ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ
ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

Μεταπτυχιακή Διπλωματική Εργασία

Μέθοδοι για την Ταξινόμηση μη Ισορροπημένων Δεδομένων με Μηχανές Διανυσματικής Υποστήριξης

ΔΡΟΣΟΥ Π. ΚΡΥΣΤΑΛΛΕΝΙΑ

Επιβλέπων Καθηγητής:

Κουκουβίνος Χρήστος, Καθηγητής Ε.Μ.Π

Αθήνα, 2015

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΔΙΕΠΙΣΤΗΜΟΝΙΚΟ – ΔΙΑΤΜΗΜΑΤΙΚΟ
ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
(Δ.Π.Μ.Σ.) "ΕΦΑΡΜΟΣΜΕΝΕΣ ΜΑΘΗΜΑΤΙΚΕΣ ΕΠΙΣΤΗΜΕΣ"

Μέθοδοι για την Ταξινόμηση μη Ισορροπημένων Δεδομένων με Μηχανές Διανυσματικής Υποστήριξης

Δρόσου Π. Κρυσταλλένια

Μεταπτυχιακή Διπλωματική Εργασία
η οποία υποβάλλεται για μερική εκπλήρωση των απαιτήσεων
στο Δ.Π.Μ.Σ. "Εφαρμοσμένες Μαθηματικές Επιστήμες".

Επιβλέπων καθηγητής

Κουκουβίνος Χρήστος Καθηγητής ΕΜΠ
Σχολή Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών

Αθήνα, 2015

NATIONAL TECHNICAL UNIVERSITY OF ATHENS

School of Applied Mathematics and Physical Science

MASTER PROGRAM IN APPLIED MATHEMATICS

Class Imbalanced Problem with Support Vector Machines

BY

Drosou P. Krystallenia

Master Thesis submitted to the Department of Mathematics of the National technical University of Athens in partial fulfillment of the requirements for the degree of Master of Science in Applied Mathematics

Athens, Greece, 2015

Ευχαριστίες

Στο σημείο αυτό θα ήθελα να ευχαριστήσω αρκετούς ανθρώπους που με βοήθησαν στην διεξαγωγή του μεταπτυχιακού αλλά και στην συγγραφή της παρούσας εργασίας. Η δημιουργία και παρουσίαση αυτής της εργασίας θα ήταν αδύνατη χωρίς την υποστήριξη και την υπομονή τους.

Πρώτα απ' όλους θα ήθελα να ευχαριστήσω εκ βαθέων και να εκφράσω την βαθύτατη εκτίμησή μου στον επιβλέποντά μου, καθηγητή του Εθνικού Μετσόβιου Πολυτεχνείου, κ. Χρήστο Κουκουβίνο. Η ενθάρρυνση και η πολύτιμη βοήθεια του καθ' όλη την διάρκεια των σπουδών μου, αλλά και η συνεχής καθοδήγησή του μου επέτρεψαν να ολοκληρώσω με επιτυχία τη συγκεκριμένη εργασία και τις μεταπτυχιακές σπουδές μου.

Παράλληλα ιδιαίτερες ευχαριστίες θα ήθελα να εκφράσω στα ιδρύματα από τα οποία έλαβα υποτροφίες που μου επέτρεψαν την επιτυχή παρακολούθηση και ολοκλήρωση των μεταπτυχιακών σπουδών μου. Αρχικά εκ βαθέων ευχαριστίες οφείλω στο Ιδρυμα Λοχαγού Φανουράκη, τον αείμνηστο ιδρυτή κ. Σωκράτη Φανουράκη και το διοικητικό συμβούλιο για την τιμή και την εμπιστοσύνη στο πρόσωπό μου. Επίσης ευχαριστώ ιδιαίτερα το Ιδρυμα Κρατικών Υποτροφιών για την χρηματοδότηση των συγκεκριμένων σπουδών. Η ολοκλήρωση της διπλωματικής εργασίας συγχρηματοδοτήθηκε μέσω της Πράξης Πρόγραμμα χορήγησης υποτροφιών ΙΚΥ για Μεταπτυχιακές Σπουδές Πρώτου Κύκλου (Μάστερ) - Οριζόντια Πράξη, από πόρους του ΕΠ «Εκπαίδευση και Δια Βίου Μάθηση» του Ευρωπαϊκού Κοινωνικού Ταμείου (ΕΚΤ) του ΕΣΠΑ, 2007-2013.

Θα ήθελα επίσης να εκφράσω την ευγνωμοσύνη μου στα μέλη της οικογένειας μου: στους γονείς μου, Παύλο και Κατερίνα και στον αδερφό μου, Νίκο για την διαρκή υποστήριξη, την ενθάρυνση και την υπομονή τους, που μου επέτρεψαν την επιτυχή και ομαλή διεκπεραίωση των σπουδών μου.

Δρόσου Κρυσταλλένια

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

Πρόλογος

Στις μέρες μας είναι σύνηθες το φαινόμενο της αναζήτηση προτύπων και μοντέλων από ιδιαίτερα πολύπλοκες δομές. Για παράδειγμα σε τομείς όπως η τεχνολογία της πληροφορίας (information technology) και η βιοπληροφορική (bioinformatics) ερχόμαστε αντιμέτωποι με δεδομένα υψηλής διάστασης όπου η κλασική στατιστική συμπερασματολογία αποτυγχάνει κάνοντας επιτακτική την ανάγκη για υιοθέτηση νέων, καινοτόμων μεθόδων. Παράλληλα, ένα άλλο ζήτημα αποτελεί το γεγονός ότι πολλές φυσικές διεργασίες συχνά δημιουργούν προβλήματα στα οποία κάποιες παρατηρήσεις συμβαίνουν με μεγαλύτερη συχνότητα από κάποιες άλλες. Αυτές οι διεργασίες οδηγούν σε μη ισορροπημένη κατανομή μεταξύ των κλάσεων του προς μελέτη προβλήματος δημιουργώντας και πάλι προβλήματα στην κλασική στατιστική μοντελοποίηση. Ιδιαίτερο ενδιαφέρον αποτελεί το εν λόγω πρόβλημα, που σχετίζεται με την ανισορροπία μεταξύ των κλάσεων, σε προβλήματα ταξινόμησης και συγκεκριμένα στη δυαδική ταξινόμηση. Αυτές οι διεργασίες που οδηγούν σε διαφορετική κατανομή μεταξύ των κλάσεων επηρεάζουν τον ταξινομητή κάνοντας τον μεροληπτικό υπέρ της κλάσης των δεδομένων που αποτελεί την πλειοψηφία στο πρόβλημα που μελετούμε και αυτό διότι οι κλασικές μέθοδοι ταξινόμησης, δηλαδή οι συμβατικοί ταξινομητές υποθέτουν ότι υπάρχει μία κανονική κατανομή μεταξύ των δύο κλάσεων του προβλήματος. Η ποσότητα αλλά και η ποικιλία των πεδίων που σχετίζονται με το πρόβλημα της ανισορροπίας μεταξύ των κλάσεων είχε ως αποτέλεσμα την παρακίνηση της ερευνητικής κοινότητας να αντιμετωπίσει αυτό το μείζον ζήτημα προσελκύοντας όλο και περισσότερους επιστήμονες να ερευνήσουν τα προβλήματα των μη ισορροπημένων κατανομών σε προβλήματα ταξινόμησης.

Περιεχόμενα

Περίληψη ..... **xxi**

Abstract ..... **xxiii**

Πίνακας Περιεχομένων ..... **xi**

Κατάλογος Πινάκων ..... **xv**

Κατάλογος Σχημάτων ..... **xvii**

ΚΕΦΑΛΑΙΟ 1 .....

Το Πρόβλημα της Ανισορροπίας Μεταξύ των Κλάσεων (The problem of handling the class imbalance problem) **25**

1.1 Εισαγωγή 25

1.2 Ταξινόμηση των δεδομένων 26

1.3 Ορισμός του προβλήματος της ανισορροπίας των κλάσεων 27

1.4 Χαρακτηριστικά των μη ισορροπημένων δεδομένων 31

1.5 Χειρισμός του προβλήματος των μη ισορροπημένων κλάσεων 33

ΚΕΦΑΛΑΙΟ 2 .....

Μέθοδοι Χειρισμού Μη Ισορροπημένων Δεδομένων (Techniques for imbalanced data set problems) **35**

2.1 Εισαγωγικά στοιχεία 35

2.2 Προσεγγίσεις σε επίπεδο δεδομένων (data level approaches) 37

2.2.1 Εισαγωγή-Αλλαγή της κατανομής των κλάσεων 37

2.2.2 Μέθοδοι 38

2.2.3 Αλγόριθμοι 40

2.3 Προσέγγιση σε επίπεδο ταξινομητών (Classifier level approaches) 45

2.3.1 Μέθοδοι 46

2.3.2 Αλγόριθμοι 50
2.4 Cost sensitive approaches 50
2.4.1 Μέθοδοι 51
2.4.2 Αλγόριθμοι 52
2.5 Feature selection approaches 52
2.5.1 Μέθοδοι 53
2.5.2 Αλγόριθμοι 53
2.6 Προσέγγιση Συνόλου μεθόδων (Ensemble level approaches) 54
2.6.1 Μέθοδοι 55
2.6.2 Αλγόριθμοι 56
2.7 Γενικά βήματα για την επίλυση προβλήματων μη ισορροπημένων δεδομένων 59

ΚΕΦΑΛΑΙΟ 3 .....

Το πρόβλημα της ανισσοροπίας των κλάσεων με Μηχανές Διανυσματικής Υποστήριξης (class imbalance problem with support vector machine learning) 61

3.1 Εισαγωγή 61
3.2 Εισαγωγή στις Μηχανές Διανυσματικής Υποστήριξης 63
3.2.1 Η SVM μέθοδος για την δυαδική ταξινόμηση 63
3.2.2 Η SVM μέθοδος για την παλινδρόμηση 71
3.2.3 Το SVM ως ποινικοποιημένη μέθοδος 73
3.2.4 Πυρήνες 76
3.2.5 Επιλογή παραμέτρων για τις SVM 78
3.3 Προσεγγιστικές Μηχανές Διανυσματικής Υποστήριξης (PSVM) 79
3.3.1 Γραμμικό PSVM 79
3.3.2 Μη γραμμικό Proximal Support vector machine (NPSVM) 80
3.4 SVM και μη ισορροπημένα δεδομένα 81
3.4.1 Αδυναμία του προβλήματος βελτιστοποίησης του μαλακού περιθωρίου (Weakness of the soft margin optimization problem 81
3.4.2 Το μη ισορροπημένο ποσοστό των διανυσμάτων υποστήριξης (The imbalanced support-vector ratio) 82
3.4.3 Αποτελεσματικότητα της εξισορρόπησης των κλάσεων 82
3.4.4 Υποθέσεις 86

3.5 Εξωτερική μάθηση μη ισορροπημένων δεδομένων για τις SVMs: Μέθοδοι προεπεξεργασίας των δεδομένων 87
3.5.1 Resampling methods 87
3.5.2 Ensemble learning methods 88
3.6 Εσωτερική μάθηση μη ισορροπημένων δεδομένων για τις SVMs: Αλγοριθμικές Μέθοδοι 89
3.6.1 Different Error Costs (DEC) Cost sensitivity SVM (TCSVM) for imbalanced data 89
3.6.2 One class learning 90
3.6.3 z-SVM 91
3.6.4 Modified proximal SVM 92
3.6.5 Μέθοδοι Τροποποίησης του πυρήνα (Kernel modification methods) 93

ΚΕΦΑΛΑΙΟ 4 .....
Μέτρα Απόδοσης σε Μη Ισορροπημένα Πεδία (Evaluation Criteria in Imbalanced Domains) 95
4.1 Εισαγωγή 95
4.2 Μεροληψία, Διασπορά και περιπλοκότητα μοντέλου 95
4.3 Διασταυρωμένη επικύρωση (Cross-validation) 99
4.4 Πίνακας Συνάφειας 100
4.5 Καμπύλες ROC 104
4.5.1. Γραφήματα και Ερμηνεία 105
4.5.2. Η περιοχή κάτω από την ROC καμπύλη (AUC) 106
4.6 Precision and Recall 106
4.7 Γεωμετρικός Μέσος (GMean) 108
4.8 Μέτρα Ευαίσθητου Κόστους (Cost sensitive measures) 108
4.8.1. Cost curve 108
4.8.2. Cost matrix 110

ΚΕΦΑΛΑΙΟ 5 .....
Πειραματικά Αποτελέσματα (Experimental Results) 113
5.1 Εφαρμογή σε μικρά σύνολα δεδομένων 114

5.1.1. Γραμμική περίπτωση 118
5.1.2. Μη γραμμική περίπτωση (Non-linear Case) 120
5.1.3 Imbalanced Methods (Two cost/weight SVM and Modified Proximal SVM) 125
5.1.4 Comparisons 131
5.2 Εφαρμογή σε πραγματικά ιατρικά δεδομένα υψηλής διάστασης 137
5.2.1 Περιγραφή των δεδομένων 138
5.2.2 Κλασικό SVM (Standard SVM) 138
5.2.3 Προσεγγίσεις για μη ισορροπημένα δεδομένα 141
5.2.4 Πειραματικά αποτελέσματα και συγκρίσεις 152

ΚΕΦΑΛΑΙΟ 6 .....
ΣΥΜΠΕΡΑΣΜΑΤΑ 155
6.1 Συμπεράσματα πρώτης μελέτης 156
6.2 Συμπεράσματα δεύτερης μελέτης 158
ΒΙΒΛΙΟΓΡΑΦΙΑ 161
ΠΑΡΑΡΤΗΜΑ Α 173
ΠΑΡΑΡΤΗΜΑ Β 177
Pima Indians Diabetes Data Set (UCI Repository) 178
Blood Transfusion (UCI Repository) 179
Thyroid Disease (New Thyroid) data set 180

Κατάλογος Πινάκων

Πίνακας 2.1 Πίνακας κόστους..... 46

Πίνακας 3.1: Συναρτήσεις ελαχιστοποίησης για τις διάφορες συναρτήσεις απώλειας .. 75

Πίνακας 4.1 Πίνακας Συνάφειας..... 101
Πίνακας 4.2 Συγκεντρωτική εικόνα των μέτρων απόδοσης..... 103

Πίνακας 5.1: Περιγραφή των τεσσάρων συνόλων δεδομένων 118
Πίνακας 5.2: PSVM and SVM training and testing correctness and running times using a linear classifier. Execution times include ten-fold training. Best results are in bold. 119
Πίνακας 5.3: Ορθότητα ταξινόμησης του PSVM and του SVM στο σύνολο εκπαίδευσης και στο σύνολο δοκιμής χρησιμοποιώντας 10-fold cross validation καθώς επίσης και οι χρόνοι εκτέλεσης χρησιμοποιώντας μη γραμμικούς ταξινομητές. Οι χρόνοι περιλαμβάνουν το 10-fold training. Η τιμή του g που έδωσε την καλύτερη επίδοση χρησιμοποιήθηκε σε κάθε περίπτωση. Τα καλύτερα αποτελέσματα είναι με έντονα γράμματα. 124
Πίνακας 5.4: Απόδοση του PSVM με πολυωνυμικό πυρήνα 125
Πίνακας 5.5: Ορθότητα ταξινόμησης στο σύνολο εκπαίδευσης και δοκιμής και χρόνοι εκτέλεσης χρησιμοποιώντας μεθόδους για μη ισορροπημένα δεδομένα, γραμμικό TCSVM, μη γραμμικό TCSVM και MPSVM. Τα καλύτερα αποτελέσματα φαίνονται με έντονα γράμματα. 130
Πίνακας 5.6: Ορθότητα ταξινόμησης του PSVM και του SVM στο σύνολο εκπαίδευσης και στο σύνολο δοκιμής χρησιμοποιώντας 10-fold cross validation καθώς επίσης και οι χρόνοι εκτέλεσης χρησιμοποιώντας γραμμικό πυρήνα. Τα καλύτερα αποτελέσματα είναι με έντονα γράμματα. 131
Πίνακας 5.7: Ορθότητα ταξινόμησης του PSVM και του SVM στο σύνολο εκπαίδευσης και στο σύνολο δοκιμής χρησιμοποιώντας 10-fold cross validation καθώς επίσης και οι χρόνοι εκτέλεσης χρησιμοποιώντας μη γραμμικό πυρήνα. Τα καλύτερα αποτελέσματα είναι με έντονα γράμματα..... 132
Πίνακας 5.8: Ορθότητα ταξινόμησης του TCSVM, του μη γραμμικού TCSVM και του MPSVM στο σύνολο εκπαίδευσης και στο σύνολο δοκιμής χρησιμοποιώντας 10-fold cross validation καθώς επίσης και οι χρόνοι εκτέλεσης. Τα καλύτερα αποτελέσματα είναι με έντονα γράμματα. 134
Πίνακας 5.9: Ορθότητα ταξινόμησης του TCSVM, του μη γραμμικού TCSVM και του MPSVM στο σύνολο εκπαίδευσης και στο σύνολο δοκιμής θεωρώντας το μέτρο του Γεωμετρικού Μέσου χρησιμοποιώντας 10-fold cross validation. Περιλαμβάνονται και οι χρόνοι εκτέλεσης. Τα καλύτερα αποτελέσματα είναι με έντονα γράμματα. 134

Επιπλέον, όπως προκύπτει από τους πίνακες που παρουσιάστηκαν παραπάνω (Πίνακας 5.6, 5.7, 5.8, 5.10) το SVM και το PSVM συγκέντρωσαν υψηλότερες τιμές ακρίβειας στις περισσότερες περιπτώσεις, σε σύγκριση με τις μεθόδους μάθησης ευαίσθητου κόστους. Ωστόσο, από την άποψη του γεωμετρικού μέσου (GM), το TCSVM και το MPSVM έχουν σαφώς καλύτερες επιδόσεις σε σχέση με τις συμβατικές μεθόδους οι οποίες δεν λαμβάνουν υπόψη τη μη ισορροπημένη αναλογία μεταξύ των δύο κλάσεων. 135
Πίνακας 5.11: Περιγραφή του ιατρικού συνόλου δεδομένων..... 138
Πίνακας 5.12: επιλογή παραμέτρου στο SVM με Γκαουσιανό πυρήνα (RBF). 139
Πίνακας 5.13: Σύγκριση της επίδοσης του κλασικού SVM για διαφορετικούς πυρήνες, στο ιατρικό σύνολο δεδομένων 141
Πίνακας 5.14: Συγκρίσεις επίδοσης για το C και το TC SVM (γραμμική περίπτωση) 144
Πίνακας 5.15: Συγκρίσεις επίδοσης σε πιο εύρωστα μέτρα για το C και το TC SVM (γραμμική περίπτωση)..... 145
Πίνακας 5.16: Κριτήρια επίδοσης για τις δύο διαφορετικές SVM τεχνικές (μη γραμμική περίπτωση)..... 146
Πίνακας 5.17: Κριτήρια επίδοσης για τις δύο διαφορετικές SVM τεχνικές (μη γραμμική περίπτωση)..... 146
Πίνακας 5.18: Grid search για διαφορετικούς συνδυασμούς του SMOTE SVM και της τυχαίας υποδειγματοληψίας. 150
Πίνακας 5.19: Σύγκριση του ποσοστού επιτυχίας στην κλάση μειωψηφίας για διαφορετικά ποσοστά υποδειγματοληψίας αλλάζοντας το ποσοστό υπερδειγματοληψίας 151
Πίνακας 5.20: Γεωμετρικός Μέσος του συνόλου εκπαίδευσης που αποκτήθηκαν από τις 4 διαφορετικές μεθόδους..... 152
Πίνακας 5.21: Γεωμετρικός Μέσος του συνόλου εκπαίδευσης που αποκτήθηκαν από τις 4 διαφορετικές μεθόδους (μη-γραμμική περίπτωση) 153

Κατάλογος Σχημάτων

Σχήμα 1.1: Παραδείγματα Πεδίων Μη Ισορροπημένων Δεδομένων..... 29
Σχήμα 1.2: Χαρακτηριστικά του μη ισορροπημένου συνόλου δεδομένων..... 31

