train	Test	train	Test	train	Test
Linear	0.7827	0.768	0.9048	0.9029	0.4978	0.4533
Polynomial	0.708	0.696	1.0000	0.9943	0.02667	0.0000
Radial	0.7627	0.752	0.9848	0.9771	0.2444	0.2267
Sigmoid	0.712	0.72	0.99429	1.0000	0.05333	0.06667

ΣΧΗΜΑ 7.3.1: Πίνακας αξιολόγησης

Από το ΣΧΗΜΑ (7.3.1) προκύπτει ότι:

- Σε κάθε περίπτωση, η ακρίβεια (accuracy) του μοντέλου είναι ικανοποιητική. Ο γραμμικός (linear) πυρήνας, στο σύνολο εκπαίδευσης, μας δίνει το καλύτερο αποτέλεσμα (0.7827) καθώς επίσης και το δεύτερο καλύτερο αποτέλεσμα στο σύνολο δοκιμής (0.768).
- Η δείκτης ευαισθησίας (sensitivity) επίσης του μοντέλου, σε κάθε περίπτωση, είναι αξιοσημείωτος, καθώς ένας δείκτης πάνω από 0.9 για ένα μοντέλο σημαίνει ότι το μοντέλο διαπράττει ελάχιστες ψευδώς αρνητικές προβλέψεις.
- Η εξειδίκευση (specificity) του μοντέλου θα μπορούσε να μας προβληματίσει, καθώς οι τιμές που παρατηρούμε είναι χαμηλές (η υψηλότερη τιμή είναι 0.4978). Με άλλα λόγια θα λέγαμε ότι το μοντέλο μας διαπράττει αρκετές ψευδώς θετικές προβλέψεις.

Με μία πιο προσεκτική ματιά, μπορούμε να συμπεράνουμε ότι οι χαμηλές τιμές οφείλονται στην ανισορροπία του συνόλου των δεδομένων. Πιο συγκεκριμένα, η δίτιμη μεταβλητή απόκρισης y (με 1 και 0 , για τους ''καλοπληρωτές'' και τους ''κακοπληρωτές αντιστοίχως) χωρίζεται σε δύο κλάσεις όπου η μία είναι πολύ μεγαλύτερη καθώς από τις 1.000 παρατηρήσεις, 700 (ή 70,0%) έχουν μικρό πιστωτικό κίνδυνο (καλοπληρωτές=1) και 300 (ή 30,0%) έχουν μεγάλο πιστωτικό κίνδυνο (κακοπληρωτές=0).

Στο σημείο αυτό θα μπορούσαμε να πούμε ότι ο γραμμικός πυρήνας είναι οι καλύτερος για την ταξινόμηση των δεδομένων μας καθώς παρουσιάζει αυξημένη ακρίβεια και ευαισθησία και την μεγαλύτερη ειδικότητα μεταξύ των υπολοίπων.

7.4 ΔΕΥΤΕΡΗ ΕΦΑΡΜΟΓΗ

7.4.1 ΤΟ ΣΥΝΟΛΟ ΔΕΔΟΜΕΝΩΝ IRIS DATA SET

Αυτή είναι ίσως η πιο γνωστή βάση δεδομένων που μπορεί να βρεθεί στη βιβλιογραφία της αναγνώρισης προτύπων (pattern recognition). Το άρθρο του Fisher είναι κλασικό σε αυτόν τον τομέα και γίνονται συχνά αναφορές σε αυτό μέχρι και σήμερα. (Βλέπε Duda & Hart, για παράδειγμα.)

Τα δεδομένα που θα χρησιμοποιηθούν στη συνέχεια διαφέρουν από αυτά που παρουσιάζονται στο άρθρο του Fisher. Το τριακοστό πέμπτο δείγμα θα πρέπει να είναι: 4.9, 3.1, 1.5 ,0.2, ‘’Iris Setosa’’ όπου το σφάλμα είναι στο τέταρτο χαρακτηριστικό. Ενώ το τριακοστό όγδοο δείγμα θα πρέπει να είναι: 4.9, 3.6, 1.4, 0.1, ‘’Iris Setosa’’ όπου τα λάθη είναι στο δεύτερο και τρίτο χαρακτηριστικό.

7.4.2 ΠΕΡΙΓΡΑΦΗ ΤΗΣ ΕΦΑΡΜΟΓΗΣ ΚΑΙ ΤΟΥ ΣΥΝΟΛΟΥ ΔΕΔΟΜΕΝΩΝ (IRIS DATA SET)

Τα δεδομένα που κατεβάσαμε (Iris Data Set) προέρχονται από την τοποθεσία:

UCI Machine Learning Repository, Creator: R.A. Fisher, Donor: Michael Marshall (MARSHALL%PLU **'@'** io.arc.nasa.gov)

Το σύνολο δεδομένων που χρησιμοποιήσαμε έχει 150 εγγραφές (instances). Τα δεδομένα περιλαμβάνουν 3 κλάσεις (‘’Iris Setosa’’, ‘’Iris Versicolour’’, ‘’Iris Virginica’’) των 50 περιπτώσεων η κάθε μία, όπου κάθε κλάση αναφέρεται σε ένα είδος του φυτού ίρις. Μία κλάση είναι γραμμικά διαχωρίσιμη από τις άλλες 2, οι άλλες, όμως, δεν είναι γραμμικά διαχωρίσιμες μεταξύ τους. Υπάρχουν 4 επεξηγηματικές μεταβλητές (αριθμητικές) και 1 μεταβλητή απόκρισης (response). Σκοπός μας είναι να προβλέψουμε την κλάση (κατηγορία) του φυτού ίρις. Για το σκοπό αυτό χρησιμοποιείται μια τεχνική εξόρυξης δεδομένων, οι μηχανές διανυσματικής υποστήριξης (SVMs). Έχουμε, λοιπόν, τον παρακάτω πίνακα δεδομένων:

ΠΙΝΑΚΑΣ ΔΕΔΟΜΕΝΩΝ

X1	**Μήκος σέπαλου**
X2	**Πλάτος σέπαλου**
X3	**Μήκος πέταλου**
X4	**Πλάτος πέταλου**
Y	**Κατηγορία φυτού (Setosa, Versicolour, Virginica)**

Όπως προαναφέρθηκε, πρόκειται για ένα σύνολο δεδομένων που αποτελείται από 150 εγγραφές, που χωρίζονται με τυχαίο τρόπο σε σύνολο εκπαίδευσης που αποτελείται από το

75% των περιπτώσεων (training set : $\approx 75\% \times 150 \cong 113$) και σύνολο δοκιμής που αποτελείται από το 25% των περιπτώσεων ($test\ set$: $\approx 25\%150 \cong 37$) για να εκτιμήσουμε την απόδοση των ταξινομητών σε νέα δεδομένα.