Σχήμα 2.1: Ταξινόμηση των Τεχνικών για την επίλυση των μη ισορροπημένων δεδομένων..... 36
Σχήμα 2.2: Συνθετικά παραδείγματα με τον αλγόριθμο υπερδειγματοληψίας, SMOTE43
Σχήμα 2.3: Σύνολο μεθόδων για τα μη ισορροπημένα δεδομένα..... 55
Σχήμα 2.4: Βήματα για την επίλυση προβλήματων μη ισορροπημένων δεδομένων..... 59

Σχήμα 3.1: Υπερεπίπεδο ανάμεσα σε δύο γραμμικά διαχωρισμένες κλάσεις 65
Σχήμα 3.2: Για C=10000, το 62% των παρατηρήσεων είναι σημεία υποστήριξης..... 67
Σχήμα 3.3: Για C=0.01 το 85% των παρατηρήσεων είναι σημεία υποστήριξης..... 68
Σχήμα 3.4 : Υπερεπίπεδο διαμέσου δύο μη γραμμικά διαχωρίσιμων κλάσεων..... 68
Σχήμα 3.5: SVM με πολυωνυμικό πυρήνα 70
Σχήμα 3.6: SVM με γκαουσιανό (radial basis) πυρήνα 71
Σχήμα 3.7 : Παλινδρόμηση με ε- insensitive σωλήνα (tube) 72
Σχήμα 3.8: Συναρτήσεις απώλειας 74
Σχήμα 3.9: Διχοτόμηση δεδομένων, ανασχηματισμός με τη χρήση του πυρήνα RBF… 77
Σχήμα 3.10: Παράδειγμα του προβλήματος ανισορροπίας στις SVM..... 83
Σχήμα 3.11: Κινήσεις του ορίου απόφασης από τον αλγόριθμο SMOTE 84
Σχήμα 3.12: Κινήσεις του ορίου απόφασης με τυχαία υποδειγματοληψία..... 86

Σχήμα 4.1: Συμπεριφορά του σφάλματος του συνόλου δοκιμών και του συνόλου εκπαίδευσης καθώς ποικίλει η περιπλοκότητα του μοντέλου. Οι ανοιχτόχρωμες μπλε καμπύλες δείχνουν το σφάλμα της εκπαίδευσης err, ενώ οι ανοιχτόχρωμες κόκκινες καμπύλες δείχνουν το υποθετικό σφάλμα δοκιμών Err_T για 100 σετ εκπαίδευσης μεγέθους 50 το καθένα, καθώς η πολυπλοκότητα του μοντέλου μεγαλώνει. Οι συμπαγείς καμπύλες δείχνουν το αναμενόμενο σφάλμα δοκιμών Err και το αναμενόμενο σφάλμα εκπαίδευσης E(err). 96
Σχήμα 4.2: Υποθετική καμπύλη μάθησης για έναν ταξινομητή σε ένα συγκεκριμένο έργο: ένα γράφημα 1 - Err σε σχέση με το μέγεθος του συνόλου εκπαίδευσης N. Με ένα σύνολο δεδομένων από 200 παρατηρήσεις, μία 5-fold διασταυρωμένη επικύρωση θα χρησιμοποιεί σύνολα εκπαίδευσης μεγέθους 160, τα οποία θα συμπεριφέρονται σαν το πλήρες σύνολο. Ωστόσο, με ένα σύνολο δεδομένων των 50 παρατηρήσεων η 5-fold διασταυρωμένη επικύρωση θα χρησιμοποιεί σύνολα εκπαίδευσης μεγέθους 40, και αυτό θα είχε ως αποτέλεσμα μία σημαντική υπερεκτίμηση του σφάλματος πρόβλεψης. 100
Σχήμα 4.3 Μέτρα απόδοσης για τα μη ισορροπημένα δεδομένα 102

Σχήμα 4.4 Ο χώρος της ROC καμπύλης και το γράφημα 4 παραδειγμάτων πρόβλεψης 105
Σχήμα 4.5: (a) Δύο καμπύλες ROC που διασταυρώνονται – (b) Αντίστοιχες καμπύλες κόστους..... 109
Σχήμα 4.6 Δύο καμπύλες ROC – (b) Αντίστοιχες καμπύλες κόστους..... 109

Σχήμα 5.1: Κατανομή των κλάσεων στο Blood Transfusion σύνολο δεδομένων (οι κόκκινες κουκίδες αναφέρονται στην κλάση -1(πλειοψηφία), οι μπλε κουκίδες στην κλάση 1(μειοψηφία))..... 114
Σχήμα 5.2: Κατανομή των κλάσεων στο Pima Indians Diabetes σύνολο δεδομένων (οι κόκκινες κουκίδες ανφέρονται στην κλάση -1(πλειοψηφία), οι μπλε κουκίδες στην κλάση 1(μειοψηφία))..... 115
Σχήμα 5.3: Κατανομή των κλάσεων στο Thyroid σύνολο δεδομένων (οι κόκκινες κουκίδες ανφέρονται στην κλάση -1(πλειοψηφία), οι μπλε κουκίδες στην κλάση 1(μειοψηφία))..... 116
Σχήμα 5.4: Κατανομή των κλάσεων στο Real Medical σύνολο δεδομένων (οι κόκκινες κουκίδες ανφέρονται στην κλάση -1(πλειοψηφία), οι μπλε κουκίδες στην κλάση 1(μειοψηφία))..... 117
Σχήμα 5.5: Μέτρα για διαφορετικές τιμές της παραμέτρου του κόστους (το κόκκινο είναι για τον GM, το μαύρο για την ακρίβεια, το μπλε για την ειδικότητα και το πράσινο για την ευαισθησία) χρησιμοποιώντας το γραμμικό SVM. Στον παραπάνω πίνακα παρουσιάζονται τα αποτελέσματα στο σύνολο δοκιμών. 119
Σχήμα 5.6: Αναζήτηση πλέγματος (Grid search) για την παράμετρο γάμμα. Το αριστερό πάνελ αναφέρεται στην ορθότητα του συνόλου δοκιμής του μη γραμμικού SVM με Gaussian πυρήνα και ο δεξιός πίνακας αναφέρεται στο PSVM με ένα γκαουσιανό πυρήνα στο Blood transfusion σύνολο δεδομένων. Ο πράσινος συμπαγής κύκλος αντιπροσωπεύει την καλύτερη τιμή της g για το SVM (gamma=2) και για το PSVM (gamma=0.5). 121
Σχήμα 5.7: Αναζήτηση πλέγματος (Grid search) για την παράμετρο γάμμα για τον SVM ταξινομητή. Παρουσιάζεται η ορθότητα του συνόλου δοκιμής του μη γραμμικού SVM με Gaussian πυρήνα στο Pima Indians Diabetes σύνολο δεδομένων (gamma=0.03125). Ο πράσινος συμπαγής κύκλος αντιπροσωπεύει την καλύτερη τιμή της g για το SVM..... 122
Σχήμα 5.8: Αναζήτηση πλέγματος (Grid search) για την παράμετρο γάμμα για τον SVM ταξινομητή. Παρουσιάζεται η ορθότητα του συνόλου δοκιμής του μη γραμμικού SVM με Gaussian πυρήνα (gamma=0.25000) στο Thyroid σύνολο δεδομένων. Ο πράσινος συμπαγής κύκλος αντιπροσωπεύει την καλύτερη τιμή της g για το SVM..... 122
Σχήμα 5.9: Αναζήτηση πλέγματος (Grid search) για την παράμετρο γάμμα. Το αριστερό πάνελ αναφέρεται στην ορθότητα του συνόλου δοκιμής του μη γραμμικού SVM με Gaussian πυρήνα και ο δεξιός πίνακας αναφέρεται στο PSVM με ένα γκαουσιανό πυρήνα στο ιατρικό σύνολο δεδομένων. Ο πράσινος συμπαγής κύκλος αντιπροσωπεύει την καλύτερη τιμή της g για το SVM (gamma= 0.0625) και για το PSVM (gamma=0.0714). 123

Σχήμα 5.10: Αναζήτηση πλέγματος (Grid search) για τις παραμέτρους των μεθόδων για μη ισορροπημένα δεδομένα χρησιμοποιώντας 10-fold cross validation. Το αριστερό πάνελ αναφέρεται στην ορθότητα του συνόλου δοκιμής θεωρώντας το μέτρο του Γεωμετρικού Μέσου (Geometric mean) δείχνοντας την απόδοση του TCSVM με γραμμικό πυρήνα. Το δεξί πάνελ αναφέρεται στο MPSVM με ένα γραμμικό πυρήνα θεωρώντας ως μέτρο τον Γεωμετρικό Μέσο. Οι μέθοδοι εφαρμόστηκαν στο Blood Transfusion σύνολο δεδομένων. Ο πράσινος συμπαγής κύκλος αντιπροσωπεύει το καλύτερο ποσοστό για τις παραμέτρους και των δυο μεθόδων. 126
Σχήμα 5.11: Αναζήτηση πλέγματος (Grid search) για τις παραμέτρους των μεθόδων για μη ισορροπημένα δεδομένα χρησιμοποιώντας 10-fold cross validation. Το αριστερό πάνελ αναφέρεται στην ορθότητα του συνόλου δοκιμής θεωρώντας το μέτρο του Γεωμετρικού Μέσου (Geometric mean) δείχνοντας την απόδοση του TCSVM με γραμμικό πυρήνα. Το δεξί πάνελ αναφέρεται στο MPSVM με ένα γραμμικό πυρήνα θεωρώντας ως μέτρο τον Γεωμετρικό Μέσο. Οι μέθοδοι εφαρμόστηκαν στο Pima Indians Diabetes σύνολο δεδομένων. Ο πράσινος συμπαγής κύκλος αντιπροσωπεύει το καλύτερο ποσοστό για τις παραμέτρους και των δυο μεθόδων. 127
Σχήμα 5.12: Αναζήτηση πλέγματος (Grid search) για τις παραμέτρους χρησιμοποιώντας 10-fold cross validation. Το γράφημα αναφέρεται στην ορθότητα στο σύνολο δοκιμής θεωρώντας το μέτρο του Γεωμετρικού Μέσου (Geometric mean) δείχνοντας την απόδοση του TCSVM με γραμμικό πυρήνα στο σύνολο δεδομένων Thyroid. Ο πράσινος συμπαγής κύκλος αντιπροσωπεύει το καλύτερο ποσοστό για την παράμετρο της μεθόδου TCSVM. 128
Σχήμα 5.13: Αναζήτηση πλέγματος (Grid search) για τις παραμέτρους των μεθόδων για μη ισορροπημένα δεδομένα χρησιμοποιώντας 10-fold cross validation. Το αριστερό πάνελ αναφέρεται στην ορθότητα του συνόλου δοκιμής θεωρώντας το μέτρο του Γεωμετρικού Μέσου (Geometric mean) δείχνοντας την απόδοση του TCSVM με γραμμικό πυρήνα. Το δεξί πάνελ αναφέρεται στο MPSVM με ένα γραμμικό πυρήνα θεωρώντας ως μέτρο τον Γεωμετρικό Μέσο. Οι μέθοδοι εφαρμόστηκαν στο Real Medical σύνολο δεδομένων. Ο πράσινος συμπαγής κύκλος αντιπροσωπεύει το καλύτερο ποσοστό για τις παραμέτρους και των δυο μεθόδων. 129
Σχήμα 5.14: Απόδοση του γραμμικού SVM για διαφορετικές τιμές του κόστους βάσει του ποσοστού σφάλματος. 139
Σχήμα 5.15: Επίδοση του SVM με κανονικό (RBF) πυρήνα για διαφορετικές τιμές της παραμέτρου gamma 140
Σχήμα 5.16: Γεωμετρικός μέσος (y-άξονας) αλλάζοντας το κόστος της κλάσης πλειοψηφίας (x-άξονας) 142
Σχήμα 5.17: Επίδοση βάσει των τριών μέτρων (Accuracy, Sensitivity, Specificity) αλλάζοντας το κόστος της κλάσης πλειοψηφίας (x-axis) (συμπαγής κόκκινη γραμμή: TCSVM; διακεκκομένη γραμμή: Classic SVM) 143
Σχήμα 5.18: Επίδοση ως προς την ακρίβεια, την ευαισθησία και την ειδικότητα της majority κλάσης (x-άξονας) 143
Σχήμα 5.19: Τρια διαφορετικά μέτρα (y-άξονας) σε σχέση με τα διανύσματα υποστήριξης (x-άξονας) για το TCSVM. 144
Σχήμα 5.20: Σύγκριση των Roc καμπύλων για τη γραμμική περίπτωση 145

Σχήμα 5.21: Σύγκριση των καμπύλων ROC για τη μη γραμμική περίπτωση στο σύνολο ελέγχου. 147
Σχήμα 5.22: ROC καμπύλες για όλους τους πυρήνες για τις μεθόδους C και TC..... 148
Σχήμα 5.23: Τιμές του γεωμετρικού μέσου στο σύνολο εκπαίδευσης σε σχέση με την αύξηση των συνθετικών παραδειγμάτων στην κλάση μειωψηφίας από την μέθοδο SMOTE..... 149
Σχήμα 5.24: Γραφική απεικόνηση του ποσοστού επιτυχίας στην κλάση μειωψηφίας για διαφορετικά ποσοστά υποδειγματοληψίας αλλάζοντας το ποσοστό υπερδειγματοληψίας. 151
Σχήμα 5.25: Μέγεθος των δειγμάτων εκπαίδευσης 153
Σχήμα 5.26 Σύγκριση των τιμών του Γεωμετρικού Μέσου για τα σύνολα εκπαίδευσης, για τις 5 διαφορετικές μεθόδους. Έγινε χρήση του Γκαουσιανού πυρήνα στο ιατρικό σύνολο δεδομένων. 154

Μέθοδοι για την Ταξινόμηση μη Ισορροπημένων Δεδομένων με Μηχανές Διανυσματικής Υποστήριξης

Λέξεις κλειδιά: *Μηχανές διανυσματικής υποστήριξης, πρόβλημα ανισορροπίας των κλάσεων, μέθοδοι εξισορρόπησης, ιατρικά δεδομένα*

Περίληψη

Στην παρούσα μεταπτυχιακή εργασία θα ασχοληθούμε με το πρόβλημα που σχετίζεται με την ανισορροπία μεταξύ των κλάσεων σε προβλήματα ταξινόμησης και συγκεκριμένα στη δυαδική ταξινόμηση. Βασικός ταξινομητής που θα χρησιμοποιήσουμε είναι οι Μηχανές Διανυσματικής Υποστήριξης. Στην συγκεκριμένη εργασία ερχόμαστε αντιμέτωποι με διάφορες μεθόδους και αλγορίθμους από το πεδίο της μηχανικής μάθησης έχοντας ως στόχο την εύρεση ενός μοντέλου το οποίο θα έχει καλή προβλεπτική ικανότητα σε προβλήματα με μη ισορροπημένες κλάσεις μεταξύ των δεδομένων. Δεδομένου ότι στην περίπτωση των ιατρικών δεδομένων το πρόβλημα αυτό αποτελεί τον κανόνα και όχι την εξαίρεση, εφαρμόζουμε τις μεθόδους σε πέντε σύνολα δεδομένων που προέρχονται από τον ιατρικό κλάδο. Οι διάφορες μέθοδοι εξισορρόπησης που χρησιμοποιούμε αναφέρονται τόσο σε επιπέδο προ-επεξεργασίας των δεδομένων όσο και σε αλγοριθμικό επίπεδο. Ιδιαίτερα σημαντικό φαίνεται να μελετήσουμε τα μικρά μη ισορροπημένα σύνολα εκπαίδευσης ξεχωριστά από τα μεγάλα λόγω χρήζουν διαφορετικές μεταχείρισης. Ως εκ τούτου το πρώτο μέρος της ανάλυσης πραγματοποιείται σε μικρότερα σύνολα ιατρικών δεδομένων και το δεύτερο μέρος σε ένα μεγαλύτερο σύνολο δεδομένων.

Πιο συγκεκριμένα, το πρώτο κεφάλαιο ασχολείται με την βασική ιδέα του προβλήματος της ανισορροπίας μεταξύ των κλάσεων σε προβλήματα ταξινόμησης και το δεύτερο με τις τεχνικές και μεθόδους χειρισμού των μη ισορροπημένων δεδομένων. Το κεφάλαιο 3 παρουσιάζει το πρόβλημα της ανισσοροπίας των κλάσεων με τη χρήση των μηχανών διανυσματικής υποστήριξης και παρουσιάζονται διάφορες μέθοδοι που θα εφορμοστούν και στην πράξη αποδεικνύοντας τη σημαντικότητα των παρουσιαζόμενων τεχνικών. Στο τέταρτο κεφάλαιο αναφερόμαστε στην αξιολόγηση ενός μοντέλου ταξινόμησης με τη χρήση πολλαπλής διασταυρωμένης επικύρωσης, στην απόδοση των προαναφερθέντων ταξινομητών αλλά και σε διάφορα μέτρα απόδοσης που είναι απαραίτητο να ληφθούν υπόψη όταν αντιμετωπίζουμε προβλήματα με μη ισορροπημένα δεδομένα. Το κεφάλαιο 5 αποτελείται από δύο μέρη. Το πρώτο μέρος περιλαμβάνει την ανάλυση τεσσάρων συνόλων ιατρικών δεδομένων με τη χρήση αλγορίθμων εξισορρόπησης και το δεύτερο μέρος παρουσιάζει την ανάλυση ενός μεγάλου συνόλου ιατρικών δεδομένων με τη χρήση τόσο αλγοριθμικών όσο και μεθόδων προ-επεξεργασίας των δεδομένων. Στο έκτο και τελευταίο κεφάλαιο παρουσιάζονται τα συμπεράσματα της παραπάνω μελέτης.

Class Imbalanced Problem with Support Vector Machines

Keywords: *Support Vector Machines, class imbalanced problem, reweighted methods, medical data*

Abstract

In this master thesis we will deal with the problem associated with imbalance between classes in classification problems and specifically on binary classification. The main classifier we will use is the Support Vector Machines. In this work we are confronted with different methods and algorithms from the field of machine learning with the aim of finding a model that will have good predictive ability in problems with unbalanced distributions between classes. Since, in medical diagnosis problems this fact constitutes the rule rather than an exception; we applied many reweighted methods on five datasets from the medical diagnosis field. Various rebalancing methods are used and reported at both on pre-processing and on algorithmic level. The first part of the analysis performed in smaller datasets of medical data and the second part into a larger data set.

More specifically, the first chapter deals with the basic idea of the problem of imbalanced classes in classification problems and the second one with techniques and methods for handling such data. Chapter 3 presents the class imbalanced problem with Support Vector Machines presenting various methods that will be applied in practice demonstrating the significance of the presented techniques. In the fourth chapter we refer to the evaluation of a classification model using multiple cross-validation, the performance of the aforementioned classifiers and various performance measures need to be taken into account when facing problems with imbalanced data. Chapter 5 consists of two parts. The first part comprises the analysis of four sets of medical data using reweighted algorithms and the second part presents the analysis of a large set of medical data using both algorithmic and pre-processing methods. The sixth and final chapter presents the conclusions of the above study.