7.4.3 ΧΕΙΡΙΣΜΟΣ ΤΟΥ ΣΥΝΟΛΟΥ ΔΕΔΟΜΕΝΩΝ (IRIS DATA SET) ΣΤΗΝ R

Κάνοντας χρήση των πακέτων e1071, caret, θα χρησιμοποιήσουμε τους έτοιμους αλγόριθμούς που υπάρχουν για ταξινόμηση με Μηχανές Διανυσματικής Υποστήριξης (SVMs). Πρώτα, θα εκπαιδεύσουμε τον ταξινομητή μας και έπειτα θα τον αξιολογήσουμε. Για το σκοπό αυτό, θα εξετάσουμε τη συμπεριφορά του ταξινομητή μας με καθέναν από τους τέσσερις πυρήνες: γραμμικό (linear), πολυωνυμικό (polynomial), ακτινωτό (radial), και σιγμοειδή (sigmoid) τόσο στο σύνολο εκπαίδευσης όσο και στο σύνολο δοκιμής. Στη συνέχεια, με κατάλληλα μέτρα (Confusion Matrix) θα εκτιμήσουμε την απόδοσή του.

Κάνοντας χρήση της C-ταξινόμησης (C-classification) και διατηρώντας την παράμετρο του κόστους σταθερή και ίση με τη μονάδα (cost=1) παίρνουμε τα αποτελέσματα του πίνακα του ΣΧΗΜΑΤΟΣ (7.4.1):

	Accuracy	
	Train	Test
Linear	0.9823	0.973
Polynomial	0.9469	0.9189
Radial	0.9735	0.973
Sigmoid	0.9115	0.8378

ΣΧΗΜΑ 7.4.1: Πίνακας αξιολόγησης

Από τον πίνακα του ΣΧΗΜΑΤΟΣ (7.4.1) μπορούμε να παρατηρήσουμε ότι η ακρίβεια του ταξινομητή είναι αρκετά υψηλή. Παρόλα αυτά, μεγαλύτερη ακρίβεια έχουμε στην περίπτωση του γραμμικού πυρήνα (0.9823) και ακολουθεί ο ακτινωτός (0.9735), στο σύνολο εκπαίδευσης και στις δύο περιπτώσεις. Ένας επιπλέον έλεγχος της απόδοσης του ταξινομητή (βλέπε πίνακα ΣΧΗΜΑΤΟΣ (7.4.2)) καθιστά σαφή την υπεροχή του γραμμικού πυρήνα έναντι του ακτινωτού. Όπως βλέπουμε στον πίνακα του ΣΧΗΜΑΤΟΣ (7.4.2), ο γραμμικός πυρήνας παρουσιάζει μεγαλύτερη ευαισθησία και εξειδίκευση συνολικά και στις τρεις κλάσεις:

	Κλάση 1		Κλάση 2		Κλάση 3	
	Linear	Radial	linear	radial	linear	Radial
Sensitivity	1.0000	1.0000	0.9750	0.9500	0.9706	0.9706
Specificity	1.0000	1.0000	0.9863	0.9863	0.9873	0.9747

ΣΧΗΜΑ 7.4.2: Πίνακας αξιολόγησης

8

ΣΥΝΟΨΗ-ΣΥΜΠΕΡΑΣΜΑΤΑ

Η πρόοδος στην ψηφιακή απόκτηση δεδομένων και στην τεχνολογία αποθήκευσης έχει σαν αποτέλεσμα την αύξηση των τεραστίων βάσεων δεδομένων. Η ανάλυση αυτών των βάσεων για την εύρεση ανυποψίαστων σχέσεων και τη σύνοψη των δεδομένων με νέους τρόπους που να είναι κατανοητοί όσο και χρήσιμοι στον ιδιοκτήτη των δεδομένων, αποτελεί την Εξόρυξη Δεομένων (Data Mining). Ο κύριος στόχος του D.M είναι η εξαγωγή νέων πληροφοριών από τα δεδομένα. Η ανακάλυψη αυτής της γνώσης γίνεται με τεχνικές που διακρίνονται στις μεθόδους με επίβλεψη (supervised) και στις μεθόδους χωρίς επίβλεψη (unsupervised). Στη μάθηση με επίβλεψη όλες οι μέθοδοι εκμάθησης απαρτίζονται από το στάδιο της επιλογής μοντέλου (με βάση συγκεκριμένα μέτρα απόδοσης) και από το στάδιο εκτίμησης (εκτίμηση του σφάλματος πρόβλεψης του τελικού μοντέλου). Για αυτά τα δύο στάδια έχουν αναπτυχθεί αρκετά κριτήρια και μέθοδοι. Οι μέθοδοι με επίβλεψη χωρίζονται στις υποκατηγορίες της άμεσης και της έμμεσης μάθησης, στη δεύτερη κατηγορία ανήκουν οι μηχανές διανυσματικής υποστήριξης (SVMs).

Οι μηχανές διανυσματικής υποστήριξης (SVMs), είναι μια νέα μέθοδος ταξινόμησης, η οποία πετυχαίνεται από μια γραμμική ή από μια μη γραμμική διαχωριστική επιφάνεια στο χώρο εισόδου του συνόλου δεδομένων και η οποία έχει ως στόχο την ελαχιστοποίηση της προσδοκίας της εξόδου του σφάλματος δοκιμής. Τα SVMs καταγράφουν ένα δεδομένο σύνολο δυαδικών επιγεγραμμένων δεδομένων εκπαίδευσης σε ένα χώρο χαρακτηριστικών υψηλής διάστασης και διαχωρίζουν τις δύο κλάσεις των δεδομένων με ένα υπερεπίπεδο μεγίστου περιθωρίου. Βασικές έννοιες για την SVM προσέγγιση είναι η δυαδικότητα και οι πυρήνες. Το εγγενώς δυικό SVM, όμως, εφαρμόζεται και σε προβλήματα ταξινόμησης με περισσότερες από δύο κλάσεις, μέσω των μεθόδων one-against-one, one-against-all και error-correcting output codes.

Η απόδοση γενίκευσης της μαθησιακής μεθόδου σχετίζεται με την ικανότητα πρόβλεψης των ανεξαρτήτων δοκιμών δεδομένων. Η αξιολόγησης αυτής της απόδοσης είναι εξαιρετικά σημαντική στην πράξη καθώς καθοδηγεί την επιλογή της μαθησιακής μεθόδου ή του μοντέλου και μας δίνει ένα μέτρο της ποιότητας του επιλεχθέντος μοντέλου. Για την αποτίμηση των αλγορίθμων εκμάθησης καθώς και για την επιλογή και εκτίμηση του μοντέλου έχει αναπτυχθεί μια μεγάλη ποικιλία κριτηρίων και μεθόδων, όπως είναι τα εξής: AIC, BIC, διασταυρωμένη επικύρωση, διασταυρωμένα επικυρωμένο ποσοστό σφάλματος, ένδειξη υπολογισμένη με τη μέθοδο Laplace, ξα-εκτίμηση, GRM, MDL, bootstrap.

Στη συνέχεια εισήχθησαν κάποια νέα κριτήρια πληροφορίας για τα δυαδικά SVMs καθώς επίσης και για τα SVMs περισσοτέρων κλάσεων. Αρχικά, εισήχθη το KRIC (κριτήριο πληροφορίας συστηματοποίησης πυρήνα) που αντιστοιχεί στο κριτήριο RIC (κριτήριο συστηματοποίησης πληροφορίας) και το οποίο χρησιμοποιείται για τη ρύθμιση της

παραμέτρου στην KLR και στα SVMs. Προτάθηκε, συγκεκριμένα, η ρύθμιση των παραμέτρων συστηματοποίησης για την KLR με τη χρήση του KRIC και το αποτέλεσμα χρησιμοποιήθηκε άμεσα στα SVMs. Παρουσιάστηκε, επίσης, η προσέγγιση του KRIC με τη χρήση της μεθόδου Nystrom, η οποία ελαχιστοποιεί το υπολογιστικό του κόστος. Επίσης, διεξήχθησαν κάποια πειράματα για τη ρύθμιση της παραμέτρου πυρήνα από το KRIC. Στη συνέχεια, εισήχθησαν τα κριτήρια πληροφορίας SVMICa και SVMICb τα οποία μοιράζονται κάποιες ιδιότητες με τα AIC και BIC, αντιστοίχως. Έπειτα, τονίστηκαν τα πλεονεκτήματά τους (σε σχέση με τα ήδη αναπτυγμένα κριτήρια) που προκύπτουν από μια μελέτη προσομοίωσης, για τη σύγκριση της απόδοσης των νέων κριτηρίων σε σχέση με τις μεθόδους για την επιλογή μοντέλου στα SVMs, και από τη σύγκριση διαφόρων τεχνικών σε μερικά σύνολα αναφοράς στον πραγματικό κόσμο.

Όπως και η επιλογή μοντέλου για τα δυαδικά SVMs, έτσι και η ομόλογή της για τα multiclass SVMs, χρησιμοποιείται για την επιλογή των παραμέτρων της συνάρτησης πυρήνα και της παραμέτρου ρύθμισης που εξισορροπεί το σφάλμα εκπαίδευσης και τη μηχανική πολυπλοκότητα. Η προσέγγιση που προτάθηκε είναι η ανάπτυξη νέων κριτηρίων βασισμένων στις αρχές των επιτυχημένων κριτηρίων στα δυαδικά SVMs για multiclass ρύθμιση. Στη συνέχεια, έγινε χρήση του φράγματος ακτινωτού περιθωρίου, και συγκεκριμένα, το πρώτο Κριτήριο που εισήχθη, αναπτύχθηκε βασισμένο στο ζευγαρωτό συνδυασμό μεταξύ των φραγμάτων ακτινωτού περιθωρίου και ενός συνόλου δυαδικών SVMs για την επιλογή μοντέλου, ενώ το δεύτερο Κριτήριο που εισήχθη, όρισε νέα ακτίνα και περιθώριο για να συμβιβαστούν οι κλάσεις. Με βάση τα πειράματα που διεξήχθησαν, καταλήξαμε στο ότι και τα δύο Κριτήρια παρουσιάζουν καλή απόδοση επιλογής μοντέλου, ενώ δουλεύουν καλά και με το σφαιρικό και με τον ελλειψοειδή πυρήνα. Επίσης και τα δύο νέα Κριτήρια βοηθούν όλους τους multiclass SVM ταξινομητές να επιτύχουν αρκετά καλή απόδοση ταξινόμησης. Η μεταξύ τους σύγκριση βρίσκει το πρώτο Κριτήριο να δείχνει περιθωριακά καλύτερη απόδοση.

Τέλος, αναλύθηκαν ο Έλεγχος Στατιστικών Υποθέσεων και τα κριτήρια απόδοσης του μοντέλου-Confusion Matrix και στη συνέχεια αξιολογήθηκε πειραματικά η εφαρμογή των μεθόδων Διανυσματικής Υποστήριξης σε προβλήματα ταξινόμησης δύο και τριών κλάσεων. Για να παραχθεί μια αμερόληπτη εκτίμηση, για την ποιότητα ταξινόμησης του κάθε μοντέλου, χρησιμοποιώντας τη μέθοδο της διάκρισης, οι τιμές των κριτηρίων απόδοσης υπολογίζονται με βάση ένα σύνολο δεδομένων (το σύνολο δοκιμής) που δε χρησιμοποιήθηκε στη διαδικασία μοντελοποίησης και αφέθηκε στην άκρη για αυτό το σκοπό. Στη συνέχεια, εξετάστηκε η συμπεριφορά του ταξινομητή με καθένα από τους τέσσερις πυρήνες: το γραμμικό, τον πολυωνυμικό, τον ακτινωτό και το σιγμοειδή. Στις εφαρμογές που διεξήχθησαν λάβαμε υπόψη μας ότι, ένας ταξινομητής θα πρέπει να παρέχει υψηλές τιμές των κριτηρίων απόδοσης του Confusion Matrix και πιο συγκεκριμένα της ακρίβειας, της ευαισθησίας και της ειδικότητας.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Blomgren P. (2013). Quasi Newton Methods-The BFGS methods.Numerical Optimization. Lectures #18. pp. 19.

[2] Claeskens G., Croux C., Van Kerckhoven J. (2008). An information criterion for variable selection in Support Vector Machines. Journal of Machine Learning Research 9 . pp. 541-558

[3] Christianini N. (2001). Support Vector and Kernel Machines. Tutorial at http://wwwsupport-vector.net/tutorial.html. ICML 2001. pp. 28.