ΚΕΦΑΛΑΙΟ 1

Το Πρόβλημα της Ανισορροπίας Μεταξύ των Κλάσεων
(The problem of handling the class imbalance problem)

1.1Εισαγωγή

Συχνά οι φυσικές διεργασίες δημιουργούν σύνολα δεδομένων στα οποία δεν υπάρχει κανονική κατανομή μεταξύ των κλάσεων, δηλαδή κάποιες παρατηρήσεις παράγονται με μεγαλύτερη συχνότητα από κάποιες άλλες. Έτσι δημιουργείται το πρόβλημα της ανισορροπίας μεταξύ των κλάσεων. Σε γενικές γραμμές το πρόβλημα των μη ισορροπημένων κλάσεων παρατηρείται σε δύο περιπτώσεις προβλημάτων: είτε διότι υπάρχει μία φυσική ανισορροπία μεταξύ των δύο κλάσεων[1], είτε λόγω της σπανιότητας των περιπτώσεων[2] (π.χ. παραδείγματα ή δείγματα). Γενεσιουργές αιτίες αυτής της ανισορροπίας θα μπορούσε να είναι η έλλειψη περιστατικών στη φύση για συγκεκριμένα φαινόμενα ή ενδεχομένως ανεπαρκή χρήματα ή χρόνος για τη συλλογή ενός επαρκούς όγκου δεδομένων. Κατά τα τελευταία έτη, πολλοί ερευνητές έχουν μελετήσει το πρόβλημα της ανισορροπίας μεταξύ των κλάσεων. Ο Weiss (2004) παρουσίασε μια μελέτη σχετική με το πεδίο της μάθησης από μη ισορροπημένα σύνολα δεδομένων. Το έργο του εστιάστηκε ιδιαίτερα σε προβλήματα σχετικά με τον εντοπισμό σπάνιων περιπτώσεων στην εξόρυξη δεδομένων, δηλαδή περιπτώσεων που εμφανίζονται ελάχιστες φορές. Ο Weiss όρισε δύο τύπους «σπανιότητας», (1) τις σπάνιες κλάσεις (κατηγορίες) που αναφέρονται στη μεταβλητή κλάσης ή αλλιώς στη μεταβλητή απόκρισης και (2) τις σπάνιες περιπτώσεις (πειραματικές εκτελέσεις) που αναφέρονται

[1] Κλάσεις ή κατηγορίες της μεταβλητής απόκρισης.
[2] Περιπτώσεις ή πειραματικές εκτελέσεις ή δείγμα. Για παράδειγμα το πλήθος των ασθενών σε ένα πείραμα ιατρικών δεδομένων.

στο μέγεθος του δείγματος. Μια σπάνια κατηγορία περιέχει σχετικά μικρότερο αριθμό περιπτώσεων από τις άλλες κατηγορίες, ενώ μια σπάνια περίπτωση καταδεικνύει ένα μικρό υποσύνολο του χώρου των δεδομένων.

Οι αλγόριθμοι εκμάθησης χωρίς επίβλεψη, όπως είναι η ομαδοποίηση (clustering) μπορεί να βοηθήσουν στον προσδιορισμό μιας σπάνιας περίπτωσης. Γενικότερα, η ανισορροπία των κλάσεων σχετίζεται με τις σπάνιες κλάσεις καθώς επίσης σχετίζεται και με προβλήματα στην ταξινόμηση. Ο Weiss υποστήριξε ότι τα τυπικά μέτρα αξιολόγησης (όπως είναι η ακρίβεια) δεν μπορούν να περιγράψουν επαρκώς τη σημασία αυτής της σπανιότητας, έτσι που η εξόρυξη δεδομένων (data mining) δεν είναι πιθανό να χειριστεί σπάνιες κλάσεις και σπάνιες περιπτώσεις. Οι Monard et al. (2002), συζήτησαν διάφορα θέματα που σχετίζονται με τη μάθηση όταν η κατανομή των κλάσεων είναι ασύμμετρη, όπως τη σχέση μεταξύ της ευαίσθητης-με-κόστος μάθησης[3] (cost-sensitive learning) και της κατανομής των κλάσεων. Παράλληλα σχολίασαν τους περιορισμούς που παρουσιάζει το μέτρο της ακρίβειας και το ποσοστό σφάλματος στη μέτρηση της απόδοσης των ταξινομητών.

1.2 Ταξινόμηση των δεδομένων

Η ταξινόμηση δεδομένων είναι μία διαδικασία η οποία βρίσκει τις κοινές ιδιότητες μεταξύ ενός συνόλου αντικειμένων σε μία βάση δεδομένων και ταξινομεί τα αντικείμενα αυτά σε διαφορετικές κλάσεις (τάξεις) σύμφωνα με ένα μοντέλο ταξινόμησης. Για να κατασκευάσουμε ένα τέτοιο μοντέλο ταξινόμησης, μία δειγματική βάση δεδομένων $E = \{t_1, t_2, \dots, t_n,\}$ θεωρείται ως το σύνολο εκπαίδευσης (training set) στο οποίο κάθε εγγραφή αποτελείται από το ίδιο σύνολο πολλαπλών χαρακτηριστικών όπως οι εγγραφές σε μία μεγάλη βάση δεδομένων W και επιπρόσθετα κάθε εγγραφή έχει μία γνωστή ετικέτα (label) κλάσης. Το σύνολο των κλάσεων το συμβολίζουμε με $C = \{c_1, c_2, \dots, c_n,\}$.

Ο αντικειμενικός σκοπός της ταξινόμησης είναι πρώτον να αναλύσει τα δεδομένα του συνόλου εκπαίδευσης και δεύτερον να αναπτύξει μία ακριβή περιγραφή ή ένα ακριβές μοντέλο για κάθε κλάση χρησιμοποιώντας τα χαρακτηριστικά που είναι διαθέσιμα στα δεδομένα. Με άλλα λόγια το πρόβλημα της κατηγοριοποίησης έγκειται στον ορισμό μίας απεικόνισης $f : E \longrightarrow C$ όπου κάθε εγγραφή t_i ανατίθεται σε μία κλάση c_i. Οι περιγραφές κλάσεων που προκύπτουν χρησιμοποιούνται στη συνέχεια για να ταξινομήσουν μελλοντικά δεδομένα (test set) στη βάση δεδομένων W ή για να αναπτύξουν μια καλύτερη περιγραφή την οποία ονομάζουμε «κανόνες ταξινόμησης» για κάθε κλάση στη βάση δεδομένων. Επομένως, μπορούμε να θεωρήσουμε ότι με την ταξινόμηση διαμερίζουμε το σύνολο Ε σε κλάσεις ισοδυναμίας και επιπλέον ότι το

[3] cost-sensitive learning: το είδος αυτού του τύπου μάθησης δηλώνει ότι η διαδικασία και τα αποτελέσματα της μάθησης επηρεάζονται από κάποιο ή κάποια κόστη που συμπεριλαμβάνονται στο μόντέλο της μάθησης

πρόβλημα της πρόβλεψης είναι ένα πρόβλημα ταξινόμησης όπου έχουμε άπειρο αριθμό κλάσεων.

Η ταξινόμηση βρίσκει πολλές εφαρμογές σε διάφορους τομείς όπως στην ιατρική διάγνωση και στο marketing και αποτελεί αντικείμενο μελέτης για τη στατιστική, τη μηχανική μάθηση και βέβαια το data mining. Πρόκειται για μάθηση με επίβλεψη (supervised learning) καθώς οι ομάδες ταξινόμησης είναι εκ των προτέρων γνωστές και το πραγματικό αποτέλεσμα κάθε υποδείγματος είναι επίσης γνωστό. Επομένως, είναι δυνατό να μετράμε το βαθμό αξιοπιστίας σε μη χρησιμοποιημένα για τη διαμόρφωση της αντίληψης δεδομένα ή υποκείμενα, ανάλογα με το βαθμό αποδοχής της περιγραφής.

Η τυπική προσέγγιση που χρησιμοποιούν οι τεχνικές ταξινόμησης είναι η δημιουργία ενός μοντέλου μέσω της αξιολόγησης του συνόλου δεδομένων εκπαίδευσης και η εφαρμογή του μοντέλου σε νέα δεδομένα. Οι πιο κοινές τεχνικές είναι τα δέντρα αποφάσεων (decision trees), τα νευρωνικά δίκτυα (Neural Networks), οι μηχανές διανυσματικής υποστήριξης (Support Vector Machines), η λογιστική παλινδρόμηση (logistic regression) και τα Bayesian Network Models.

1.3 Ορισμός του προβλήματος της ανισορροπίας των κλάσεων

Το πρόβλημα της ανισορροπίας στα σύνολα δεδομένων εμφανίζεται στην ταξινόμηση, όπου ο αριθμός των παρατηρήσεων που ανήκουν σε μία κατηγορία/κλάση είναι κατά πολύ μικρότερος από εκείνον των άλλων κατηγοριών. Η κύρια πρόκληση στο πρόβλημα της ανισορροπίας είναι ότι η κλάση που αποτελεί την μειοψηφία είναι και η κλάση ενδιαφέροντος. Ωστόσο η κλασική μοντελοποίηση των αλγορίθμων ταξινόμησης υποθέτει ότι έχουμε κανονική κατανομή μεταξύ των κλάσεων με αποτέλεσμα οι ταξινομητές να είναι μεροληπτικοί απέναντι στην κλάση πλειοψηφίας αγνοώντας την μειοψηφική κλάση. Αυτό σημαίνει ότι οι ταξινομητές έχουν την τάση να επικεντρώνονται στην κλάση πλειοψηφίας και να αγνοούν την μειοψηφική κλάση η οποία τις περισσότερες φορές αντιπροσωπεύει και την κλάση του ενδιαφέροντος. Στη μάθηση μηχανής, τα μη ισορροπημένα σύνολα δεδομένων αποτελούν ένα κρίσιμο πρόβλημα το οποίο κατακλύζει πλήθος εφαρμογών όπως είναι η ανίχνευση των δόλιων κλήσεων, η βιοιατρική, η μηχανική, η τηλεπισκόπηση, η επιστήμη των ηλεκτρονικών υπολογιστών αλλά και οι κατασκευαστικές βιομηχανίες. Έχουν προταθεί πολλές προσεγγίσεις προκειμένου να ξεπεραστούν τα προβλήματα που δημιουργούνται από το εν λόγω πρόβλημα. Εκτενέστερα θα ασχοληθούμε στη συνέχεια.

Τα τελευταία έτη το πρόβλημα της ανισορροπίας μεταξύ των κλάσεων έχει τύχει σημαντικής προσοχής σε τομείς όπως η Μηχανική Μάθηση (Machine Learning) και η αναγνώριση προτύπων (Pattern Recognition). Ένα σύνολο δεδομένων που αποτελείται από δύο κλάσεις (δηλαδή η μεταβλητής απόκρισης y είναι Binary/δυαδική) θεωρείται μη ισορροπημένο, όταν μία από τις κλάσεις του σε μεγάλο βαθμό υπο-εκπροσωπείται σε αντίθεση με την άλλη κλάση, που αποτελεί την κλάση πλειοψηφίας. Η εν λόγω

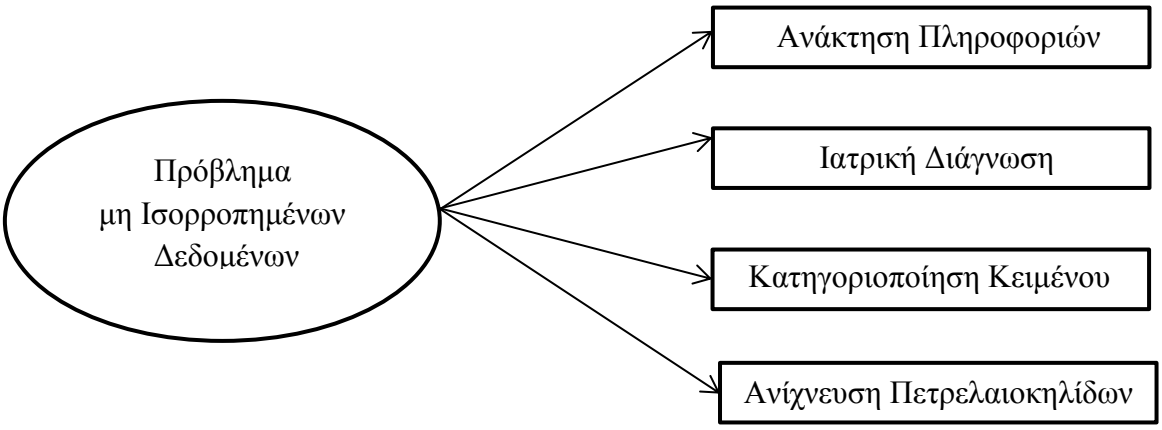
κατάσταση θα μπορούσαμε να πούμε ότι αποτελεί ένα πρόβλημα και μάλιστα ζωτικής σημασίας σε πολλές εφαρμογές του πραγματικού κόσμου, όπου είναι ιδιάιτερα δαπανηρό το να ταξινομηθούν εσφαλμένα παραδείγματα από την κλάση της μειοψηφίας. Χαρακτηριστικά παραδείγματα αποτελούν η ανίχνευση δόλιων τηλεφωνημάτων (fraudulent telephone calls), η διάγνωση σπάνιων νόσων (diagnosis of rare diseases), ανάκτηση πληροφοριών (information retrieval), κατηγοριοποίηση κειμένου (text categorization) και οι εργασίες φιλτραρίσματος (filtering tasks) (García et al.(2007)). Για την αντιμετώπιση αυτού του προβλήματος έχουν ήδη προταθεί διάφορες προσεγγίσεις, οι οποίες μπορούν να κατηγοριοποιηθούν σε δύο ομάδες:

1. *Εσωτερικές προσεγγίσεις (internal approaches).*
   Δημιουργία καινοτόμων αλγορίθμων ή αλλαγή υπαρχόντων έτσι ώστε να ληφθεί υπόψη το πρόβλημα της ανισορροπίας μεταξύ των κλάσεων.
2. *Εξωτερικές προσεγγίσεις (external approaches)*
   Προ-επεξεργασία των δεδομένων, ώστε να μειωθεί η επίδραση που προκαλείται από την ανισορροπία μεταξύ των κλάσεων.

Οι εσωτερικές προσεγγίσεις έχουν το μειονέκτημα ότι αποτελούν μία αλγοριθμική προσέγγιση, ενώ οι εξωτερικές προσεγγίσεις είναι ανεξάρτητες από τον ταξινομητή που χρησιμοποιείται κάτι που τις κάνει περισσότερο ευέλικτες. Για το λόγο αυτό η μέθοδος CO^2RBFN εφαρμόζεται στην επίλυση του προβλήματος ταξινόμησης με μη ισορροπημένες κλάσεις (Perez-Godoy et al. (2010)). Στις περισσότερες εφαρμογές, η ακριβής κατάταξη των παραδειγμάτων που αποτελούν την κλάση μειοψηφίας είναι πιο σημαντική από αυτών της κλάσης πλειοψηφίας. Για παράδειγμα, στην πρόβλεψη αλληλεπιδράσεων πρωτεΐνης-πρωτεΐνης, ο αριθμός των μη αλληλεπιδράσεων πρωτεΐνης είναι μεγαλύτερος από τον αριθμό των πρωτεϊνών που αλληλεπιδρούν. Επίσης, στα προβλήματα ανάλυσης ιατρικών δεδομένων, ο αριθμός των περιπτώσεων που έχουν τη νόσο είναι συνήθως μικρότερος από τον αριθμό των περιπτώσεων που δεν έχουν την ασθένεια (Thanathamathee και Lursinsap (2013).

Η υψηλή δραστηριότητα της προόδου στα προβλήματα μάθησης με μη ισορροπημένα δεδομένα παραμένει γνωστή ανάμεσα σε όλες τις τρέχουσες εξελίξεις και αποτελεί ένα δύσκολο πεδίο έρευνας. Η ικανότητα των μη ισορροπημένων δεδομένων να μειώνουν σημαντικά την απόδοση των περισσοτέρων τυποποιημένων αλγορίθμων μάθησης αποτελεί ένα θεμελιώδες ζήτημα όταν αντιμετωπίζουμε ένα πρόβλημα μάθησης με μη ισορροπημένα δεδομένα. Όπως ήδη έχουμε αναφέρει το πρόβλημα αυτό μπορουμε να το συναντήσουμε σε πολλά διαφορετικά είδη πεδίων. Με σκοπό να αναδείξουμε τις επιπτώσεις που επιφέρει το προαναφερθέν πρόβλημα παρουσιάζουμε ορισμένα από τα πεδία, όπως η ιατρική διάγνωση, η κατηγοριοποίηση κειμένου, η ανίχνευση πετρελαιοκηλίδων σε εικόνες ραντάρ και η ανάκτηση πληροφοριών.

Αυτά παριστάνονται στο Σχήμα 1.1.



Σχήμα 1.1: Παραδείγματα Πεδίων Μη Ισορροπημένων Δεδομένων

Ανάκτηση πληροφοριών (Information Retrieval)

Τα IR πιθανοτικά μοντέλα από την οπτική της ταξινόμησης των προτύπων έχουν δείξει ότι είναι ιδιαίτερα αποδοτικά στη φύση. Στα μοντέλα IR διερευνάται τόσο η εφαρμογή των διαφορετικών ταξινομητών όπως των Μηχανών Διανυσματικής Υποστήριξης (SVMs) όσο και των MEs. Οι εντυπωσιακές θεωρητικές ιδιότητες και η σημαντική χρησιμότητα τους σε αυτά τα μοντέλα βρίσκεται στην ικανότητά τους να μαθαίνουν αυτόματα από μια σειρά από χαρακτηριστικά που επηρεάζουν τη συνάφεια. Τα πειράματα στην κατά περίπτωση (ad-hoc) ανάκτηση πληροφοριών αποδεικνύουν ότι χρησιμοποιούν τον ίδιο τύπο των χαρακτηριστικών, και πως οι Μηχανές Διανυσματικής Υποστήριξης (SVMs) λειτουργούν το ίδιο καλά με τα LMs στις περισσότερες περιπτώσεις. Η ικανότητα των SVMs στην εκμάθηση μιας ποικιλίας χαρακτηριστικών (feautures) στη περίπτωση του γνωστού προβλήματος της εύρεσης σελίδας/home-page (home-page finding task), τις κάνουν να ξεπερνούν τις βασικές εκτελέσεις οι οποίες χρησιμοποιούν μόνο τα χαρακτηριστικά που βασίζονται στο περιεχόμενο περίπου στο 50% των MRR. (Nallapati (2004))

Ιατρική Διάγνωση (Medical diagnosis)

Ένα άλλο σημαντικό ζήτημα είναι ότι τα ιατρικά σύνολα δεδομένων που χρησιμοποιούνται για τη μηχανική μάθηση πρέπει να είναι αντιπροσωπευτικά της γενικής συχνότητας εμφάνισης της νόσου που βρίσκεται υπό μελέτη. Στην ιατρική διαγνωστική περιοχή REMED μπορούν να χρησιμοποιηθούν ιδιαίτερα ανταγωνιστικοί αλγόριθμοι. Ωστόσο, η REMED δεν προσποιείται ότι είναι η λύση της μηχανικής μάθησης στην ιατρική διαγνωστική, αλλά μια καλή προσέγγιση με τα προτιμητέα χαρακτηριστικά για να αποκρυπτογραφήσει ιατρικές αναλυτικές αρμοδιότητες, τη δυνατότητα κατανόησης της διαγνωστικής γνώσης, την καλή απόδοση, την ικανότητα να διαφωτίσει διάφορες αποφάσεις, και την ικανότητα του αλγορίθμου να μειώσει τον

αριθμό των δοκιμών που απαιτούνται για την απόκτηση αξιόπιστης διάγνωσης (Mena and Gonzalez (2006)).

Κατηγοριοποίηση Κειμένου (text categorization)

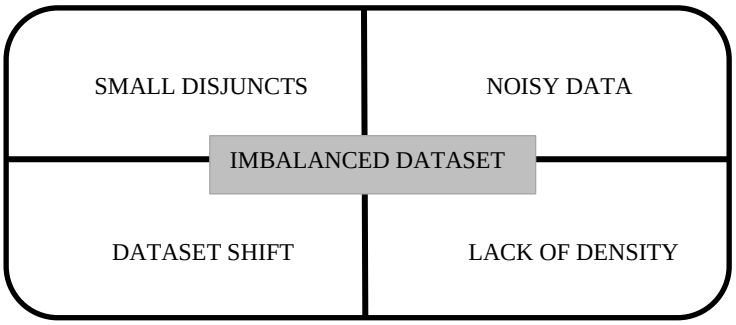
Οι στρατηγικές δειγματοληψίας, όπως η Υπερδειγματοληψία (Oversampling) και η Υποδειγματοληψία (Subsampling) είναι σύγχρονες στην αντιμετώπιση του προβλήματος της ανισορροπίας μεταξύ των κλάσεων. Οι τρεις τύποι των ταξινομητών όπως είναι τα SVM, τα KNN και τα Naive-Bayes, εφαρμόζονται στην αναζήτηση της επιστημονικής βάσης δεδομένων PubMed. Στην ταξινόμηση των βιοϊατρικών κειμένων χρησιμοποιούνται τρεις τύποι λεξικών. Τα πειράματα διεξάγονται με τρία διαφορετικά λεξικά, όπως είναι το NLPBA, το BioCreative, και ένα κατά περίπτωση (ad-hoc) υποσύνολο της UniProt βάσης δεδομένων, χρησιμοποιώντας τους παραπάνω ταξινομητές και τις στρατηγικές δειγματοληψίας. Τα καλύτερα αποτελέσματα ελήφθησαν με τα λεξικά NLPBA και Protein και τον ταξινομητή SVM χρησιμοποιώντας τη μέθοδο εξισορρόπησης της Υποδειγματοληψίας. Αυτά τα αποτελέσματα συγκρίθηκαν με εκείνα που λαμβάνονται με τη χρήση των δημόσιων corpus TREC Genomics 2005 (Borrajo et al.(2011)).