[4] Christianini N., Saw-Taylor J. (2000). An introduction to Support Vector Machines and other kernel-based methods. Cambridge University Press. pp. 137-138.

[5] Fawcett, T., (2003). ROC Graphs: Notes and Practical Considerations for Data Mining Researchers, Intelligent Enterprise Technologies Laboratory HP Laboratories Palo Alto HPL-2003-4 January 7th.

[6] Fletcher, T. (2009). Support vector machines explained. *Tutorial paper., Mar*. pp.28

[7] Gao F. (1999). Iterated ranGACV: a computational Proxy for the Comparative Lullback-Leibler Distance. TECHNICAL REPORT NO.1011. Department of Statistics University of Winscosin-Madison. pp.1-3.

[8] Hamel, L. H. (2011). *Knowledge discovery with support vector machines* (Vol. 3). John Wiley & Sons. pp. 106-107

[9] Hand D., Mannila H. , Smyth P. (2001). Principles of Data Mining. A Bradford Book, The MIT Press Cambridge, Massachusetts London England. pp. 5-7

[10] Hastie T., Tibshirani R., (2009). Statistical Learning and Data Mining III, Cross Validation and Bootstrap, Georgetown University Conference Center, Georgetown DC. pp. 3-4.

[11] Hastie T., Tibshirani R., Friedman J. (2008). The Elements of Statistical Learning. Second Edition. Springer Series in Statistics. pp. 132-135, 219-237, 249-252, 417-421, 423-428.

[12] Hlavac V. (2002) Vapnik-Chervovenkis learning theory. Talk at http://cmp.felk.cvut/~hlavac. pp. 8-22.

[13] Joachims T. (2000). *Estimating the Generalization Performance of a SVM Efficiently*. Proceedings of the International Conference on Machine Learning (ICML). Morgan Kaufman.

[14] Kearns M., Mansour Y., Andrew Y.NG, Ron D. (1995). An experimental and Theoretical Comparison of Model Selection Methods. Colt '95 Proceedings of the eighth annual Conference on Computational Learning Theory. ACM, New York, USA. pp. 21-30.

[15] Kobayashi K., Komaki F.(2005). Information Criteria for Kernel Machines. Mathematical Engineering Technical Reports. METR 2005-23.

[16] Lin, C. J. (2006). A guide to support vector machines. *Department of Computer Science, National Taiwan University*. pp. 25-27

[17] Refaeilzadeh P., Tang L. and Liu H. (2009). Cross Validation. Encyclopedia of Database Systems. Springer.

[18] Rifkin R. (2008). Multiclass Classification. 9.520 Class 06. pp. 12, 18, 23.

[19] Tin-Yau Kwok J. (2000). The Evidence Framework Applied to Support Vector Machines. IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. XX, NO. X.

[20] Vanderbeghe L. (2013-2014). Quasi-Newton methods. Optimization Methods for large-scale systems . EE236C. pp. 5.

[21] Wang L., Xue P., LukChan K. (2008). Two Criteria for Model Selection in Multiclass Support Vector Machines. IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART B: CYBERNETICS, VOL. 38, NO. 6.

[22] Wahba G. (2008). Support Vector Machines, Reproducing Kernel Hilbert Spaces and the Randomized GACV. TECHNICAL REPORT NO.984rr. Department of Statistics University of Winscosin-Madison.

[23] Witten, I. H., & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann. pp. 41-60

[24] Xia F. (2008). Reducing multiclass to Binary. Ling 572, Advanced Statistical Methods for NLP. University of Washington Courses. pp. 10-17.

[25] Yaeger L. (2006). Intro to Genetic Algorithms. Artificial Life as an approach of Artificial Intelligence. Lecture 03, I400/I590. pp. 4-5

[26] Youn, E. S. (2002). *Feature Selection in Support Vector Machines*. Thesis (M. Sc) - University of Florida. pp. 4

[27] B.Zabinsky Z. (2003). Stochastic Adaptive Search for Global Optimization. XVIII. Kluwer Academic Publishers. pp. 10-11.

[28] Zhang X., Lu X., Shi Q., Xu X., H. E Leung, N Harris L., D Iglehart J., Miron A., Liu J.S and Wong W.H. (2006). Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data. BMC Bioinformatics, BioMed Central. pp. 9-13.

[29]Γιαννούλη Δ. (2014). Εφαρμογές των Μηχανών Διανυσματικής Υποστήριξης σε Προβλήματα Ταξινόμησης και Παλινδρόμησης. Διπλωματική Εργασία στο ΕΜΠ.

[30] Ζαρωτιάδου, Η.Α. (2011). Εφαρμογή των SVMs και k-NN Τεχνικών σε προβλήματα Στατιστικού Ελέγχου Ποιότητας. Διπλωματική Εργασία στο ΕΜΠ.

[31]] Τοπούζης Κ. (2012). Τεχνικές εξόρυξης δεδομένων και Λογιστική Παλινδρόμηση στη βαθμολόγηση πιστοληπτικής ικανότητας (credit scoring). Διπλωματική Εργασία στο ΕΜΠ.

R TUTORIAL

[32] Kuhn, M. (2013). A Short Introduction to the caret Package.

[33] Meyer, D., Dimitriadou, E., Hornik, K., Lin, C., & Meyer, M. D. (2014). Package ‘e1071’.

[34] Venables, W. N., Smith, D. M., & R Development Core Team. (2005). An introduction to R.

ΔΙΑΔΥΚΤΙΑΚΟΙ ΙΣΤΟΤΟΠΟΙ

[35] http://archive.ics.uci.edu/ml/

[36] http://www.learningace.com/

[37] http://www.wikipedia.org

ΠΑΡΑΡΤΗΜΑΤΑ

ΠΑΡΑΡΤΗΜΑ Α

ΑΠΟΔΕΙΞΗ ΤΗΣ ΣΧΕΣΗΣ (5.1.7.2)

Αποδεικνύουμε τη σχέση (5.1.7.2):

$$trace\left[\left(\widetilde{K}diag(t)+\lambda I_l\right)^{-1}\left(\widetilde{K}diag(m)^2-\frac{1}{l}\widetilde{K}mm^T\right)\right]$$
$$=trace\left[\left(\widetilde{U}^T diag(t)\widetilde{U}+\lambda\tilde{\Lambda}^{-1}\right)^{-1}\widetilde{U}^T\left(diag(m)^2-\frac{1}{l}mm^T\right)\widetilde{U}\right]$$