Ανίχνευση Πετρελαιοκηλίδων (Detection of oil spills)

Το πρόβλημα των μη ισορροπημένων συνόλων προκύπτει πιο συχνά σε εφαρμογές και μειώνει σημαντικά την απόδοση των συμβατικών τεχνικών ταξινόμησης. Έχουν προταθεί πολλές μέθοδοι για την αντιμετώπιση του προβλήματος των μη ισορροπημένων κλάσεων. Τουλάχιστον μια μεγάλης κλίμακας συγκριτική μελέτη απαιτείται για να αξιολογήσει τα συγκριτικά πλεονεκτήματα αυτών των μεθόδων αλλά και το πως αυτές λειτουργούν. Πολλές μέθοδοι που δημιουργήθηκαν, όπως για παράδειγμα ο αλγόριθμος SHRINK, μπορούν αναμφίβολα να βελτιωθούν με περαιτέρω έρευνα. Ιδιαίτερα σημαντικό φαίνεται να μελετήσουμε τα μικρά μη ισορροπημένα σύνολα εκπαίδευσης ξεχωριστά από τα μεγάλα. Η μάθηση από τα ομαδοποιημένα παραδείγματα είναι ένα άλλο ζήτημα που απαιτεί περαιτέρω έρευνα. Τέτοιου είδους μάθηση σχετίζεται με τα θέματα της μάθησης στην εμφάνιση των περιστάσεων, όπως οι ομάδες συχνά χαρακτηρίζουν το άγνωστο πλαίσιο εντός του οποίου συλλέχθηκαν τα παραδείγματα εκπαίδευσης (Kubat, et al. (1998)).

1.4 Χαρακτηριστικά των μη ισορροπημένων δεδομένων

Κάθε σύνολο δεδομένων που εμφανίζει μία άνιση κατανομή μεταξύ των κλάσεων μπορεί να θεωρηθεί ως μη ισορροπημένο (He και Garcia (2009)). Μία μορφή της ανισορροπίας αναφέρεται ως ανισορροπία μεταξύ των κλάσεων και δεν είναι ασυνήθιστο μεταξύ των μη ισορροπημένων κλάσεων η ταξινόμηση της τάξεως του 100:1, 1000:1, και 10000:1, όπου σε κάθε περίπτωση, μία κλάση υπερτερεί κατά πολύ της άλλης (He and Shen (2007) Kubat et al.(1998), Pearson et al.(2003)).



Σχήμα 1.2: Χαρακτηριστικά του μη ισορροπημένου συνόλου δεδομένων

Το πρόβλημα που σχετίζεται με τη χρήση των δεδομένων εν γένει γίνεται διαφορετικό σε αυτό το πρόβλημα ταξινόμησης. Αυτό διευκολύνει την ανάπτυξη των σημερινών μοντέλων σχετικά με: την έλλειψη πυκνότητας πιθανότητας στα δεδομένα εκπαίδευσης (*lack of density*), την παρουσία των μικρών disjuncts (*small disjuncts*), την ταυτοποίηση των θορυβωδών δεδομένων (*noisy data*), την μετατόπιση του συνόλου δεδομένων μεταξύ της κατανομής εκπαίδευσης (training) και δοκιμής (test) και τη σημασία των οριακών περιπτώσεων. Όλα αυτά απεικονίζονται στο Σχήμα 1.2 (Lopez et al.(2013)).

Small disjuncts

Η παρουσία των μη ισορροπημένων κλάσεων συνδέεται στενά με το πρόβλημα των μικρών disjuncts. Το πρόβλημα αυτό παρουσιάζεται, όταν οι εκτελέσεις/concepts παρουσιάζονται σε μικρές συστάδες, το οποίο προκύπτει ως άμεσο αποτέλεσμα της υπο-εκπροσώπησης των επιμέρους εκτελέσεων (Jo και Japkowicz (2004)). Το πρόβλημα των μικρών disjuncts γίνεται εντονότερο για τους αλγορίθμους ταξινόμησης που βασίζονται στη μεθοδολογία-προσέγγιση του «διαίρει και βασίλευε». Αυτό συνίσταται στην υποδιαίρεση του αρχικού προβλήματος σε μικρότερα προβλήματα, όπως είναι η διαδικασία που χρησιμοποιείται στα δέντρα απόφασης (decision trees), και μπορεί να οδηγήσει σε "κατακερματισμό" των δεδομένων, δηλαδή, στην εξασφάλιση διαχωρισμού των δεδομένων με λίγες αναπαραστάσεις των εκτελέσεων/περιπτώσεων (Weiss 2004).

Lack of density

Ένα από τα μεγαλύτερα προβλήματα που μπορεί να προκύψουν στην ταξινόμηση είναι το μικρό μέγεθος του δείγματος (Raudys και Jain (1991)). Το ζήτημα αυτό σχετίζεται με την «έλλειψη πυκνότητας» ή «έλλειψη πληροφορίας», όπου οι αλγόριθμοι επαγωγής δεν έχουν αρκετά δεδομένα για να μπορούν να κάνουν γενικεύσεις σχετικά με την κατανομή των δειγμάτων, μια κατάσταση που γίνεται όλο και πιο δύσκολη με την παρουσία δεδομένων υψηλής διάστασης (Parpoula et al. (2013)) και την παρουσία μη ισορροπημένων συνόλων δεδομένων (Drosou et al. (2014)). Ο συνδυασμός των μη ισορροπημένων δεδομένων και το πρόβλημα του μικρού μεγέθους του δείγματος αποτελεί μια νέα πρόκληση για την ερευνητική κοινότητα (Wasikowski και Chen (2010)).

Noisy data

Στην περίπτωση των μη ισορροπημένων δεδομένων, η παρουσία του θορύβου έχει μεγαλύτερη επίδραση στις κλάσεις μειοψηφίας απ’ ότι στις κλάσεις που αποτελούν την πλειοψηφία, δηλαδή στις συνήθεις περιπτώσεις (Weiss (2004)). Οι Khoshgoftaar et al. (2011), παρουσίασαν μια μελέτη σχετικά με την σημασία των δεδομένων θορύβου και των μη ισορροπημένων δεδομένων χρησιμοποιώντας τεχνικές bagging και boosting. Τα αποτελέσματα έδειξαν την καλή προσαρμογή της προσέγγισης bagging χωρίς αντικατάσταση, και οι συγγραφείς συνιστούν τη χρήση τεχνικών μείωσης του θορύβου πριν από την εφαρμογή για διαδικασιών boosting.

Dataset shift

Το πρόβλημα της μετατόπισης του συνόλου δεδομένων (Alaiz-Rodriguez και Japkowicz (2008)), ορίζεται ως η περίπτωση όπου τα δεδομένα εκπαίδευσης και δοκιμής ακολουθούν διαφορετικές κατανομές. Είναι ένα γνωστό πρόβλημα που μπορεί να επηρεάσει όλα τα είδη των προβλημάτων ταξινόμησης, και συχνά φαίνεται να οφείλεται σε ζητήματα μεροληψίας στην επιλογή του δείγματος. Το θέμα του Dataset shift σχετίζεται άμεσα με την ταξινόμηση των μη ισορροπημένων κλάσεων, διότι σε εξαιρετικά μη ισορροπημένα πεδία, η κλάση μειοψηφίας είναι συνήθως ευαίσθητη σε σφάλματα ταξινόμησης, λόγω του χαμηλού αριθμού παραδειγμάτων (Moreno-Torres και Herrera (2010)).

1.5 Χειρισμός του προβλήματος των μη ισορροπημένων κλάσεων

Οι διάφορες προσεγγίσεις που χρησιμοποιούνται για την αντιμετώπιση του προβλήματος της ανισορροπίας μεταξύ των κλάσεων (class imbalance problem) μπορούν να ομαδοποιηθούν σε τρεις κατηγορίες:

- Αλλαγή της κατανομής των κλάσεων (τροποποιώντας τα ίδια τα δεδομένα έτσι ώστε να εξισορροπηθεί η ασυμμετρία των συνόλων δεδομένων (data level),
- Προσαρμογή των ταξινομητών (προσαρμόζοντας βασικούς αλγορίθμους ταξινόμησης σε μη ισορροπημένα σύνολα δεδομένων αποδίδοντας ένα κόστος (cost) ή αλλιώς ένα βάρος (weight) στις λανθασμένα ταξινομημένες περιπτώσεις), και
- Σύνολο μεθόδων μάθησης (χρησιμοποιώντας ένα συνδυασμό πολλαπλών ταξινομητών με πολλαπλά σύνολα δεδομένων).

Στο δεύτερο κεφάλαιο γίνεται μία πιο αναλυτική αναφορά και μία πιο λεπτομερής και εν μέρη διαφορετική κατηγοριοποίηση των μεθόδων αλλά και των αλγορίθμων που μας βοηθούν να χειριστούμε τέτοιου είδους προβλήματα.

ΚΕΦΑΛΑΙΟ 2

Μέθοδοι Χειρισμού Μη Ισορροπημένων Δεδομένων
(Techniques for imbalanced data set problems)

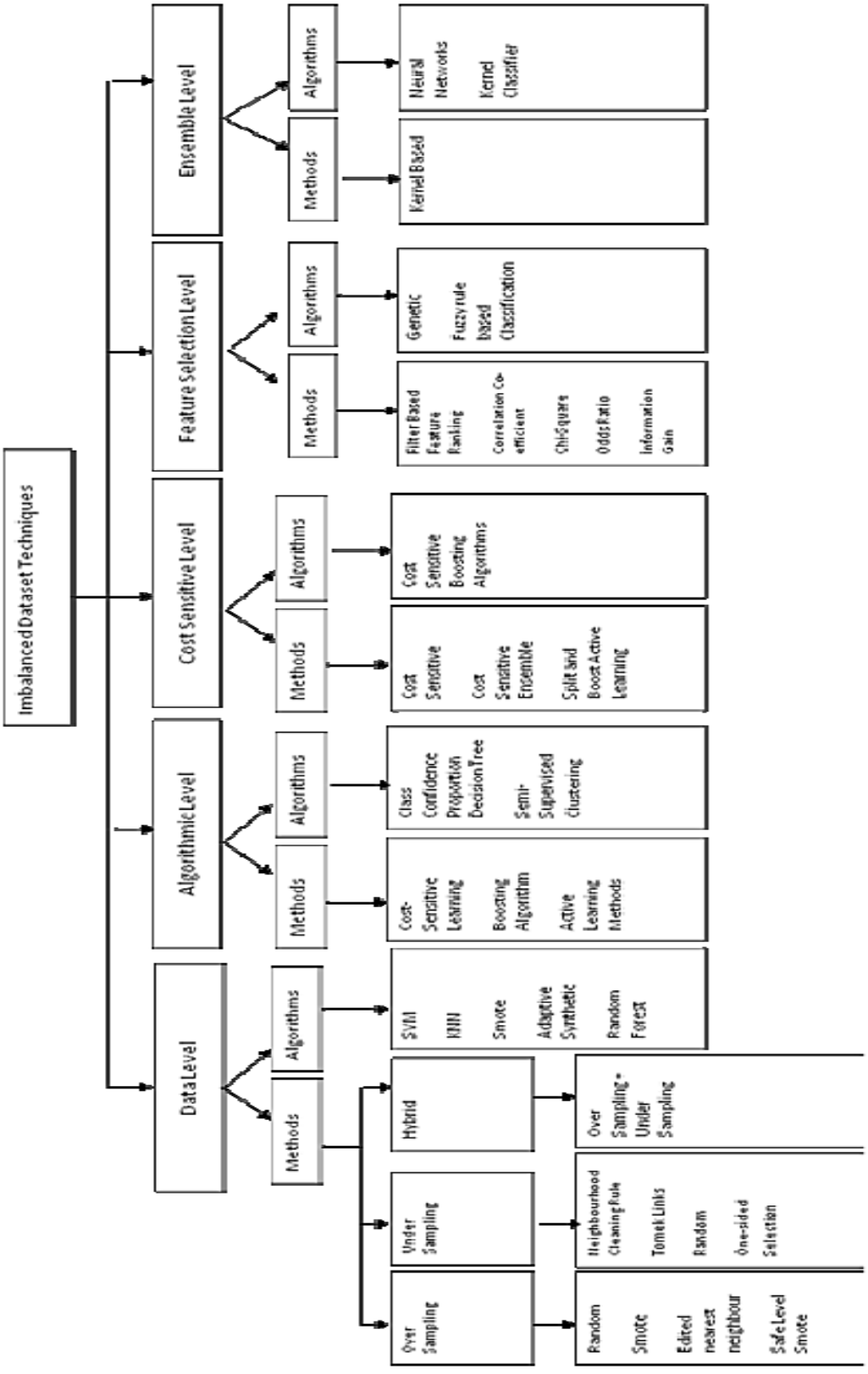
2.1 Εισαγωγικά στοιχεία

Οι βασικοί αλγόριθμοι μηχανικής μάθησης αδυνατούν να διαχειριστούν προβλήματα με μη ισορροπημένα δεδομένα με την έννοια ότι δεν μπορούν να ταξινομήσουν το σύνολο δεδομένων με ιδιαίτερη επιτυχία. Αυτό συμβαίνει διότι το σφάλμα ταξινόμησης στην κλάση πλειοψηφίας κυριαρχεί του σφάλματος ταξινόμησης στην κλάση μειοψηφίας. Αυτή η κυριαρχία οδηγεί στην προώθηση της διαχωριστικής συνάρτησης (ή συνάρτησης απόφασης) μακριά από την κλάση πλειοψηφίας έτσι ώστε να μειωθεί το σφάλμα ταξινόμησης κατά τη διαδικασία προσαρμογής του βάρους (weight adjusting process). Ως εκ τούτου, τα δεδομένα του συνόλου δοκιμής ταξινόμουνται εσφαλμένα πιο συχνά στην κλάση μειοψηφίας από εκείνα που ανήκουν στην κλάση πλειοψηφίας.

Οι τεχνικές για να χειριστούμε το πρόβλημα της ανισορροπίας των κλάσεων μπορούν να κατηγοριοποιηθούν ως τεχνικές

- στο επίπεδο των δεδομένων (***data level***),
- σε αλγοριθμικό επίπεδο (***algorithmic level***),
- στο επίπεδο του ευαίσθητου κόστους (***cost sensitive level***)
- σε επίπεδο επιλογής χαρακτηριστικών και (***feature selection level***)
- στο επίπεδο ενός συνόλου μεθόδων μάθησης (***ensemble level***).

Αυτό φαίνεται παραστατικά στο ακόλουθο Σχήμα (Σχήμα 2.1)



Σχήμα 2.1 Ταξινόμηση των Τεχνικών για την επίλυση των μη ισορροπημένων δεδομένων

2.2 Προσεγγίσεις σε επίπεδο δεδομένων (data level approaches)

Η προσέγγιση αυτή λειτουργεί σε ένα στάδιο προ-επεξεργασίας (pre-processing), απευθείας στο χώρο των δεδομένων, και προσπαθεί να εξισορροπήσει τις κατανομές μεταξύ των κλάσεων. Είναι αυτόνομη διαδικασία με την έννοια ότι δεν αποτελεί ένα πραγματικό στάδιο της διαδικασίας ταξινόμησης και ως εκ τούτου οι τεχνικές που την απαρτίζουν μπορούν να χρησιμοποιηθούν με περισσότερη ευελιξία. Οι πιο αξιόλογες προσεγγίσεις χρησιμοποιούν μια στρατηγική υπερδειγματοληψίας (Oversampling) που εισάγει τεχνητά παραδείγματα μέσα στο χώρο των δεδομένων. Η πιο γνωστή τεχνική είναι η SMOTE (Chawla et al. (2002)), αν και υπάρχουν, πιο πρόσφατες, βελτιωμένες εναλλακτικές λύσεις, όπως η ADASYN (He et al. (2008)), ή η RAMO (Chen et al. (2010)). Οι μέθοδοι υπερδειγματοληψίας ωστόσο μπορεί επίσης να οδηγήσουν σε άλλα προβλήματα, όπως είναι η μετατόπιση της κατανομής των κλάσεων μετά από πολλές επαναλήψεις (Krawczyk et al. (2012)).

2.2.1 Εισαγωγή-Αλλαγή της κατανομής των κλάσεων

Η ενότητα αυτή αφορά στην πρώτη κατηγορία αντιμετώπισης του προβλήματος ανισορροπίας των κλάσεων. Η προσέγγιση αφορά στην αλλαγή της κατανομής των κλάσεων στο επίπεδο των δεδομένων προκειμένου να τροποποιήσουμε την κατανομή στα σύνολα δεδομένων εκπαίδευσης. Δεδομένου ότι υπάρχουν πολύ περισσότερες περιπτώσεις (πειραματικές εκτελέσεις) που ανήκουν στην κλάση πλειοψηφίας (majority class) από την κλάση μειοψηφίας (minority class), η κατανομή των κλάσεων μπορεί να εξισορροπηθεί χρησιμοποιώντας μεθόδους υπο-δειγματοληψίας (under-sampling) της κλάσης πλειοψηφίας, υπερ-δειγματοληψίας (over-sampling) της κλάσης μειοψηφίας αλλά και συνδυασμό αυτών των δύο ή κάποια άλλη μέθοδο δειγματοληψίας. Μελέτες έχουν δείξει ότι ένα ισορροπημένο σύνολο δεδομένων παρέχει βελτιωμένη απόδοση ταξινόμησης σε σχέση με ένα μη ισορροπημένο σύνολο δεδομένων. Έχουν υπάρξει πολλές μελέτες σχετικά με την αλλαγή της κατανομής των κλάσεων όπως των Laurikkala (2001) και των Estabrooks et al. (2004). Επιπρόσθετα, ο Weiss (2003) διερεύνησε την επίδραση της κατανομής των κλάσεων στην περίπτωση της ταξινόμησης με δέντρα αποφάσεων, αλλάζοντας την κατανομή των κλάσεων για την επίτευξη διαφορετικών ποσοστών μεταξύ της κλάσης μειοψηφίας και πλειοψηφίας και μετρώντας την απόδοση χρησιμοποιώντας την ακρίβεια (accuracy) και την περιοχή κάτω από την ROC καμπύλη (AUC: Area under the Curve).

Θα μπορούσαμε να αναφέρουμε τρεις βασικές τεχνικές που χρησιμοποιούνται στην εξισορρόπηση των κλάσεων. Η κατηγοριοποίηση μπορεί να γίνει ως εξής:

- heuristic και non-heuristic under-sampling (ευρετική και μη-ευρετική υπό-δειγματοληψία)

- heuristic και non-heuristic over-sampling (ευρετική και μη-ευρετική υπερ-δειγματοληψία)
- advanced sampling, (προηγμένη δειγματοληψία).

Ο Japkowicz (2000) σύγκρινε πολλαπλές μεθόδους εξισορρόπησης και κατέληξε στο συμπέρασμα ότι τόσο οι τεχνικές υπο-δειγματοληψίας όσο και οι τεχνικές υπερ-δειγματοληψίας είναι πολύ αποτελεσματικές για την αντιμετώπιση του προβλήματος ανισορροπίας μεταξύ των κλάσεων.