Έστω $M \coloneqq diag(m)$ $\kappa\alpha\iota$ $T \coloneqq diag(t)$. Χρησιμοποιώντας το λήμμα της αντιστροφής του πίνακα,

$$trace\left[(KT+\lambda I_l)^{-1}K\left(M^2-\frac{1}{l}mm^T\right)\right]$$
$$=trace\left[\frac{1}{\lambda}T^{-1}\left\{T-T\widetilde{U}\left(\lambda\tilde{\Lambda}^{-1}+\widetilde{U}^T T\widetilde{U}\right)^{-1}\widetilde{U}^T T\right\}K\left(M^2-\frac{1}{l}mm^T\right)\right]$$
$$=\frac{1}{\lambda}\left[\left\{I_l-\widetilde{U}\left(\lambda\tilde{\Lambda}^{-1}+\widetilde{U}^T T\widetilde{U}\right)^{-1}\widetilde{U}^T T\right\}K\left(M^2-\frac{1}{l}mm^T\right)\right]$$
$$=\frac{1}{\lambda}trace\left[\widetilde{U}\tilde{\Lambda}^{-1}\widetilde{U}^T\left(M^2-\frac{1}{l}mm^T\right)\right]-\frac{1}{\lambda}trace\left[\left(\lambda\tilde{\Lambda}^{-1}+\widetilde{U}^T T\widetilde{U}\right)^{-1}\widetilde{U}^T T\widetilde{U}\tilde{\Lambda}\widetilde{U}^T\left(M^2-\right.\right.$$
$$1lmmTU$$
(5.1.7.3)

Ο δεύτερος όρος της πιο δεξιάς πλευράς αξιολογείται ως εξής:

$$\frac{1}{\lambda}trace\left[\left(\lambda\tilde{\Lambda}^{-1}+\widetilde{U}^T T\widetilde{U}\right)^{-1}\widetilde{U}^T T\widetilde{U}\tilde{\Lambda}\widetilde{U}^T\left(M^2-\frac{1}{l}mm^T\right)\widetilde{U}\right]$$

$$= trace\left[\left(\lambda\tilde{\Lambda}^{-1} + \tilde{U}^T T\tilde{U}\right)^{-1}\left(\lambda\tilde{\Lambda}^{-1} + \tilde{U}^T T\tilde{U}\right)\tilde{\Lambda}\tilde{U}^T\left(M^2 - \frac{1}{l}mm^T\right)\tilde{U}\right]$$
$$-trace\left[\lambda\left(\lambda\tilde{\Lambda}^{-1} + \tilde{U}^T T\tilde{U}\right)^{-1}\tilde{U}^T\left(M^2 - \frac{1}{l}mm^T\right)\tilde{U}\right]$$
$$= trace\left[\tilde{\Lambda}\tilde{U}^T\left(M^2 - \frac{1}{l}mm^T\right)\tilde{U}\right] - trace\left[\lambda\left(\lambda\tilde{\Lambda}^{-1} + \tilde{U}^T T\tilde{U}\right)^{-1}\tilde{U}^T\left(M^2 - \frac{1}{l}mm^T\right)\tilde{U}\right]$$

Αντικαθιστώντας αυτή στην (5.1.7.3),

$$trace\left[(KT + \lambda I_l)^{-1}K\left(M^2 - \frac{1}{l}mm^T\right)\right]$$
$$\frac{1}{\lambda}trace\left[\tilde{U}\tilde{\Lambda}^{-1}\tilde{U}^T\left(M^2 - \frac{1}{l}mm^T\right) - \frac{1}{\lambda}trace\left[\tilde{\Lambda}\tilde{U}^T\left(M^2 - \frac{1}{l}mm^T\right)\tilde{U}\right]\right]$$
$$+trace\left[\left(\lambda\tilde{\Lambda}^{-1} + \tilde{U}^T T\tilde{U}\right)^{-1}\tilde{U}^T\left(M^2 - \frac{1}{l}mm^T\right)\tilde{U}\right]$$
$$= trace\left[\left(\lambda\tilde{\Lambda}^{-1} + \tilde{U}^T T\tilde{U}\right)^{-1}\tilde{U}^T\left(M^2 - \frac{1}{l}mm^T\right)\tilde{U}\right]$$

Ως εκ τούτου, η (5.1.7.2) ισχύει.

ΠΑΡΑΡΤΗΜΑ Β

ΣΕΙΡΙΑΚΗ ΕΛΑΧΙΣΤΗ ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ (SEQUENTIAL MINIMAL OPTIMISATION(SMO))

Ο αλγόριθμος SMO προέρχεται από τη λήψη της ιδέας της αποσύνθεσης στα άκρα και τη βελτιστοποίηση ενός ελάχιστου υποσυνόλου μόνο δύο σημείων σε κάθε επανάληψη. Η ισχύς αυτής της τεχνικής βρίσκεται στο γεγονός ότι το πρόβλημα βελτιστοποίησης για τα δύο σημεία δεδομένων παραδέχεται μια αναλυτική λύση, εξαλείφοντας την ανάγκη χρήσης ενός βελτιστοποιητή επαναληπτικού τετραγωνικού προγράμματος.

Η προϋπόθεση , ότι η συνθήκη $\sum_{i=1}^{l} a_i y_i = 0$, ενισχύεται μέσω του ότι οι επαναλήψεις υπαινίσσονται ότι ο πιο μικρός αριθμός των πολλαπλασιαστών που μπορούν να βελτιστοποιηθούν σε κάθε βήμα είναι δύο: οποτεδήποτε ένας πολλαπλασιαστής ενημερώνεται, τουλάχιστον ένας άλλος πολλαπλασιαστής πρέπει να προσαρμοστεί για να διατηρηθεί η συνθήκη αληθής.

Σε κάθε βήμα, ο SMO επιλέγει δύο στοιχεία a_i και a_j για να τα βελτιστοποιήσει από κοινού, βρίσκει τις βέλτιστες τιμές για αυτές τις δύο παραμέτρους δεδομένου ότι όλες οι άλλες είναι σταθερές, και ενημερώνει το διάνυσμα α, ανάλογα. Η επιλογή αυτών των δύο σημείων καθορίζεται από τους πειραματισμούς (heuristic), ενώ η βελτιστοποίηση των δύο πολλαπλασιαστών διεξάγεται αναλυτικά.