2.2.2 Μέθοδοι

2.2.2.1 Μέθοδοι δειγματοληψίας (Sampling Methods)

Ο Jong Myong Choi πρότεινε μια επαναληπτική μέθοδο δειγματοληψίας (iterative sampling methodology) που χρησιμοποιήθηκε για την παραγωγή μικρότερων συνόλων μάθησης με την εξάλειψη των περιττών περιπτώσεων. Η μέθοδος αυτή ενσωματώνει τους μηχανισμούς της πληροφόρησης και της αντιπροσωπευτικής υπό-δειγματοληψίας για να επιταχύνει τη διαδικασία μάθησης των μη ισορροπημένων δεδομένων με SVM. Ως εκ τούτου για μεγάλης κλίμακας μη ισορροπημένα σύνολα δεδομένων, η μεθοδολογία της δειγματοληψίας παρέχει μια πολυμήχανη και αποτελεσματική λύση στο πρόβλημα της ανισορροπίας με τη χρήση των SVM (Choi (2010)).

Στη συνέχεια δίνουμε μία απλή περιγραφή των δύο βασικών μεθόδων δειγματοληψίας η οποία θα φανεί ιδιαίτερα χρήσιμη για την κατανόηση και εφαρμογή πιο σύνθετων μεθόδων.

Υπερδειγματοληψία (Over-sampling)

Ο μηχανισμός αυτής της μεθόδου είναι η προσθήκη ενός είτε τυχαία επιλεγμένου είτε κατευθυνόμενου δείγματος, έστω του συνόλου E, επιπλέον περιπτώσεων (π.χ. διπλασιασμός των περιπτώσεων) από την κλάση μειοψηφίας (minority class) του αρχικού σύνολου, S. Με τον τρόπο αυτό, ο αριθμός των συνολικών περιπτώσεων (instances) μειοψηφίας αυξάνεται κατά E και ως αποτέλεσμα, η κατανομή των κλάσεων είναι πιο ισορροπημένη. Αυτό παρέχει έναν μηχανισμό για διαφοροποίηση του βαθμού ισορροπίας της κατανομής των κλάσεων σε οποιοδήποτε επιθυμητό επίπεδο ισορροπίας. Η μέθοδος της υπερδειγματοληψίας δεν αυξάνει την λαμβάνουσα πληροφορία, αντιθέτως με τον διπλασιασμό αυξάνεται το βάρος/βαρύτητα (weight) των περιπτώσεων της κλάσης μειοψηφίας. Μία απλή μέθοδος υπερδειγματοληψίας είναι η μέθοδος της τυχαίας υπερδειγματοληψίας. Το βασικό πρόβλημα της υπερδειγματοληψίας είναι ότι εν γένει προκύπτει ένα πρόβλημα υπερπροσαρμογής (overfitting), το οποίο έχει ως επακόλουθο ο κανόνας ταξινόμησης να γίνει πάρα πολύ συγκεκριμένος: ακόμη και αν η ακρίβεια για το σύνολο εκπαίδευσης (train set) είναι υψηλή, η απόδοση ταξινόμησης για

τα νέα σύνολα δεδομένων δοκιμής (test set) κατά πάσα πιθανότητα θα είναι χειρότερη. Με την προσάρτηση των διπλών δεδομένων στο αρχικό σύνολο δεδομένων, ορισμένα από τα δεδομένα που έχουν αντιγραφεί γίνονται πάρα πολύ συγκεκριμένα και οι ταξινομητές θα παράγουν πολλαπλές κλάσεις για τα διπλότυπα (duplicate) δεδομένα (Kubat και Martin, 1997). Υπάρχουν διάφορες μέθοδοι υπερδειγματοληψίας που προσφέρουν μία πιο βελτιωμένη απόδοση από την απλή τυχαία δειγματοληψία όπως είναι ο αλγόριθμος SMOTE, η Borderline SMOTE, η ADASYN, η SPIDER2 και πολλές άλλες με κάποιες από τις οποίες θα ασχοληθούμε και στη συνέχεια.

Υποδειγματοληψία (Undersampling)

Ενώ με τη μέθοδο της υπερδειγματολειψίας προσθέταμε παραδείγματα[4] στο αρχικό σύνολο δεδομένων, στην περίπτωση της υποδειγματοληψίας αφαιρούμε περιπτώσεις από την τάξη πλειοψηφίας, διατηρώντας παράλληλα όλες τις περιπτώσεις της κλάσης μειοψηφίας λόγω της σπάνιας εμφάνισης τους, άρα και της λιγοστής πληροφορίας που παρέχουν. Μια απλή μέθοδος υποδειγματοληψίας της τάξης πλειοψηφίας είναι η τυχαία υποδειγματοληψία (undersampling), μια non-heuristic (μη-ευρετική) μέθοδος που ισορροπεί τις κατανομές ισορροπίας με την επιλογή και την τυχαία αφαίρεση περιπτώσεων πλειοψηφίας. Υπάρχουν πολλές μέθοδοι που προτάθηκαν για την υποδειγματοληψία της κλάσης πλειοψηφίας όπως η Neighborhood Cleaning Rule (NCL), η Condensed nearest Neighbor rule + Tomek links (CNN_TL), η Class Purity Maximization (CPM), η Under-sampling Based on Clustering (SBC), η Tomek Links (TL) και άλλες, κάποιες εκ των οποίων θα συζητήσουμε στη συνέχεια.

2.2.2.2 Προσαρμοστικές μέθοδοι δειγματοληψίας και δημιουργίας συνθετικών δεδομένων (Adaptive sampling methods and synthetic data generation)

Η πρόθεση αυτών των μεθόδων είναι να παρέχουν μια ισορροπημένη κατανομή από τεχνικές υπέρ-δειγματοληψίας ή/και υπο-δειγματοληψίας με σκοπό τη βελτίωση της γενικής ταξινόμησης. Σε ότι αφορά τις συνθετικές μεθόδους δειγματοληψίας, η τεχνική SMOTE δημιουργεί συνθετικά δεδομένα στην κλάση της μειοψηφίας με την επιλογή μερικών από τους κοντινότερους γείτονες των δεδομένων της μειοψηφίας και την παραγωγή συνθετικών δεδομένων μειοψηφίας, μαζί με τις γραμμές μεταξύ των δεδομένων της μειοψηφίας και τους κοντινότερους γείτονες της κλάσης μειοψηφίας. Οι προσαρμοστικές μέθοδοι δειγματοληψίας προτάθηκαν για την παραγωγή συνθετικών δεδομένων. Μία άλλη ιδέα που ανήκει στην συγκεκριμένη κατηγορία μεθόδων είναι αυτή του Borderline-smote που εξηγούμε εκτενέστερα στη συνέχεια (Estabrooks και Japkowiczv (2004)).

[4] Παραδείγματα=περιπτώσεις=πειραματικές εκτελέσεις

2.2.3 Αλγόριθμοι

2.2.3.1 Μηχανές Διανυσματικής Υποστήριξης (Support Vector Machines)

Οι Xiao-yan και Hong-bing (2007) πρότειναν μία τροποποιημένη έκδοση των μηχανών διανυσματικής υποστήριξης, συγκεκριμένα του proximal SVM το λεγόμενο MPSVM που εκχωρεί διαφορετικούς συντελεστές ποινής στα θετικά και στα αρνητικά δείγματα, αντίστοιχα με την προσθήκη ενός νέου διαγώνιου πίνακα στο αρχικό πρόβλημα βελτιστοποίησης. Ο αλγόριθμος RICA χρησιμοποιείται για να επιλέξουμε τις βέλτιστες παραμέτρους για να πάρουμε την υψηλότερη γενικευμένη απόδοση (Xiao-yan και Hong-bing (2007)).

Οι Diao et al. (2012) πρότειναν μία μέθοδο υποδειγματοληψίας για να συμπιέσουν και να ισορροπήσουν το σύνολο εκπαίδευσης που χρησιμοποιούν οι συμβατικοί ταξινομητές SVM «πετώντας» την ελάχιστη δυνατή πληροφορία. Το κλειδί της εποπτικής μεθόδου είναι ότι μπορούν να οικοδομήσουν ένα trade-off μεταξύ του μεγέθους του συνόλου εκπαίδευσης και της απώλειας πληροφοριών με τον προσεκτικό ορισμό ενός μέτρου ομοιότητας μεταξύ των δεδομένων του δείγματος. Τα πειράματά τους έδειξαν ότι ο ταξινομητής SVM παρέχει μια βελτιωμένη απόδοση εφαρμόζοντας την προσέγγιση της συμπίεσης και της εξισορρόπησης των δεδομένων (Diao et al. (2012)).

2.2.3.2 K- nearest neighbor

Αρκετές heuristic μέθοδοι υπο-δειγματοληψίας έχουν προταθεί τα τελευταία χρόνια. Αυτές οι μέθοδοι βασίζονται σε μία από τις δύο διαφορετικές υποθέσεις του μοντέλου θορύβου:

- η μία είναι ότι οι περιπτώσεις κοντά σε ένα όριο απόφασης μεταξύ δύο κλάσεων θεωρούνται θορυβώδεις,
- ενώ η άλλη θεωρεί ότι οι περιπτώσεις που έχουν περισσότερους γείτονες από διαφορετικές κλάσεις είναι θορυβώδεις.

Αφού η τυχαία υπο-δειγματοληψία οδηγεί σε απώλεια, πιθανών, χρήσιμων δεδομένων, κάποιες ευρετικές (heuristic) μέθοδοι υπό-δειγματοληψίας προσπαθούν να αφαιρέσουν περιττές περιπτώσεις που δεν θα επηρεάσουν την ακρίβεια ταξινόμησης του συνόλου εκπαίδευσης.

CNN (Heuristic μέθοδος)

Ο Hart (1968) εισήγαγε ενάν αλγόριθμο συμπύκνωσης του συνόλου εκπαίδευσης, Condensed Nearest Neighbor Rule (CNN), προκειμένου να βρει ένα συνεπές υποσύνολο του δειγματικού συνόλου που μπορεί να ταξινομήσει σωστά όλες τις υπόλοιπες περιπτώσεις του συνόλου εκπαίδευσης. Ο αλγόριθμος χρησιμοποιεί δύο σύνολα, που

ονομάζονται S και T. Αρχικά, το πρώτο δείγμα του συνόλου εκπαίδευσης τοποθετείται στο σύνολο S, ενώ τα υπόλοιπα δείγματα του συνόλου εκπαίδευσης τοποθετούνται στο συνολό T. Στη συνέχεια, εκτελείται ένα πέρασμα μέσα από το T. Κατά τη διάρκεια της σάρωσης, κάθε φορά που ένα σημείο στο T ταξινομείται εσφαλμένα χρησιμοποιώντας το S ως σύνολο εκπαίδευσης, μεταφέρεται από το T στο S. Μετά από την ταξινόμηση, η διαδικασία επαναλαμβάνεται μέχρις ότου να μην υπάρχουν σημεία που να μεταφέρονται από το σύνολο T στο S. Το κίνητρο για αυτό το heuristic είναι ότι τα λανθασμένα ταξινομημένα δεδομένα βρίσκονται κοντά στο όριο απόφασης.

Edited Nearest Neighbor Rule (ENN, Heuristic μέθοδοι)

Ο Wilson (1972) εισήγαγε τη μέθοδο Edited Nearest Neighbor (ENN) για να αφαιρέσει οποιαδήποτε περίπτωση της οποίας η ετικετά της κλάσης είναι διαφορετική από την κλάση τουλάχιστον δύο από τους τρεις πλησιέστερους γείτονες. Η ιδέα πίσω από αυτή την τεχνική είναι να αφαιρεθούν οι περιπτώσεις από την κλάση πλειοψηφίας που είναι κοντά ή γύρω από τη διαχωριστική γραμμή των διαφορετικών κλάσεων βασιζόμενη στην έννοια του πλησιέστερου γείτονα (NN), προκειμένου να αυξηθεί η ακρίβεια ταξινόμησης των περιπτώσεων της κλάσης μειοψηφίας παρά των περιπτώσεων της κλάσης πλειοψηφίας.

Συνδέσεις Tomek (Heuristic μέθοδος)

Με τον ίδιο τρόπο, ο Tomek (1976) πρότεινε μια αποτελεσματική μέθοδο για την εξάλειψη των δεδομένων στις επικαλυπτόμενες περιοχές. Δοθέντων δύο περιπτώσεων x και y που έχουν μια διαφορετική ετικέτα κλάσης και χωρίζονται από μια απόσταση $d\,(x,y)$, το ζεύγος (x,y) ονομάζεται *σύνδεση Tomek* αν δεν υπάρχει παράδειγμα z τέτοιο ώστε $d\,(x,z) < d\,(x,y)$ ή $d\,(y,z) > d\,(x,y)$. Περιπτώσεις ή παραδείγματα που συμμετέχουν στις συνδέσεις Tomek θεωρούνται είτε οριακά είτε θορυβώδη.

OSS (Heuristic μέθοδος)

Οι Kubat και Martin (1997) πρότειναν την μονόπλευρη δειγματοληψία (one-sided sampling OSS) για την ανίχνευση λιγότερων περιπτώσεων για εκμάθηση. Η τεχνική αυτή έχει ως στόχο να κρατήσει όλες τις περιπτώσεις της κλάσης μειοψηφίας, δεδομένου ότι είναι σπάνιες, (έστω και αν ορισμένες από αυτές μπορεί να είναι θορυβώδεις) και αντ' αυτού «κλαδεύει» μόνο τις περιπτώσεις της κλάσης πλειοψηφίας. Αρχικά ξεκινάει με ένα υποσύνολο (C) του συνόλου εκπαίδευσης (S) που περιέχει όλες τις περιπτώσεις της κλάσης μειοψηφίας, $C \subseteq S$, και χρησιμοποιώντας τον κανόνα του 1-πλησιέστερου γείτονα (1-Nearest Neighbor), χρησιμοποιώντας περιπτώσεις του συνόλου C, ταξινομεί τις περιπτώσεις του συνόλου S. Στη συνέχεια, όλες οι εσφαλμένα ταξινομημένες περιπτώσεις μετακινούνται στο σύνολο C και τότε όλες οι περιπτώσεις της κλάσης πλειοψηφίας που συμμετέχουν στις συνδέσεις Tomek από το C αφαιρούνται επειδή πιστεύεται ότι είναι οριακές ή/και θορυβώδης.

Class Purity Maximization (CPM)

Ο αλγόριθμος CPM ακολουθεί μια επαναληπτική διαδικασία. Αρχικά, καθορίζει δύο δείγματα κεντρικών σημείων. Ένα από αυτά αντιπροσωπεύει την κλάση μειοψηφίας, ενώ το άλλο αντιπροσωπεύει την κλάση πλειοψηφίας. Στη συνέχεια, χρησιμοποιεί αυτά τα δύο σημεία για να χωρίσει αυτό το σύνολο δεδομένων σε δύο συστάδες (clusters) C_1 και C_2. Τότε, υπολογίζει την «ακαθαρσία» της κάθε συστάδας. Τέλος, κάνει μία σύγκριση μεταξύ της ακαθαρσίας του γονέα και της ακαθαρσίας της προκύπτουσας συστάδας. Αυτή η διαδικασία επαναλαμβάνεται μέχρι να πραγματοποιηθεί η συνθήκη τερματισμού. Η συνθήκη τερματισμού επιτυγχάνεται όταν μία από τις συστάδες έχει μικρότερη ακαθαρσία από το γονέα της είτε όταν φτάσουμε σε μονοσύνολο (Yoon & Kwek, 2005).

Condensed nearest Neighbor rule + Tomek links (CNN_TL)

Η πρώτη μέθοδος CNN χρησιμοποιείται για να μειώσει το υποσύνολο που αποτελεί την κλάση πλειοψηφίας με την αφαίρεση δειγμάτων που βρίσκονται μακριά από το σύνορο απόφασης. Στη συνέχεια, οι συνδέσεις Tomek εφαρμόζονται για να αφαιρέσουν τα θορυβώδη δείγματα και τα δείγματα πλειοψηφίας που βρίσκονται κοντά στο σύνορο απόφασης (Fernándeza et al.(2008)).

Σχετικά πρόσφατα έχουν παρουσιαστεί δυο εργασίες, τις οποίες παραθέτουμε στη συνέχεια όπου παρουσιάζονται αποδοτικοί αλγόριθμοι με τη χρήση του KNN.

CCW (class confidence weights)

Οι Liu και Chawla παρουσίασαν μια νέα μέθοδο KNN όπου προτείνεται η στρατηγική της στάθμισης για την αντιμετώπιση του προβλήματος της ανισορροπίας. Συγκεκριμένα πρότειναν τη μέθοδο CCW (class confidence weights) που χρησιμοποιεί την πιθανότητα που αποδίδουν οι τιμές των χαρακτηριστικών (επεξηγηματικών μεταβλητών) στις ετικέτες των κλάσεων για να δώσουν βάρος στα δείγματα του KNN. Το σημαντικό πλεονέκτημα της CCW είναι ότι είναι σε θέση να διορθώσει την εγγενή μεροληψία (bias) στην κλάση πλειοψηφίας έτσι ώστε οι αλγόριθμοι KNN να είναι προσιτοί σε διάφορες διαστάσεις. Τόσο η θεωρητική μελέτη όσο και τα ολοκληρωμένα πειράματά τους, επιβεβαιώνουν τους ισχυρισμούς τους (Liu και Chawla (2011)).

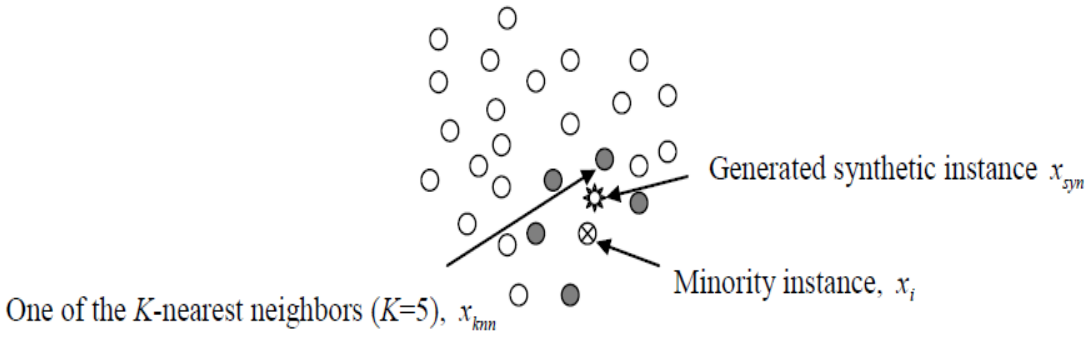
K-means cluster

Ο Yang Yong πρότεινε μία μέθοδο δειγματοληψίας με βάση τις συστάδες του αλγορίθμου K-means (K-means cluster) βασιζόμενος παράλληλα και στο γενετικό αλγόριθμο. Πιο συγκεκριμένα, χρησιμοποίησε τον αλγόριθμο K-means για να ομαδοποιήσει την κλάση μειοψηφίας του δείγματος, και σε κάθε ομάδα (cluster) χρησιμοποίησε το γενετικό αλγόριθμο για να αποκτήσει το νέο δείγμα και να συνεχίσει με την έγκυρη επικύρωση (valid authetication). Στο τέλος, μέσα από τη χρήση του KNN και των SVM απέδειξε την ισχύ της τεχνικής του με πειράματα προσομοίωσης (Yang

Yong(2012)).