Παρόλη την ανάγκη περισσοτέρων επαναλήψεων για σύγκλιση, κάθε επανάληψη χρησιμοποιεί τόσο λίγες λειτουργίες που ο αλγόριθμος παρουσιάζει μια συνολική επιτάχυνση ορισμένων τάξεων μεγέθους. Εκτός του χρόνου σύγκλισης, άλλα σημαντικά χαρακτηριστικά του αλγορίθμου είναι ότι δε χρειάζεται να αποθηκεύσει στη μνήμη τον πίνακα πυρήνα, από τη στιγμή που δε συμπεριλαμβάνονται λειτουργίες πίνακα, ότι δε χρησιμοποιεί άλλα πακέτα και το ότι είναι αρκετά εύκολο να εφαρμοστεί. Να σημειωθεί ότι από τη στιγμή που ο τυπικός SMO δε χρησιμοποιεί έναν κρυμμένο στη μνήμη πίνακα πυρήνα, η εισαγωγή του θα μπορούσε να χρησιμοποιηθεί για να επιτευχθεί περαιτέρω επιτάχυνση, εις βάρος της αυξημένης πολυπλοκότητας χώρου.

ΠΑΡΑΡΤΗΜΑ Γ (ΕΦΑΡΜΟΓΕΣ)

Γ.1 GERMAN CREDIT DATA

Γ.1.1ΠΥΡΗΝΕΣ

ΓΡΑΜΜΙΚΟΣ/ΜΟΝΤΕΛΟ 1

```
Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  linear
       cost:  1
      gamma:  0.02040816

Number of Support Vectors:  410
```

ΠΟΛΥΩΝΥΜΙΚΟΣ/ΜΟΝΤΕΛΟ2

```
Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  polynomial
       cost:  1
     degree:  3
      gamma:  0.02040816
     coef.0:  0

Number of Support Vectors:  524
```

ΑΚΤΙΝΩΤΟΣ/ΜΟΝΤΕΛΟ 3

```
Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1
```

```
     gamma:  0.02040816

Number of Support Vectors:  481
```

ΣΙΓΜΟΕΙΔΗΣ/ΜΟΝΤΕΛΟ 4

```
Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  sigmoid
       cost:  1
     gamma:  0.02040816
    coef.0:  0

Number of Support Vectors:  458
```

Γ.1.2 ΣΥΝΟΛΟ ΕΚΠΑΙΔΕΥΣΗΣ

ΓΡΑΜΜΙΚΟΣ/ΜΟΝΤΕΛΟ 1

```
Confusion Matrix and Statistics

          Reference
Prediction   1   2
         1 475 113
         2  50 112

               Accuracy : 0.7827
                 95% CI : (0.7514, 0.8117)
    No Information Rate : 0.7
    P-Value [Acc > NIR] : 2.246e-07

                  Kappa : 0.4375
 Mcnemar's Test P-Value : 1.197e-06

            Sensitivity : 0.9048
            Specificity : 0.4978
         Pos Pred Value : 0.8078
         Neg Pred Value : 0.6914
             Prevalence : 0.7000
         Detection Rate : 0.6333
   Detection Prevalence : 0.7840
      Balanced Accuracy : 0.7013
```

```
'Positive' Class : 1
```

ΠΟΛΥΩΝΥΜΙΚΟΣ/ ΜΟΝΤΕΛΟ 2

```
Confusion Matrix and Statistics

          Reference
Prediction   1   2
         1 525 219
         2   0   6

               Accuracy : 0.708
                 95% CI : (0.674, 0.7403)
    No Information Rate : 0.7
    P-Value [Acc > NIR] : 0.3322

                  Kappa : 0.0369
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 1.00000
            Specificity : 0.02667
         Pos Pred Value : 0.70565
         Neg Pred Value : 1.00000
             Prevalence : 0.70000
         Detection Rate : 0.70000
   Detection Prevalence : 0.99200
      Balanced Accuracy : 0.51333

       'Positive' Class : 1
```

ΑΚΤΙΝΩΤΟΣ/ΜΟΝΤΕΛΟ 3

```
Confusion Matrix and Statistics

          Reference
Prediction   1   2
         1 517 170
         2   8  55

               Accuracy : 0.7627
                 95% CI : (0.7306, 0.7927)
    No Information Rate : 0.7
```

```
   P-Value [Acc > NIR] : 7.76e-05

                  Kappa : 0.2886
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9848
            Specificity : 0.2444
         Pos Pred Value : 0.7525
         Neg Pred Value : 0.8730
             Prevalence : 0.7000
         Detection Rate : 0.6893
   Detection Prevalence : 0.9160
      Balanced Accuracy : 0.6146

       'Positive' Class : 1
```

ΣΙΓΜΟΕΙΔΗΣ/ΜΟΝΤΕΛΟ 4

```
Confusion Matrix and Statistics

          Reference
Prediction   1   2
         1 522 213
         2   3  12

               Accuracy : 0.712
                 95% CI : (0.6781, 0.7442)
    No Information Rate : 0.7
    P-Value [Acc > NIR] : 0.25

                  Kappa : 0.0649
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.99429
            Specificity : 0.05333
         Pos Pred Value : 0.71020
         Neg Pred Value : 0.80000
             Prevalence : 0.70000
         Detection Rate : 0.69600
   Detection Prevalence : 0.98000
      Balanced Accuracy : 0.52381

       'Positive' Class : 1
```

Γ.1.3 ΣΥΝΟΛΟ ΔΟΚΙΜΗΣ

ΓΡΑΜΜΙΚΟΣ/ΜΟΝΤΕΛΟ 1

```
Confusion Matrix and Statistics

          Reference
Prediction   1   2
         1 158  41
         2  17  34

               Accuracy : 0.768
                 95% CI : (0.7107, 0.8189)
    No Information Rate : 0.7
    P-Value [Acc > NIR] : 0.010119

                  Kappa : 0.392
 Mcnemar's Test P-Value : 0.002527

            Sensitivity : 0.9029
            Specificity : 0.4533
         Pos Pred Value : 0.7940
         Neg Pred Value : 0.6667
             Prevalence : 0.7000
         Detection Rate : 0.6320
   Detection Prevalence : 0.7960
      Balanced Accuracy : 0.6781

       'Positive' Class : 1
```