2.2.3.3 Synthetic Minority Over-sampling Technique (SMOTE)

Για να αποφευχθεί το πρόβλημα της υπερπροσαρμογής στην περίπτωση της εφαρμογής της μεθόδου της υπερ-δειγματοληψίας, οι Chawla et αl. (2002) πρότειναν μία heuristic μέθοδο υπερδειγματολειψίας, που ονομάζεται Synthetic Minority Over-sampling Technique (SMOTE) η οποία φαίνεται να λειτουργεί καλά σε διάφορες εφαρμογές. Η SMOTE θεωρείται ότι είναι μία από τις τελευταίες και αποδοτικότερες προσεγγίσεις για τη μάθηση με μη ισορροπημένα δεδομένα. Αυτή η μέθοδος παράγει συνθετικά (synthetic) δεδομένα βασιζόμενη στο χώρο των χαρακτηριστικών[5] μεταξύ των υπαρχόντων minority περιπτώσεων θεωρώντας τους k-πλησιέστερους γείτονες για κάθε περίπτωση (πειραματική εκτέλεση) της κλάσης μειοψηφίας (minority class). Προκειμένου να δημιουργήσει ένα συνθετικό παράδειγμα, η μέθοδος SMOTE βρίσκει τους k-πλησιέστερους γείτονες για κάθε περίπτωση της κλάσης μειοψηφίας, επιλέγει τυχαία έναν από αυτούς, και στη συνέχεια πολλαπλασιάζει την αντίστοιχη διαφορά του διανύσματος των χαρακτηριστικών (feature vector) με ένα τυχαίο αριθμό μεταξύ 0 και 1 για να παραχθεί μία νέα περίπτωση της κλάσης μειοψηφίας στη γειτονιά. Στο Σχήμα 2.2 παρουσιάζουμε ένα παράδειγμα της διαδικασίας SMOTE.



$$x_{syn} = x_i + (x_{knn} - x_i) \times \alpha$$

α is a random number between 0 and 1

Σχήμα 2.2: Συνθετικά παραδείγματα με τον αλγόριθμο υπερδειγματοληψίας, SMOTE

Με αυτή τη μέθοδο υπερδειγματοληψίας που δημιουργεί συνθετικά δεδομένα, αποφεύγεται το πρόβλημα της υπερπροσαρμογής. Επιπρόσθετα, η μέθοδος οδηγεί τα όρια απόφασης για την κλάση της μειοψηφίας να κινηθούν προς την κατεύθυνση της κλάσης πλειοψηφίας.

[5] Χώρος των επεξηγηματικών μεταβλητών (feature space)

Borderline SMOTE

Ως μια παραλλαγή του SMOTE, οι Han et al. (2005) εισήγαγαν το Borderline SMOTE το οποίο δημιουργεί συνθετικές περιπτώσεις της κλάσης μειοψηφίας μόνο κοντά στο όριο απόφασης δεδομένου ότι οι περιπτώσεις εκείνες είναι πιο πιθανό να ταξινομηθούν εσφαλμένα. Τα αποτελέσματα της προσέγγισης αυτής ήταν καλύτερα σε σύγκριση με το κλασικό SMOTE και την τυχαία υπερδειγματοληψία χρησιμοποιώντας για ταξινομητές τα δέντρα απόφασης (Decision Tree).

2.2.3.4 ADASYN

Οι He et al. (2008) εισήγαγαν μια μέθοδο δειγματοληψίας που παράγει συνθετικές περιπτώσεις, και ονομάζεται Adaptive Synthetic Sampling (ADASYN). Η μέθοδος ADASYN χρησιμοποιεί την κατανομή πιθανότητας των περιπτώσεων μειοψηφίας ως κριτήριο για να αποφασίσει αυτόματα τον αριθμό των συνθετικών δειγμάτων που δημιουργούνται για κάθε περίπτωση της κλάσης μειοψηφίας. Η μέθοδος ADASYN δημιουργεί μία νέα περίπτωση με τον υπολογισμό του ποσοστού μεταξύ των περιπτώσεων της κλάσης μειοψηφίας και της κλάσης πλειοψηφίας στους k-πλησιέστερους γείτονες της κάθε περίπτωσης μειοψηφίας. Ως αποτέλεσμα, παράγονται περισσότερες συνθετικές περιπτώσεις, για την κλάση μειοψηφίας, που είναι πιο δύσκολο να «μάθουν» συγκριτικά με τις περιπτώσεις της κλάσης πλειοψηφίας που είναι πιο εύκολο να «μάθουν». Αυτή η προσέγγιση βελτίωσε τη μάθηση με τα μη ισορροπημένα σύνολα δεδομένων βασιζόμενη στις κατανομές των δεδομένων, μειώνοντας τη μεροληψία της κατανομής των κλάσεων και προσαρμόζοντας το όριο απόφασης μετατοπίζοντάς το, έτσι ώστε να δοθεί μεγαλύτερη προσοχή στις περιπτώσεις που είναι πιο δύσκολο να «μάθουν» (minority class).

Κατά συνέπεια, η προσέγγιση ADASYN λαμβάνει την καλύτερη δυνατή μάθηση σε σχέση με τις κατανομές των δεδομένων με δύο τρόπους: (1) τη μείωση της μεροληψίας που εισήχθη από την ανισορροπία των κλάσεων, και (2) προσαρμοστικά μετατοπίζοντας το όριο απόφασης της ταξινόμησης προς τα δύσκολα παραδείγματα. Η ανάλυση με τη χρήση προσομοιώσεων σε διάφορα σύνολα δεδομένων δείχνουν την αποτελεσματικότητα αυτής της μέθοδο με βάση πέντε μέτρα αξιολόγησης.

2.2.3.5 Random Forest

Οι Yao et al. (2013) πρότειναν τον αλγόριθμο Random Forest βασιζόμενοι στην δειγματοληψία με επανάθεση. Συγκεκριμένα εξήγαγαν τυχαία πολλαπλά υποσύνολα παραδειγμάτων με επανάθεση από την κλάση πλειοψηφίας, και ο αριθμός των εξαγόμενων υποσυνόλων παραδειγμάτων είναι ο ίδιος με τον αριθμό των παραδειγμάτων που ανήκουν στην κλάση μειοψηφίας. Στη συνέχεια, κατασκεύασαν τα πολλαπλά νέα σύνολα εκπαίδευσης συνδυάζοντας το κάθε εξαγόμενο υποσύνολο παραδειγμάτων της κλάσης πλειοψηφίας και της κλάσης μειοψηφίας αντίστοιχα. Στη συνέχεια πάνω στα νέα σύνολα δεδομένων εκπαίδευσης, εκπαιδεύτηκαν πολλαπλοί ταξινομητές Random Forest. Τα αποτελέσματα που εξήγαγαν για πέντε σύνολα δεδομένων από την βάση UCI και από ένα πραγματικό ιατρικό σύνολο δεδομένων έδειξαν ότι αυτή η μέθοδος θα μπορούσε να χειριστεί αποτελεσματικά το πρόβλημα της ανισορροπίας μεταξύ των κλάσεων αλλά και ότι ο βελτιωμένος αλγόριθμος Random Forest ξεπερνάει όχι μόνο τον αρχικό αλγόριθμο Random Forest αλλά και άλλες μεθόδους που έχουν προταθεί κατά καιρούς στη βιβλιογραφία.

2.3 Προσέγγιση σε επίπεδο ταξινομητών (Classifier level approaches)

Η εξισορρόπηση της κατανομής των δεδομένων είτε μέσω υπερ-δειγματοληψίας είτε μέσω υπο-δειγματοληψίας έχει κάποια επιτυχία στο χειρισμό τέτοιου είδους προβλημάτων, αλλά η εφαρμογή αυτών των μεθόδων είναι συνήθως υπολογιστικά ακριβή. Επίσης, η αλλαγή της κατανομής των κλάσεων στο επίπεδο των δεδομένων δεν οδηγεί πάντα σε καλύτερη απόδοση ταξινόμησης. Ένας ταξινομητής δεν είναι επηρεάζεται πάντα από την κατανομή των κλάσεων.

Οι Drummond και Holte (2003) παρατήρησαν ότι η υπερ-δειγματοληψία δεν παρέχει αποτελεσματική βελτίωση της απόδοσης ή ότι δεν υπάρχει κάποια αλλαγή στην ταξινόμηση των δεδομένων. Αντιθέτως, η υπερδειγματοληψία «κλαδεύει» λιγότερο από την υπο-δειγματοληψία χρησιμοποιώντας τις προκαθορισμένες (default) παραμέτρους του αλγορίθμου C4.5. Μια τροποποίηση των παραμέτρων του C4.5 βελτιώνει την απόδοση της ταξινόμησης και παράλληλα αποφεύγεται το πρόβλημα της υπερ-προσαρμογής κατά τη διάρκεια της υπερ-δειγματοληψίας. Έτσι, ενώ οι μέθοδοι δειγματοληψίας έχουν προσπαθήσει να εξισορροπήσουν την κατανομή των κλάσεων εξετάζοντας τις αναλογίες μεταξύ των παραδειγμάτων της κάθε κλάσης/κατηγορίας στην αρχική κατανομή των δεδομένων, άλλες προσεγγίσεις έχουν εισήχθη για τη μάθηση των μη ισορροπημένων δεδομένων.

Η προσπάθεια να προσαρμόσουμε τους υπάρχοντες αλγόριθμους για το πρόβλημα των μη ισορροπημένων συνόλων δεδομένων και να τους κάνουμε μεροληπτικούς ευνοώντας την κλάση μειοψηφίας είναι γνωστή ως προσέγγιση στο επίπεδο των ταξινομητών

(Classifier level approaches). Εδώ, χρειάζεται λίγο περισσότερο σε βάθος γνώση σχετικά με τη φύση των παραγόντων που χρησιμοποιούνται στις προβλέψεις και που προκαλούν την αποτυχία της ταξινόμησης στην αναγνώριση της κλάσης μειοψηφίας. Μια πιθανότητα είναι να εκτελέσουμε one-class classification, το οποίο μπορεί να μάθει τις έννοιες της τάξης μειοψηφίας λαμβάνοντας ως ακραίες τιμές τα αντικείμενα της κλάσης πλειοψηφίας.

2.3.1 Μέθοδοι

2.3.1.1 Μάθηση ευαίσθητου κόστους (Cost-sensitive learning)

Η μέθοδος του Cost-sensitive learning συνδέεται με τα μη σωστά ταξινομημένα πρότυπα. Ο πίνακας του κόστους χρησιμοποιείται για την αριθμητική απεικόνιση των συνεπειών της ταξινόμησης των παραδειγμάτων από τη μια κλάση στην άλλη (Maheshwari et al.(2011)). Κατά συνέπεια δεν εκχωρείται κάποιο κόστος για την ορθή ταξινόμηση κάποιας από τις κλάσεις και το κόστος των λανθασμένα ταξινομημένων δειγμάτων της κλάσης μειοψηφίας είναι υψηλότερο από ότι το κόστος των δειγμάτων πλειοψηφίας, για παράδειγμα, $C(Majority, Minority) > C(Minority, Majority)$. Στόχος της μεθόδου του Cost-sensitive learning είναι να ελαχιστοποιηθεί το συνολικό κόστος για το σύνολο δεδομένων εκπαίδευσης. Ο Elkan (2001) παρουσίασε ένα θεώρημα που δείχνει πώς να αλλάξουμε το ποσοστό των θετικών και αρνητικών δειγμάτων, έτσι ώστε να κάνουμε τις βέλτιστες ταξινομήσεις, ευαίσθητου-κόστους, για το μαθησιακό πρόβλημα. Ο Domingos (1999) πρότεινε μια πιο γενική μέθοδο για να κάνει το μαθησιακό σύστημα, ένα σύστημα ευαίσθητου κόστους. Η προσέγγιση του ευαίσθητου κόστους (cost sensitive method) χρησιμοποιεί έναν πίνακα κόστους για την ποινικοποίηση εσφαλμένα ταξινομημένων περιπτώσεων, όπως φαίνεται στον Πίνακας 2.1.

Πίνακας 2.1 Πίνακας κόστους

Πραγματική	Προβλεπόμενη	
	i Κλάση	j Κλάση
i Κλάση	0	c_{ij}
j Κλάση	c_{ji}	0

Συνήθως, δεν υπάρχουν κόστη που να εφαρμόζονται στις περιπτώσεις που έχουμε σωστή ταξινόμηση, και το κόστος των μη σωστά ταξινομημένων περιπτώσεων της κλάσης μειοψηφίας είναι υψηλότερο από εκείνο των περιπτώσεων της κλάσης πλειοψηφίας.

Ο στόχος αυτής της στρατηγικής είναι να ελαχιστοποιηθεί το κόστος της εσφαλμένης ταξινόμησης. Σε ορισμένες εφαρμογές, οι cost sensitive τεχνικές έχουν καλύτερες επιδόσεις από τις μεθόδους δειγματοληψίας (McCarthy et al., 2005 και Liu et al., 2006a).

MetaCost

Η MetaCost (Domingos, 1999) είναι μια άλλη μέθοδος που σχετίζεται με την cost sensitive μάθηση. Αυτή η μέθοδος εκτιμάει τις πιθανότητες των κλάσεων με χρήση του Bagging και στη συνέχεια εκ νέου τοποθετεί ετικέτες στις περιπτώσεις της εκπαίδευσης με τις ελάχιστες αναμενόμενες κλάσεις τους, και στο τέλος, εκπαιδεύει ξανά ένα μοντέλο χρησιμοποιώντας το τροποποιημένο σύνολο εκπαίδευσης.

Cost sensitive και Naïve Bayes ή NNs

Μερικοί ταξινομητές όπως ο Naïve Bayes ταξινομητής ή κάποια Νευρωνικά Δίκτυα χρησιμοποιούν ένα σκορ για να δείξουν το βαθμό στον οποίο μια περίπτωση ανήκει σε μια κατηγορία. Αυτό το είδος της κατάταξης μπορεί να χρησιμοποιηθεί σε εναλλακτικούς ταξινομητές με την αλλαγή του κατωφλίου (threshold) για ένα παράδειγμα που ανήκει σε μια κλάση (Weiss, 2004).

Cost sensitive και discrimination

Για την δημιουργία μεροληψίας στη διαδικασία της διακριτικοποίησης (discrimination), οι Barandela et al. (2003) πρότειναν μια σταθμισμένη συνάρτηση απόστασης στην ταξινόμηση αντί να μεταβάλλουν τις κατανομές των κλάσεων με βάση τη μέθοδο ταξινόμησης του πλησιέστερου γείτονα (NN). Υποθέτοντας ότι $d_e(\cdot)$ είναι η Ευκλείδεια μετρική, x_{new} ένα νέο παράδειγμα που θέλουμε να ταξινομήσουμε, x_0 ένα δείγμα εκπαίδευσης από την κλάση i, n_i ο αριθμός των περιπτώσεων της κλάσης i και m η διάσταση της μεταβλητής εισόδου, μια σταθμισμένη συνάρτηση της απόστασης, $d_w(\cdot)$ ορίζεται ως εξής:

$$d_w(x_{new}, x_0) = \left(\frac{n_i}{n}\right)^{1/m} \times d_e(x_{new}, x_0)$$

Αυτό θα μπορούσε να αποδώσει μεγαλύτερους συντελεστές βαρύτητας για τις περιπτώσεις της κλάσης πλειοψηφίας παρά για τις περιπτώσεις της κλάσης μειοψηφίας. Κατά συνέπεια, παράγονται μικρότερες αποστάσεις στις περιπτώσεις της κλάσης μειοψηφίας παρά σε αυτές της κλάσης πλειοψηφίας. Ως αποτέλεσμα, οι γείτονες των νέων περιπτώσεων βρέθηκαν ανάμεσα στις περιπτώσεις της κλάσης μειοψηφίας, αυξάνοντας την τιμή του γεωμετρικού μέσου (gmean).

Cost sensitive και SVM

Στην περίπτωση του αλγόριθμου ταξινόμησης SVM, αυτή η μεροληπτική προσέγγιση ωθεί το υπερεπίπεδο πιο μακριά από την κλάση μειοψηφίας (θετική κλάση) για τα μη ισορροπημένα σύνολα δεδομένων. Οι Wu και Chang (2003) πρότειναν έναν μεροληπτικό αλγόριθμο για να αλλάξουν τη συνάρτηση του πυρήνα. Οι μεροληπτικοί αλγόριθμοι ταξινόμησης SVM χρησιμοποιούν μεγαλύτερες σταθερές ποινής που συνδέονται με την κλάση μειοψηφίας κάνοντας τα σφάλματα της ταξινόμησης για τις περιπτώσεις της κλάσης μειοψηφίας πολύ «ακριβότερα» από τα λάθη των περιπτώσεων της κλάσης πλειοψηφίας (Veropoulos et al., 1999) (θα επεκτεθούμε σε επόμενο κεφάλαιο).

BMPM

Οι Huang στο el. (2004) πρότειναν τη μεροληπτική μέθοδο Biased Minimax Probability Machine (BMPM) για την επίλυση της μάθησης για μη ισορροπημένα σύνολα δεδομένων. Λαμβάνοντας υπόψη τους πίνακες των μέσων τιμων και των συνδιακυμάνσεων των κλάσεων πλειοψηφίας και μειοψηφίας, ο BMPM διαμορφώνει ένα πρόβλημα βελτιστοποίησης με στόχο την εύρεση του υπερεπιπέδου απόφασης ρυθμίζοντας το κατώτερο όριο της ακρίβειας για την ταξινόμηση των μελλοντικών δεδομένων. Για παράδειγμα, αν η αντικειμενική συνάρτηση είναι η μεγιστοποίηση της ακρίβειας της ταξινόμησης για την κλάση της μειοψηφίας, η βελτιστοποίηση προσπαθεί να τη μεγιστοποιήσει, ορίζοντας ένα κατώτερο όριο της ακρίβειας ταξινόμησης για τις δύο κλάσεις. Η επίτευξη της χειρότερης ακρίβειας για την κλάση της μειοψηφίας μπορεί να αποφευχθεί, διατηρώντας παράλληλα το αποδεκτό επίπεδο ακρίβειας της κλάσης πλειοψηφίας στη μάθηση με μη ισορροπημένα δεδομένα.

One-class learning

Η μάθηση μίας κλάσης (one-class learning) είναι μια εναλλακτική λύση της διακριτικοποίησης, εφόσον το μοντέλο δημιουργείται μόνο βάση των περιπτώσεων της κλάσης στόχου. Η βασική ιδέα είναι ότι τα σύνορα μεταξύ των δύο κλάσεων υπολογίζονται από τα δεδομένα μιας κλάσης (κλάση στόχος), έτσι ώστε αυτή η προσέγγιση δεν είναι ευαίσθητη στην κατανομή της κλάσης στο σύνολο εκπαίδευσης. Ένα όριο γύρω από την κλάση στόχο ορίζεται κατά τέτοιο τρόπο ώστε τα περισσότερα από τα αντικείμενα-στόχους περιλαμβάνονται και ταυτόχρονα ελαχιστοποιείται η πιθανότητα της αποδοχής ακραίων αντικειμένων. Για παράδειγμα, οι Kubat et αl. (1998) εισήγαγαν τον αλγόριθμο SHRINK μετά από αυτή τη γενική αρχή και την εφάρμοσαν στην ανίχνευση σπάνιων πετρελαιοκηλίδων από δορυφορικές εικόνες ραντάρ. Ο στόχος ήταν να βρεθεί ο κανόνας ταξινόμησης που αναγνωρίζει καλύτερα τα θετικά παραδείγματα (πετρελαιοκηλίδες) χρησιμοποιώντας ως μέτρο απόδοσης το Γεωμετρικό Μέσο. Υποθέτοντας ότι οι αρνητικές (majority) περιπτώσεις ξεπερνούν τις θετικές (minority) περιπτώσεις, ο αλγόριθμος επισημαίνει τις μικτές περιοχές ως θετικές (minority). Αυτό αλλάζει την εστίαση της μάθησης: αναζήτηση για την *καλύτερη θετική περιοχή*, αυτή με τη μέγιστη αναλογία θετικών -αρνητικών. Οι Raskutti και Kowalczyk

(2004) μελέτησαν την εκμάθηση της μιας κλάσης (one-class learning) με εξαιρετικά μη ισορροπημένα σύνολα δεδομένων με τη χρήση ενός SVM ταξινομητή. Έδειξαν ότι η εκμάθηση μιας κλάσης είναι χρήσιμη για τα εξαιρετικά μη ισορροπημένα σύνολα δεδομένων με υψηλών διαστάσεων θορυβώδη χώρο χαρακτηριστικών.