ΠΟΛΥΩΝΥΜΙΚΟΣ/ ΜΟΝΤΕΛΟ 2

```
Confusion Matrix and Statistics

          Reference
Prediction   1   2
         1 174  75
         2   1   0

               Accuracy : 0.696
                 95% CI : (0.6349, 0.7524)
    No Information Rate : 0.7
    P-Value [Acc > NIR] : 0.5854
```

```
               Kappa : -0.008
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.9943
            Specificity : 0.0000
         Pos Pred Value : 0.6988
         Neg Pred Value : 0.0000
             Prevalence : 0.7000
         Detection Rate : 0.6960
   Detection Prevalence : 0.9960
      Balanced Accuracy : 0.4971

       'Positive' Class : 1
```

ΑΚΤΙΝΩΤΟΣ/ ΜΟΝΤΕΛΟ 3

```
Confusion Matrix and Statistics

          Reference
Prediction   1   2
         1 171  58
         2   4  17

               Accuracy : 0.752
                 95% CI : (0.6937, 0.8043)
    No Information Rate : 0.7
    P-Value [Acc > NIR] : 0.04044

               Kappa : 0.2566
 Mcnemar's Test P-Value : 1.685e-11

            Sensitivity : 0.9771
            Specificity : 0.2267
         Pos Pred Value : 0.7467
         Neg Pred Value : 0.8095
             Prevalence : 0.7000
         Detection Rate : 0.6840
   Detection Prevalence : 0.9160
      Balanced Accuracy : 0.6019

       'Positive' Class : 1
```

ΣΙΓΜΟΕΙΔΗΣ/ΜΟΝΤΕΛΟ 4

```
Confusion Matrix and Statistics

          Reference
Prediction   1   2
         1 175  70
         2   0   5

               Accuracy : 0.72
                 95% CI : (0.6599, 0.7747)
    No Information Rate : 0.7
    P-Value [Acc > NIR] : 0.2692

                  Kappa : 0.0909
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 1.00000
            Specificity : 0.06667
         Pos Pred Value : 0.71429
         Neg Pred Value : 1.00000
             Prevalence : 0.70000
         Detection Rate : 0.70000
   Detection Prevalence : 0.98000
      Balanced Accuracy : 0.53333

       'Positive' Class : 1
```

Γ.2 IRIS DATA SET

Γ.2.1 ΠΥΡΗΝΕΣ

ΓΡΑΜΜΙΚΟΣ/ΜΟΝΤΕΛΟ 1

```
Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  linear
       cost:  1
      gamma:  0.25

Number of Support Vectors:  21
```

ΠΟΛΥΩΝΥΜΙΚΟΣ/ΜΟΝΤΕΛΟ 2

```
Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  polynomial
       cost:  1
     degree:  3
      gamma:  0.25
     coef.0:  0

Number of Support Vectors:  40
```

ΑΚΤΙΝΩΤΟΣ/ΜΟΝΤΕΛΟ 3

```
Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1
      gamma:  0.25

Number of Support Vectors:  40
```

ΣΙΓΜΟΕΙΔΗΣ/ΜΟΝΤΕΛΟ 4

```
Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  sigmoid
       cost:  1
      gamma:  0.25
     coef.0:  0

Number of Support Vectors:  46
```

Γ.2.2 ΣΥΝΟΛΟ ΕΚΠΑΙΔΕΥΣΗΣ

ΓΡΑΜΜΙΚΟΣ/ΜΟΝΤΕΛΟ 1

Confusion Matrix and Statistics

	Reference		
Prediction	Iris-setosa	Iris-versicolor	Iris-virginica
Iris-setosa	39	0	0
Iris-versicolor	0	39	1

```
  Iris-virginica          0            1           33

Overall Statistics

               Accuracy : 0.9823
                 95% CI : (0.9375, 0.9978)
    No Information Rate : 0.354
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9734
 Mcnemar's Test P-Value : NA



Statistics by Class:

                     Class: Iris-setosa Class: Iris-versicolor
Sensitivity                     1.0000                 0.9750
Specificity                     1.0000                 0.9863
Pos Pred Value                  1.0000                 0.9750
Neg Pred Value                  1.0000                 0.9863
Prevalence                      0.3451                 0.3540
Detection Rate                  0.3451                 0.3451
Detection Prevalence            0.3451                 0.3540
Balanced Accuracy               1.0000                 0.9807
                     Class: Iris-virginica
Sensitivity                         0.9706
Specificity                         0.9873
Pos Pred Value                      0.9706
Neg Pred Value                      0.9873
Prevalence                          0.3009
Detection Rate                      0.2920
Detection Prevalence                0.3009
Balanced Accuracy                   0.9790
```

ΠΟΛΥΩΝΥΜΙΚΟΣ/ΜΟΝΤΕΛΟ 2

```
Confusion Matrix and Statistics

                 Reference
Prediction       Iris-setosa Iris-versicolor Iris-virginica
  Iris-setosa             39             0            0
  Iris-versicolor          0            40            6
```

```
  Iris-virginica          0            0           28

Overall Statistics

               Accuracy : 0.9469
                 95% CI : (0.888, 0.9803)
    No Information Rate : 0.354
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9198
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Iris-setosa Class: Iris-versicolor
Sensitivity                     1.0000                 1.0000
Specificity                     1.0000                 0.9178
Pos Pred Value                  1.0000                 0.8696
Neg Pred Value                  1.0000                 1.0000
Prevalence                      0.3451                 0.3540
Detection Rate                  0.3451                 0.3540
Detection Prevalence            0.3451                 0.4071
Balanced Accuracy               1.0000                 0.9589
                     Class: Iris-virginica
Sensitivity                         0.8235
Specificity                         1.0000
Pos Pred Value                      1.0000
Neg Pred Value                      0.9294
Prevalence                          0.3009
Detection Rate                      0.2478
Detection Prevalence                0.2478
Balanced Accuracy                   0.9118
```

ΑΚΤΙΝΩΤΟΣ/ ΜΟΝΤΕΛΟ 3

```
Confusion Matrix and Statistics

                 Reference
Prediction       Iris-setosa Iris-versicolor Iris-virginica
  Iris-setosa            39             0            0
  Iris-versicolor         0            38            1
  Iris-virginica          0             2           33
```

Overall Statistics

Accuracy : 0.9735
95% CI : (0.9244, 0.9945)
No Information Rate : 0.354
P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.9601
Mcnemar's Test P-Value : NA

Statistics by Class:

	Class: Iris-setosa	Class: Iris-versicolor	Class: Iris-virginica
Sensitivity	1.0000	0.9500	0.9706
Specificity	1.0000	0.9863	0.9747
Pos Pred Value	1.0000	0.9744	0.9429
Neg Pred Value	1.0000	0.9730	0.9872
Prevalence	0.3451	0.3540	0.3009
Detection Rate	0.3451	0.3363	0.2920
Detection Prevalence	0.3451	0.3451	0.3097
Balanced Accuracy	1.0000	0.9682	0.9726

ΣΙΓΜΟΕΙΔΗΣ/ΜΟΝΤΕΛΟ 4

Confusion Matrix and Statistics

Prediction	Reference: Iris-setosa	Iris-versicolor	Iris-virginica
Iris-setosa	39	0	0
Iris-versicolor	0	36	6
Iris-virginica	0	4	28

Overall Statistics

Accuracy : 0.9115

95% CI : (0.8433, 0.9567)
No Information Rate : 0.354
P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.8667
Mcnemar's Test P-Value : NA

Statistics by Class:

	Class: Iris-setosa	Class: Iris-versicolor	Class: Iris-virginica
Sensitivity	1.0000	0.9000	0.8235
Specificity	1.0000	0.9178	0.9494
Pos Pred Value	1.0000	0.8571	0.8750
Neg Pred Value	1.0000	0.9437	0.9259
Prevalence	0.3451	0.3540	0.3009
Detection Rate	0.3451	0.3186	0.2478
Detection Prevalence	0.3451	0.3717	0.2832
Balanced Accuracy	1.0000	0.9089	0.8864

Γ.2.3 ΣΥΝΟΛΟ ΔΟΚΙΜΗΣ

ΓΡΑΜΜΙΚΟΣ/ΜΟΝΤΕΛΟ 1

Confusion Matrix and Statistics

	Reference		
Prediction	Iris-setosa	Iris-versicolor	Iris-virginica
Iris-setosa	11	0	0
Iris-versicolor	0	9	0
Iris-virginica	0	1	16

Overall Statistics

Accuracy : 0.973

95% CI : (0.8584, 0.9993)
No Information Rate : 0.4324
P-Value [Acc > NIR] : 1.675e-12

Kappa : 0.9582
Mcnemar's Test P-Value : NA

Statistics by Class:

	Class: Iris-setosa	Class: Iris-versicolor	Class: Iris-virginica
Sensitivity	1.0000	0.9000	1.0000
Specificity	1.0000	1.0000	0.9524
Pos Pred Value	1.0000	1.0000	0.9412
Neg Pred Value	1.0000	0.9643	1.0000
Prevalence	0.2973	0.2703	0.4324
Detection Rate	0.2973	0.2432	0.4324
Detection Prevalence	0.2973	0.2432	0.4595
Balanced Accuracy	1.0000	0.9500	0.9762

ΠΟΛΥΩΝΥΜΙΚΟΣ/ΜΟΝΤΕΛΟ 2

Confusion Matrix and Statistics

	Reference		
Prediction	Iris-setosa	Iris-versicolor	Iris-virginica
Iris-setosa	11	0	0
Iris-versicolor	0	10	3
Iris-virginica	0	0	13

Overall Statistics

Accuracy : 0.9189
95% CI : (0.7809, 0.983)
No Information Rate : 0.4324
P-Value [Acc > NIR] : 6.343e-10

```
               Kappa : 0.878
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Iris-setosa Class: Iris-versicolor
Sensitivity                     1.0000             1.0000
Specificity                     1.0000             0.8889
Pos Pred Value                  1.0000             0.7692
Neg Pred Value                  1.0000             1.0000
Prevalence                      0.2973             0.2703
Detection Rate                  0.2973             0.2703
Detection Prevalence            0.2973             0.3514
Balanced Accuracy               1.0000             0.9444
                     Class: Iris-virginica
Sensitivity                     0.8125
Specificity                     1.0000
Pos Pred Value                  1.0000
Neg Pred Value                  0.8750
Prevalence                      0.4324
Detection Rate                  0.3514
Detection Prevalence            0.3514
Balanced Accuracy               0.9062
```

ΑΚΤΙΝΩΤΟΣ/ΜΟΝΤΕΛΟ 3

```
Confusion Matrix and Statistics

                 Reference
Prediction       Iris-setosa Iris-versicolor Iris-virginica
  Iris-setosa            11             0            0
  Iris-versicolor         0            10            1
  Iris-virginica          0             0           15

Overall Statistics

               Accuracy : 0.973
                 95% CI : (0.8584, 0.9993)
    No Information Rate : 0.4324
    P-Value [Acc > NIR] : 1.675e-12

               Kappa : 0.9588
 Mcnemar's Test P-Value : NA
```

Statistics by Class:

	Class: Iris-setosa	Class: Iris-versicolor
Sensitivity	1.0000	1.0000
Specificity	1.0000	0.9630
Pos Pred Value	1.0000	0.9091
Neg Pred Value	1.0000	1.0000
Prevalence	0.2973	0.2703
Detection Rate	0.2973	0.2703
Detection Prevalence	0.2973	0.2973
Balanced Accuracy	1.0000	0.9815

	Class: Iris-virginica
Sensitivity	0.9375
Specificity	1.0000
Pos Pred Value	1.0000
Neg Pred Value	0.9545
Prevalence	0.4324
Detection Rate	0.4054
Detection Prevalence	0.4054
Balanced Accuracy	0.9688

ΣΙΓΜΟΕΙΔΗΣ/ΜΟΝΤΕΛΟ 4

Confusion Matrix and Statistics

	Reference		
Prediction	Iris-setosa	Iris-versicolor	Iris-virginica
Iris-setosa	11	0	0
Iris-versicolor	0	6	2
Iris-virginica	0	4	14

Overall Statistics

Accuracy : 0.8378
95% CI : (0.6799, 0.9381)
No Information Rate : 0.4324
P-Value [Acc > NIR] : 4.664e-07

Kappa : 0.7477
Mcnemar's Test P-Value : NA

Statistics by Class:

	Class: Iris-setosa	Class: Iris-versicolor
Sensitivity	1.0000	0.6000
Specificity	1.0000	0.9259
Pos Pred Value	1.0000	0.7500
Neg Pred Value	1.0000	0.8621
Prevalence	0.2973	0.2703
Detection Rate	0.2973	0.1622
Detection Prevalence	0.2973	0.2162
Balanced Accuracy	1.0000	0.7630

	Class: Iris-virginica
Sensitivity	0.8750
Specificity	0.8095
Pos Pred Value	0.7778
Neg Pred Value	0.8947
Prevalence	0.4324
Detection Rate	0.3784
Detection Prevalence	0.4865
Balanced Accuracy	0.8423