2.3.1.2 Boosting Method

Το Boosting είναι μια τεχνική για την βελτίωση της απόδοσης των αδύναμων ταξινομητών. Το AdaBoost (Freund & Schapire, 1997) είναι ο πιο γνωστός αλγόριθμος boosting. Σε κάθε επανάληψη, τα βάρη τροποποιούνται με στόχο την σωστή ταξινόμηση των παραδειγμάτων στην επόμενη επανάληψη. Στο τέλος, όλα τα προσαρμοσμένα μοντέλα συμβάλλουν σε μια σταθμισμένη ψήφο για ταξινομήσουν τα μη επισημασμένα παραδείγματα. Αυτή η μέθοδος είναι πιο χρήσιμη στο πρόβλημα ανισορροπίας των κλάσεων, επειδή κυρίως τα παραδείγματα της κλάσης μειοψηφίας αναμένεται να ταξινομηθούν λανθασμένα και, ως εκ τούτου δίνονται μεγαλύτερα βάρη στις επόμενες επαναλήψεις.

Adacost

Βασιζόμενοι στον κανόνα ενημέρωσης βάρους του AdaBoost (Freund & Schapire, 1997) για τις λανθασμένα ταξινομημένες περιπτώσεις στην επαναληπτική μάθηση, οι Fan et al. (1999) πρότειναν μια διακριτική (discriminant) μέθοδο ενημέρωσης βάρους για τις λανθασμένες περιπτώσεις για τα μη ισορροπημένα σύνολα δεδομένων, η οποία ονομάζεται Adacost. Η προσέγγισή τους αφορά στην εκχώρηση μεγαλύτερων βαρών για τις εσφαλμένα ταξινομημένες περιπτώσεις που ανήκουν στην κλάση μειοψηφίας απ' ότι σε εκείνες που ανήκουν στην κλάση πλειοψηφίας και ως εκ τούτου, ο Adacost φαίνεται εμπειρικά ότι εκτελείται καλύτερα στη μείωση των συσσωρευτικών κόστων εσφαλμένης ταξινόμησης απ' ότι ο AdaBoost.

2.3.1.3 Μέθοδοι ενεργητικής μάθησης (Active learning methods)

Οι παραδοσιακές μέθοδοι ενεργητικής μάθησης χρησιμοποιούνται για την επίλυση των μη ισορροπημένων δεδομένων εκπαίδευσης. Πρόσφατα, έχουν προταθεί διάφορες προσεγγίσεις για την ενεργό μάθηση από τα μη ισορροπημένα σύνολα δεδομένων. Ως παράδειγμα μπορούμε να αναφέρουμε τη μέθοδο της ενεργής μάθησης που βασίζεται στις SVM, όπου επιλέγονται αποτελεσματικά οι περιπτώσεις από ένα τυχαίο σύνολο δεδομένων εκπαίδευσης, έτσι ώστε να μειωθεί σημαντικά το υπολογιστικό κόστος, ιδιαίτερα όταν πρόκειται για μεγάλα μη ισορροπημένα σύνολα δεδομένων (Ertekin et al. (2007)).

2.3.2 Αλγόριθμοι

2.3.2.1 Class Confidence Proportion Decision Tree (CCPDT)

Οι Liu και Chawla πρότειναν ένα νέο αλγόριθμο δέντρων απόφασης, τον Class Confidence Proportion Decision Tree (CCPDT), ο οποίος είναι εύρωστος και όχι ευαίσθητος στο μέγεθος των κλάσεων και παράγει κανόνες που είναι στατιστικά σημαντικοί. Για να δημιουργήσουν τέτοιους κανόνες που είναι στατιστικά σημαντικοί οι συγγραφείς σχεδίασαν μια καινοτόμα και αποδοτική (top-down and bottom-up approach) προσέγγιση, η οποία χρησιμοποιεί το τεστ του Fisher για να κλαδέψει τα κλαδιά του δέντρου που δεν είναι στατιστικά σημαντικά. Μαζί αυτές οι δύο αλλαγές διαφοροποιούν έναν ταξινομητή που εκτελείται στατιστικά καλύτερα όχι μόνο από τα παραδοσιακά δέντρα απόφασης, αλλά και τα δέντρα που μαθαίνουν από τα δεδομένα και που έχουν εξισορροπηθεί από γνωστές τεχνικές δειγματοληψίας. Οι ισχυρισμοί τους επιβεβαιώνονται από εκτεταμένα πειράματα και συγκρίσεις μεταξύ των αλγορίθμων C4.5, CART, HDDT και SPARCCC (Liu et al. (2010)).

2.3.2.2 Semi supervised Clustering (Ημι εποπτευόμενο Clustering)

Οι Mingwei Leng, et al., πρότειναν έναν ενεργό ημί εποπτευόμενο αλγόριθμο ομαδοποίησης που χρησιμοποιεί μια ενεργητική μέθοδο για την επιλογή των δεδομένων για να ελαχιστοποιηθεί η ποσότητα του επισημασμένων πληροφοριών, και εκτελεί ένα πολλαπλό thresold (multithreshold) για μεγέθυνση των επισημασμένων δεδομένων σε σύνολα πολλαπλών πυκνοτήτων (multidensity) και μη ισορροπημένα σύνολα δεδομένων. Τρία τυπικά σύνολα δεδομένων και ένα συνθετικό σύνολο δεδομένων χρησιμοποιούνται για να επιδείξουν τον αλγόριθμο, και τα δοκιμαστικά αποτελέσματα δείχνουν ότι ο αλγόριθμος της ημι-εποπτευόμενης ομαδοποίησης έχει μια υψηλότερη ακρίβεια και μια πιο σταθερή απόδοση σε σύγκριση με άλλους ημιεποπτευόμενους αλγορίθμους ομαδοποίησης, ιδιαίτερα όταν τα σύνολα δεδομένων είναι με πολλαπλή πυκνότητα (multidensity) και μη ισορροπημένα (Leng et al. (2013))

2.4 Cost sensitive approaches

Η προσέγγιση αυτή μπορεί να χρησιμοποιήσει και την προσέγγιση σε επίπεδο δεδομένων και τροποποιήσεις των αλγορίθμων μάθησης. Το υψηλότερο κόστος μη εσφαλμένης ταξινόμησης εκχωρείται για τα παραδείγματα της κλάσης μειοψηφίας και η ταξινόμηση εκτελείται έτσι ώστε να μειωθεί το συνολικό κόστος της μάθησης. Το κόστος συχνά προσδιορίζονται στη μορφή των πινάκων του κόστους. Η έλλειψη γνώσεως σχετικά με το πώς να ρυθμίσουμε τις πραγματικές τιμές στον πίνακα του κόστους είναι το βασικό μειονέκτημα των μεθόδων του ευαίσθητου κόστους, δεδομένου

ότι στις περισσότερες περιπτώσεις αυτό δεν είναι γνωστό από τα δεδομένα ούτε δίνεται από κάποιον εμπειρογνώμονα (Elkan (2001)).

2.4.1 Μέθοδοι

2.4.1.1 Cost-sensitive methods

Οι Cost-sensitive μέθοδοι μάθησης χρησιμοποιούν τον πίνακα κόστους για να εξετάσουν τα κόστοι που συνδέονται με μη ταξινομημένα παραδείγματα. Το Cost-sensitive νευρωνικό δίκτυο με την τεχνική του κινούμενου-κατωφλίου (Moving-threshold) προτάθηκε για να ρυθμιστεί το κατώφλι προς την ανέξοδη κλάση, έτσι ώστε τα δείγματα με υψηλό κόστος να είναι απίθανο να ταξινομηθούν εσφαλμένα. Τρεις cost sensitive boosting μέθοδοι, AdaC1, AdaC2, και AdaC3 προτάθηκαν όπου χρησιμοποιήθηκαν κόστη για να δώσουν κάποιο βάρος στην στρατηγική του αλγορίθμου boosting (Sun et al. (2007)).

2.4.1.2 Cost sensitive Ensemble Method

Οι Zhang και Wang (2013) πρότειναν μία Cost sensitive ensemble μέθοδο που βασίζεται στις SVM με ευαίσθητο κόστος και στον αλγόριθμο QBC (query-by-committee) με σκοπό την επίλυση του προβλήματος των μη ισορροπημένων κλάσεων μεταξύ των δεδομένων. Η μέθοδος αυτή πρώτα χωρίζει την κλάση πλειοψηφίας σε πολλά επιμέρους σύνολα δεδομένων, σύμφωνα με την αναλογία των μη ισορροπημένων δειγμάτων και εκπαιδεύει τους επιμέρους ταξινομητές χρησιμοποιώντας τη μέθοδο AdaBoost. Τότε, η μέθοδος παράγει υποψήφια δείγματα για εκπαίδευση με τη μέθοδο μάθησης QBC και χρησιμοποιεί τη μέθοδο SVM με ευαίσθητο κόστος για την εκμάθηση των δειγμάτων εκπαίδευσης.

2.4.1.3 Split and Boost active learning methods

Οι Yong Zhang, et al., για την επίλυση του προβλήματος των μη ισορροπημένων δεδομένων σε προβλήματα ταξινόμησης πρότειναν μια Split and Boost active learning μέθοδο (Split Boost), με βάση το SVM με ευαίσθητο κόστος (cost sensitive SVM). Η μέθοδος Split Boost χωρίζει πρώτα το σύνολο δεδομένων της κλάσης πλειοψηφίας σε πολλά υποσύνολα δεδομένων, σύμφωνα με το ποσοστό των μη ισορροπημένων δειγμάτων, και καθοδηγεί τους υπο-ταξινομητές με τη μέθοδο AdaBoost. Στη συνέχεια, το Split Boost παράγει υποψήφια δείγματα εκπαίδευσης με την ενεργή μέθοδο μάθησης QBC και χρησιμοποιεί cost sensitive SVM για την εκμάθηση των δειγμάτων εκπαίδευσης. Χρησιμοποιώντας 6 μη ισορροπημένα σύνολα δεδομένων, τα πειραματικά αποτελέσματα έδειξαν ότι η μέθοδος έχει υψηλότερα AUC, F-μέτρο, και G-μέσο από πολλές μεθόδους εκμάθησης μη ισορροπημένων κλάσεων.

2.4.2 Αλγόριθμοι

2.4.2.1 Cost sensitive boosting algorithm

Υπάρχουν τρεις τρόποι για να εισάγουμε τα στοιχεία του κόστους στον τύπο ενημέρωσης βάρους του αλγορίθμου AdaBoost: στο εσωτερικό του εκθέτη, εκτός του εκθέτη, και τόσο εντός όσο και εκτός του εκθέτη. Έχουμε λοιπόν τρεις τροποποιήσεις της εξίσωσης:

Τροποποίηση I:

$$D^{t+1}(i) = \frac{D^t(i)\exp(-a_t C_i h_t(x_i) y_i)}{Z_t}$$

Τροποποίηση II:

$$D^{t+1}(i) = \frac{C_i D^t(i)\exp(-a_t h_t(x_i) y_i)}{Z_t}$$

Τροποποίηση III:

$$D^{t+1}(i) = \frac{C_i D^t(i)\exp(-a_t C_i h_t(x_i) y_i)}{Z_t}$$

Κάθε τροποποίηση μπορεί να θεωρηθεί ως ένας νέος boosting αλγόριθμος που συμβολίζεται ως AdaC1, AdaC2 και AdaC3, αντίστοιχα. Δεδομένου ότι αυτοί οι αλγόριθμοι χρησιμοποιούν τα στοιχεία του κόστους, μπορούν επίσης να θεωρηθούν ως boosting αλγόριθμοι ευαίσθητου κόστους (Cost sensitive boosting algorithms). Για τον αλγόριθμο AdaBoost, η επιλογή της παραμέτρου ενημέρωσης βάρους είναι ζωτικής σημασίας για τη μετατροπή ενός ασθενούς αλγορίθμου μάθησης σε ισχυρό (Drummond και Holte (2006)). Όταν τα στοιχεία κόστους εισάγονται στον τύπο ενημέρωσης βάρους του αλγορίθμου AdaBoost, η επικαιροποιημένη κατανομή των δεδομένων επηρεάζεται από τα στοιχεία του κόστους. Χωρίς την εκ νέου εισαγωγή της παράμετρος ενημέρωσης, η οποία λαμβάνει υπόψη τα στοιχεία κόστους για κάθε cost sensitive boosting αλγόριθμο, η αποδοτικότητα του boosting αλγορίθμου δεν είναι εγγυημένη.

2.5 Feature selection approaches

Η κύρια ιδέα της επιλογής χαρακτηριστικών είναι να επιλέξουμε ένα υποσύνολο των χαρακτηριστικών εισόδου (των επεξηγηματικών μεταβλητών) με την εξάλειψη χαρακτηριστικών με μικρή ή καθόλου προγνωστική πληροφορία σύμφωνα με κάποιο μέτρο. Για την υιοθέτηση της επιλογής χαρακτηριστικών εντός του προβλήματος της ανισορροπίας μεταξύ των κλάσεων, υπάρχουν δύο προσεγγίσεις. Η πρώτη βασίζεται στην προσαρμογή των εκτιμήσεων της κατανομής πιθανότητας των κλάσεων. Η επόμενη

προσέγγιση βασίζεται στην αρχή νέων μέτρων επιλογής χαρακτηριστικών (Kotsiantis et al.(2006)).

2.5.1 Μέθοδοι

2.5.1.1 Information gain

Το κέρδος πληροφορίας (Moreno-Torres και Herrera (2010)) μετρά τον αριθμό των bits πληροφορίας που λαμβάνονται για την πρόβλεψη της κατηγορίας γνωρίζοντας την παρουσία ή την απουσία ενός όρου σε ένα έγγραφο. Η μέθοδος του κέρδους πληροφορίας είναι επίσης γνωστή ως Αναμενόμενη Αμοιβαία πληροφορία (Expected Mutual Information). Η τεχνική εξομάλυνσης της εκτίμησης Αναμενόμενης Πιθανοφάνειας (Expected Likelihood Estimation) χρησιμοποιήθηκε για να χειριστεί ανωμαλίες (singularities) κατά την εκτίμηση αυτών των πιθανοτήτων (Yang and Pedersen (1997)).

2.5.1.2 Odds Ratio

Η μέθοδος των συμπληρωματικών πιθανοτήτων (OR) μετρά τα odds της λέξης που συμβαίνουν στην θετική κλάση και που ομαλοποιούνται από αυτά της αρνητικής κλάσης. Η βασική σκέψη είναι ότι η κατανομή των χαρακτηριστικών επί των σχετικών εγγράφων είναι διαφορετική από την κατανομή των χαρακτηριστικών για τα μη σχετικά έγγραφα.

2.5.2 Αλγόριθμοι

2.5.2.1 Genetic Algorithms

Ο Jong Myong Choi πρότεινε μια metaheuristic προσέγγιση (Genetic Algorithm) για την υπό-δειγματοληψία ενός μη ισορροπημένου συνόλου δεδομένων στο πλαίσιο ενός ταξινομητή SVM. Ο στόχος της προσέγγισης του είναι ο εντοπισμός ενός βέλτιστου συνόλου μάθησης από μη ισορροπημένα σύνολα δεδομένων χωρίς τις εμπειρικές μελέτες που κανονικά απαιτούνται για να βρούμε τη βέλτιστη κατανομή των κλάσεων. Ενδεικτικά αποτελέσματα με πραγματικά δεδομένα δείχνουν ότι αυτή η metaheuristic υπό-δειγματοληψία εκτελείται πολύ καλά σε επανεξισορροπημένες κατανομές μεταξύ των κλάσεων. Για μεγάλης κλίμακας μη ισορροπημένα σύνολα δεδομένων, η μέθοδός

του παρέχει μία ικανή και πολύτιμη λύση για την μάθηση μη ισορροπημένων δεδομένων με έναν ταξινομητή SVM (Choi 2010).

Οι Maheshwari et al. (2011) πρότειναν Γενετικούς Αλγορίθμους (GA) για την υπερ-δειγματοληψία έτσι ώστε να μεγενθύνουν την αναλογία των αισιόδοξων δειγμάτων (των παραδειγμάτων δηλαδή που εμφανίζονται συχνότερα-κλάση πλειοψηφίας), και μετέπειτα εφάρμοσαν συσταδοποίηση (clustering) στο σύνολο δεδομένων εκπαίδευσης όπου έγινε η υπερδειγματοληψία ως μέθοδο καθαρισμού των δεδομένων για τις κλάσεις που είναι μαζί, εξαλείφοντας τα περιττά ή θορυβώδη παραδείγματα. Αυτή η προσέγγιση αναλύθηκε πλήρως και τα πειραματικά αποτελέσματα έδειξαν μια βελτίωση στην ταξινόμηση με βάση την περιοχή κάτω από τη ROC καμπύλη.

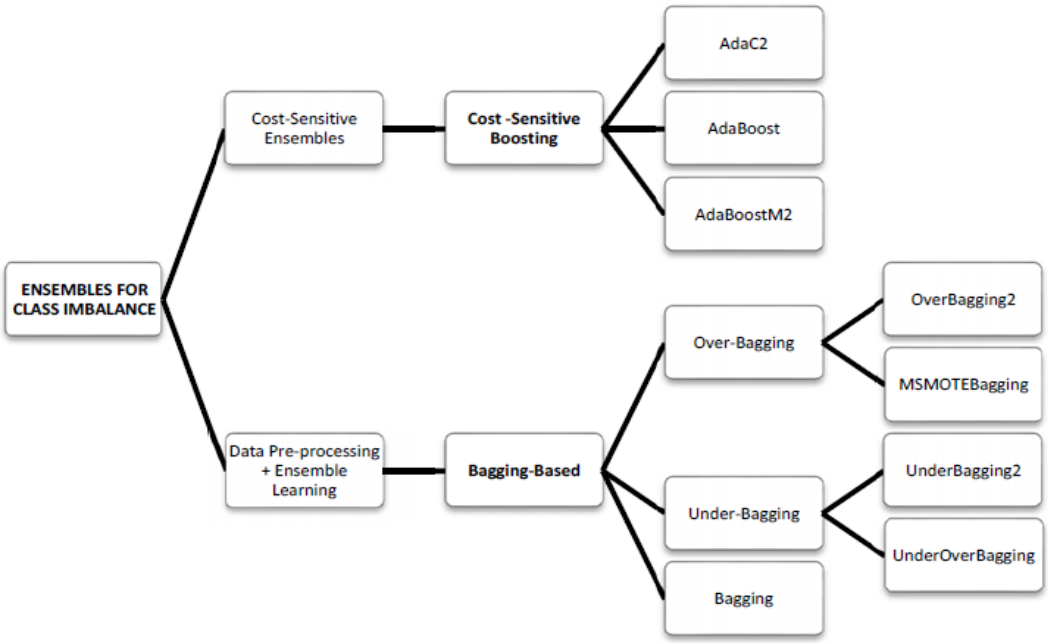
2.5.2.2 Fuzzy rule based classification system

Οι Fernandez et al. (2009) πρότειναν τη χρήση του ιεραρχικού ασαφή κανόνα που βασίζεται στο σύστημα ταξινόμησης (Fuzzy rule based classification system), το οποίο στηρίζεται στην τροποποίηση ενός απλού μοντέλου γλωσσικής ασάφειας μέσω της επέκτασης της βάσης της γνώσης με έναν ιεραρχικό τρόπο και τη χρήση μιας διαδικασίας επιλογής με ένα γενετικό κανόνα, προκειμένου να πάρει ένα συμπαγές και ακριβές μοντέλο. Η εξαιρετική απόδοση αυτής της προσέγγισης παρουσιάζεται μέσα από μια εκτεταμένη πειραματική μελέτη που πραγματοποιήθηκε σε μια μεγάλη συλλογή από μη ισορροπημένα σύνολα δεδομένων.

2.6 Προσέγγιση Συνόλου μεθόδων (Ensemble level approaches)

Η Ensemble μάθηση παρακινείται από την απώλεια πληροφορίας που εμφανίζεται στην υπο-δειγματοληψία. Στην ensemble μάθηση, οι πολλαπλοί ταξινομητές δημιουργούνται από τα υποσύνολα εκπαίδευσης από το αρχικό σύνολο δεδομένων. Στο τέλος, οι ταξινομητές συνδυάζονται σε μια διαδικασία μάθησης και η τελική ταξινόμηση καθορίζεται από ένα σύστημα ψηφοφορίας.

Οι Ensemble μέθοδοι βελτιώνουν την απόδοση του συνολικού συστήματος. Η αποτελεσματικότητα των ensemble μεθόδων εξαρτάται πολύ από την ανεξαρτησία του σφάλματος που διέπραξε η βασική μέθοδος εκμάθησης (base learner). Η απόδοση των ensemble μεθόδων εξαρτάται σε μεγάλο βαθμό από την ακρίβεια και την ποικιλομορφία της βασικής μεθόδου εκμάθησης. Η πιο απλή προσέγγιση για να δημιουργήσουμε διαφοροποιημένους ταξινομητές ως βάση είναι με το χειρισμό των δεδομένων εκπαίδευσης Zhang, et al. (2012).



Σχήμα 2.3 Σύνολο μεθόδων για τα μη ισορροπημένα δεδομένα

2.6.1 Μέθοδοι

2.6.1.1 Boosting και Bagging

Οι τεχνικές Boosting και Bagging (boostrap συσσωμάτωσης) είναι οι πιο επιτυχημένες προσεγγίσεις. Οι περισσότεροι αλγόριθμοι χρησιμοποιούν επαναληπτική μάθηση αδύναμων ταξινομητών που έχουν παραχθεί με την τοποθέτηση διαφορετικών βαρών στις περιπτώσεις εκπαίδευσης. Σε κάθε επανάληψη, το boosting αυξάνει τα βάρη για τις εσφαλμένα ταξινομημένες περιπτώσεις και μειώνει τα βάρη για τις σωστά ταξινομημένες, δίνοντας στην επόμενη επανάληψη μεγαλύτερη προσοχή στις εσφαλμένα ταξινομημένες περιπτώσεις. Το Rare-Boost κλιμακώνει τα ψευδώς-θετικά παραδείγματα ανάλογα με το πόσο καλά διαφοροποιούνται από τα αληθώς-θετικά παραδείγματα και κλίμακώνει τα ψευδώς-θετικά παραδείγματα ανάλογα με το πόσο καλά αυτά διακρίνονται από τα αληθώς-αρνητικά παραδείγματα (Joshi et al., 2001).

2.6.1.2 SMOTEBoost

Ο SMOTEBoost (Chawla et al., 2003) ασχολήθηκε με το θέμα της υπερπροσαρμογής που μπορεί να προκαλεί το boosting όπως έγινε και με τη μέθοδο της υπερ-δειγματοληψίας. Αντί της ενημέρωσης των βαρών έτσι ώστε να αλλάξει η κατανομή του συνόλου δεδομένων εκπαίδευσης, η μέθοδος αυτή προσθέτει νέες περιπτώσεις της κλάσης μειοψηφίας χρησιμοποιώντας τη μέθοδο SMOTE.

2.6.1.3 Chan και Stolfo μέθοδος

Οι Chan και Stolfo (2001) πρότειναν μια άλλη ensemble μέθοδο εννοιολογικά παρόμοια με την προσέγγιση του Bagging. Διεξήγαγαν ορισμένα προκαταρκτικά πειράματα για να προσδιορίσει μια επιθυμητή κατανομή κλάσης που αποφεύγει το πρόβλημα ανισορροπίας των κλάσεων, και στη συνέχεια υποβάλλεται σε αναδειγματοληψία ώστε να δημιουργήσει πολλαπλά σύνολα εκπαίδευσης βάσει της επιθυμητής κατανομής για την κλάση. Κάθε σύνολο εκπαίδευσης περιείχε όλες τις περιπτώσεις της κλάσης μειοψηφίας και ένα υποσύνολο των περιπτώσεων της κλάσης πλειοψηφίας. Για να χρησιμοποιήσεις όλες τις περιπτώσεις της κλάσης πλειοψηφίας, κάθε περίπτωση της κλάσης πλειοψηφίας εμφανίστηκε σε τουλάχιστον ένα σύνολο εκπαίδευσης. Τέλος, ο αλγόριθμος μάθησης εφαρμόστηκε σε κάθε σύνολο εκπαίδευσης και ένας σύνθετος «μαθητής» (learner) δημιουργήθηκε από τα αποτελέσματα της ταξινόμησης όλων των ταξινομητών.

2.6.1.4 Kernel-based methods

Στις μεθόδους που βασίζονται στον πυρήνα (kernel-based methods), υπάρχουν πολλές εργασίες όπου εφαρμόζουν δειγματοληψία και ensemble τεχνικές στις Μηχανές Διανυσματικής Υποστήριξης (SVM). Διαφορετικά σφάλματα στα κόστη προτάθηκαν για διαφορετικές κλάσεις για να κάνουν μεροληπτικές τις SVM έτσι ώστε να μετατοπιστεί το όριο απόφασης μακριά από τα θετικά παραδείγματα και να κάνουν τα θετικά παραδείγματα, να κατανέμονται με μεγαλύτερη πυκνότητα. Άλλες μέθοδοι ανέπτυξαν ένα ensemble σύστημα τροποποιώντας την κατανομή των δεδομένων και το SVM με ασύμμετρο κόστος εσφαλμένης ταξινόμησης, προκειμένου να ενισχύσουν την απόδοση (Akbani et al. (2004)).

2.6.2 Αλγόριθμοι

2.6.2.1 Αλγόριθμοι EasyEnsemble και BalanceCascade

Πρόσφατα, εισήχθησαν δύο αλγόριθμοι, ο EasyEnsemble και ο BalanceCascade (Liu et al., 2006b). Η στρατηγική των δύο αυτών μεθόδων είναι να δημιουργήσουν πολλά σύνολα εκπαίδευσης, κρατώντας όλες τις περιπτώσεις της κλάσης μειοψηφίας και εκτελώντας υπο-δειγματοληψία σε διάφορα υποσύνολα από την κλάση πλειοψηφίας. Με επανάθεση στη δειγματοληψία της κλάσης πλειοψηφίας, αυτές οι μέθοδοι ξεπερνούν την πιθανή απώλεια της πληροφορίας από την κλάση πλειοψηφίας.

EasyEnsemble

Η μέθοδος *EasyEnsemble* δημιουργεί ανεξάρτητα δείγματα (με επανάθεση) από αρκετά υποσύνολα της κλάσης πλειοψηφίας των οποίων το μέγεθος είναι ίσο με το μέγεθος της κλάσης μειοψηφίας και παράγει τους ατομικούς ταξινομητές για τα υποσύνολα. Με άλλα λόγια, ο EasyEnsemble δημιουργεί T ισορροπημένα σύνολα εκπαίδευσης. Η έξοδος (output) της μάθησης του i-οστού συνόλου εκπαίδευσης είναι ένας AdaBoost ταξινομητής H_i (i = 1, .., T). Τότε όλοι οι ταξινομητές που δημιουργούνται, $H_{i=1,\ldots,T}$, τ συνδυάζονται για την τελική απόφαση.

BalanceCascade

Η μέθοδος *BalanceCascade* μειώνει επαναληπτικά το μέγεθος της κλάσης πλειοψηφίας, βασιζόμενη στον πιο πρόσφατο ταξινομητή. Αυτός ο αλγόριθμος χρησιμοποιεί ένα εκπαιδευμένο ταξινομητή για να καθοδηγήσει τη διαδικασία δειγματοληψίας για τους επόμενους ταξινομητές. Αρχικά, εκτελεί δειγματοληψία ενός ισορροπημένου συνόλου εκπαίδευσης όπως ο EasyEnsemble. Αφού εκπαιδευτεί το ensemble Adaboost με το αρχικό, ισορροπημένο, σύνολο εκπαίδευσης, όλες οι περιπτώσεις πλειοψηφίας που έχουν ταξινομηθεί σωστά αφαιρούνται από την κλάση πλειοψηφίας. Με τον τρόπο αυτό, το σύνολο δεδομένων της κλάσης πλειοψηφίας του συνόλου εκπαίδευσης μειώνεται μετά την εκπαίδευση του κάθε AdaBoost ensemble, H_i. Αυτή η στρατηγική δειγματοληψίας μειώνει την πλεονάζουσα πληροφορία της κλάσης πλειοψηφίας και εξερευνά όσες χρήσιμες πληροφορίες είναι δυνατόν.

Εκτός από τις μεθόδους που έχουν ήδη συζητηθεί υπάρχουν και άλλες προσεγγίσεις που έχουν χρησιμοποιηθεί για την αντιμετώπιση του προβλήματος της ανισορροπίας των κλάσεων. Για παράδειγμα, η επιλογή μεταβλητών (χαρακτηριστικών) χρησιμοποιήθηκε για την επιλογή σημαντικών μεταβλητών για τις κλάσεις πλειοψηφίας και μειοψηφίας χωριστά και στη συνέχεια γίνεται ο συνδυασμός τους (Zheng et al.2004).

2.6.2.2 Νευρωνικά δίκτυα (Neural Network)

Οι Ghazikhani et al.(2013) πρότειναν μία online μέθοδο ενός συνόλου ταξινομητών NN. Τα ensemble μοντέλα είναι οι πιο συχνές μέθοδοι που χρησιμοποιούνται για την ταξινόμηση των μη-σταθερών και μη ισορροπημένων ροών δεδομένων. Η κύρια συνεισφορά είναι μια προσέγγιση δύο επιπέδων για τον χειρισμό τόσο της ανισορροπίας μεταξύ των κλάσεων όσο και της μη σταθερότητας. Στο *πρώτο* στρώμα, είναι ενσωματωμένη μέσα στην φάση της εκπαίδευσης των NNs μία μέθοδος μάθησης ευαίσθητου κόστους (cost sensitive learning), και στο *δεύτερο* στρώμα χρησιμοποιείται μία νέα μέθοδος για τη στάθμιση του συνόλου των ταξινομητών. Αυτή η μέθοδος αξιολογήθηκε σε 3 συνθετικά και 8 πραγματικά σύνολα δεδομένων και τα αποτελέσματα έδειξαν μία στατιστικά σημαντική βελτίωση σε σύγκριση με άλλες ensemble μεθόδους με παρόμοια χαρακτηριστικά.

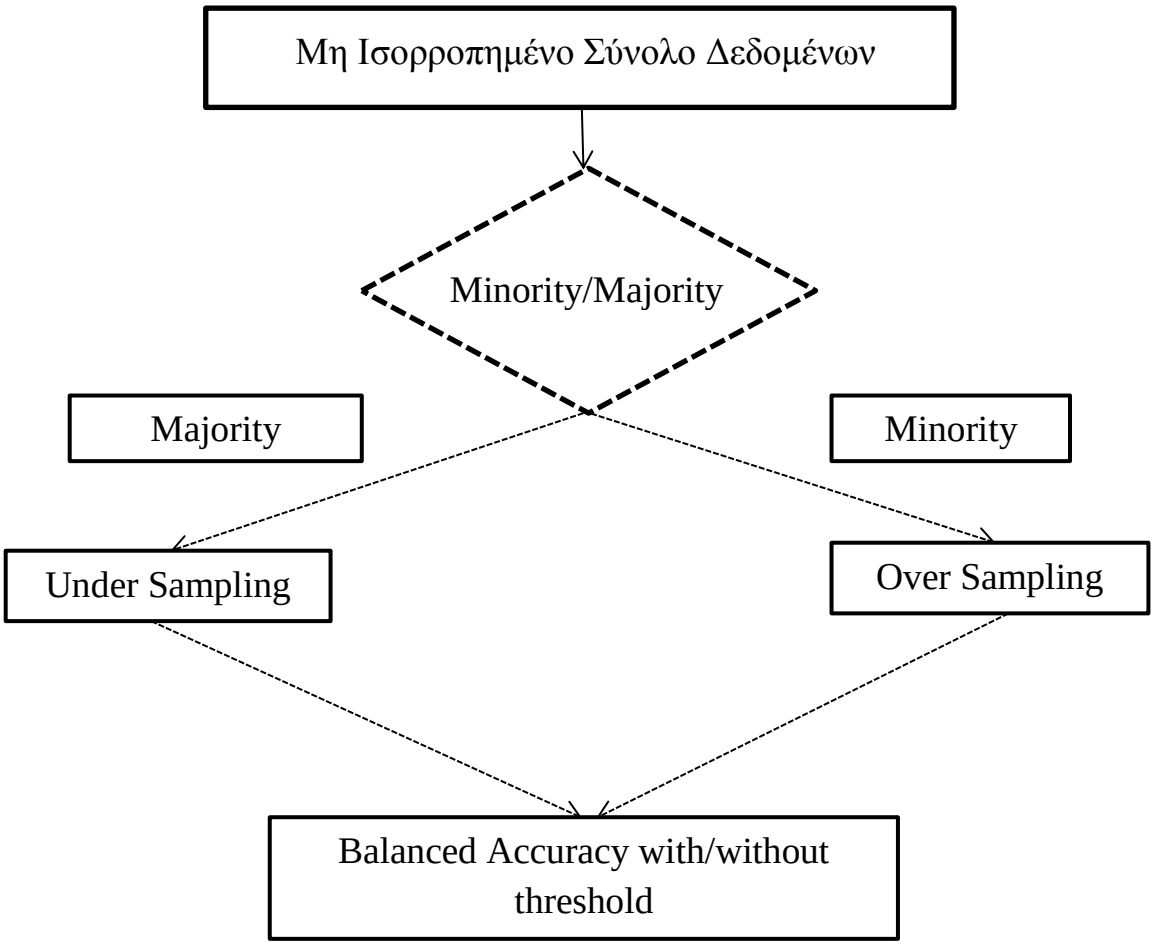
Οι Mazurowski et al. (2008), ερεύνησαν την επίδραση των μη ισορροπημένων κλάσεων στην εκπαίδευση των δεδομένων κατά την ανάπτυξη των νευρωνικών δικτύων με στόχο την ιατρική διάγνωση με τη βοήθεια υπολογιστή. Η έρευνα διεξήχθει σε ιατρικά δεδομένα με μικρό μέγεθος του δείγματος εκπαίδευσης και τεράστιο αριθμό χαρακτηριστικών και συσχετίσεις μεταξύ των χαρακτηριστικών. Διερευνήθηκαν δύο τρόποι εκπαίδευσης νευρωνικών δικτύων: η κλασική ανάστροφη διάδοση (BP) και η βελτιστοποίηση σμήνους σωματιδίων (PSO) με τη χρήση ενός κλινικού κριτηρίου για την εκπαίδευση. Έγινε μια δοκιμαστική μάθηση με τη χρήση προσομοιωμένων δεδομένων και τα συμπεράσματα επικυρώθηκαν περαιτέρω με τη εφαρμογή της μεθόδου σε πραγματικά κλινικά δεδομένα για τη διάγνωση του καρκίνου του μαστού. Τα αποτελέσματά δείχνουν ότι η απόδοση του ταξινομητή χειροτερεύει ακόμη και με μέτρια ανισορροπία μεταξύ των κλάσεων στα δεδομένα εκπαίδευσης. Περαιτέρω, φάνηκε ότι η BP γενικά προτιμάται σε σχέση με μέθοδο PSO για την έλλειψη ισορροπίας στα δεδομένα εκπαίδευσης, ιδίως όταν το δείγμα των δεδομένων είναι μικρό και ο αριθμός των χαρακτηριστικών συγκριτικά μεγάλος.

2.6.2.3 Kernel classifier identification (Ταυτοποίηση ταξινομητή με πυρήνα)

Ο Xia Χονγκ πρότεινε έναν αλγόριθμο kernel classifier identification που βασίζεται σε ένα νέο εκτιμητή κανονικοποιημένων ορθογώνιων σταθμισμένων ελαχίστων τετραγώνων (regularized orthogonal weighted least squares, ROWLS) και το κριτήριο επιλογής μοντέλου of maximal leave one out area under the curve, LOO-AUC) των καμπύλων (ROC). Φαίνεται καθαρά ότι, λόγω της μεθόδου ορθογωνοποίησης, η LOO-AUC μπορεί να μελετηθεί χρησιμοποιώντας έναν αναλυτικό τύπο βασιζόμενο στο νέο εκτιμητή κανονικοποιημένων ορθογώνιων σταθμισμένων ελαχίστων τετραγώνων της παραμέτρου, χωρίς να διαιρέσει στην πραγματικότητα το σύνολο των δεδομένων της αξιολόγησης. Αυτός ο αλγόριθμος μπορεί να επιτευχθεί με το ελάχιστο υπολογιστικό κόστος μέσω μιας σειράς ενημερώσεων του αναδρομικού τύπου αναζητώντας το μοντέλο με τη μέγιστη αυξητική LOO-AUC τιμή. Για την παρουσίαση της αποδοτικότητας του αλγορίθμου χρησιμοποιήθηκαν αριθμητικά μοντέλα (Hong, X. (2007)).

2.7 Γενικά βήματα για την επίλυση προβλήματων μη ισορροπημένων δεδομένων

Τα γενικά βήματα που μπορούν να συμμετάσχουν στην επίλυση του προβλήματος των μη ισορροπημένων δεδομένων μπορεί να περιγραφεί στο ακόλουθο σχήμα (Σχήμα 2.4) Το Σχήμα 2.4 δείχνει σαφώς ότι τα μη ισορροπημένα προβλήματα προκύπτουν στην περίπτωση που έχουμε κλάσεις μειοψηφίας ή πλειοψηφίας. Στη μειονοτική κλάση των δεδομένων πραγματοποιείται υπερδειγματοληψία για την εξισορρόπησή της. Στην περίπτωση της κλάσης πλειοψηφίας πραγματοποιείται υποδειγματοληψία για το σύνολο των δεδομένων. Τέλος, μπορούμε να βρούμε την ισορροπημένη ακρίβεια με ή χωρίς τιμή κατωφλίου για την επίλυση του προβήματος της ανισορροπίας.



Σχήμα 2.4: Βήματα για την επίλυση προβλήματων μη ισορροπημένων δεδομένων

ΚΕΦΑΛΑΙΟ 3

Το πρόβλημα της ανισσοροπίας των κλάσεων με Μηχανές Διανυσματικής Υποστήριξης
(class imbalance problem with support vector machine learning)

Οι Μηχανές Διανυσματικής Υποστήριξης (Support Vector Machines, SVM) είναι μία πολύ δημοφιλής τεχνική που ανήκει στις μεθόδους της μηχανικής μάθησης. Παρά τα θεωρητικά και πρακτικά πλεονεκτήματα τους, οι SVMs δεν παράγουν βέλτιστα αποτελέσματα με μη ισορροπημένα σύνολα δεδομένων. Δηλαδή, ένας ταξινομητής SVM που εκπαιδεύεται σε μη ισορροπημένα δεδομένα μπορεί να παράγει μοντέλα που μεροληπτούν προς την κλάση πλειοψηφίας (majority) και έχουν χαμηλή απόδοση στην κλάση της μειοψηφίας, όπως συμβαίνει και με τις περισσότερες μεθόδους ταξινόμησης. Υπάρχουν διάφορες τεχνικές προεπεξεργασίας των δεδομένων και αλγοριθμικές τεχνικές που προτείνονται στη βιβλιογραφία με σκοπό να μετριασθεί το πρόβλημα της ανισορροπίας για τις SVMs. Στο κεφάλαιο αυτό έχουμε ως στόχο να εξετάσουμε κάποιες από αυτές τις τεχνικές.

3.1 Εισαγωγή

Οι Μηχανές Διανυσματικής Υποστήριξης σαν ιδέα δημιουργήθηκαν από τον Cortes και τον Vapnik (1995) (κοίτα επίσης και Vapnik(2000)). Η SVM τεχνική βασίζεται στην

στατιστική θεωρία της μηχανικής μάθησης[6] και μπορεί να χρησιμοποιηθεί για την πρόβλεψη μελλοντικών δεδομένων. Εκπαιδεύεται από την επίλυση ενός περιορισμένου προβλήματος ταξινόμησης και υλοποιεί τη χαρτογράφηση των συντελεστών παραγωγής σε ένα υψηλό τρισδιάστατο χώρο χρησιμοποιώντας ένα σύνολο μη γραμμικών βασικών συναρτήσεων. Η SVM τεχνική μπορεί να χρησιμοποιηθεί για μια ποικιλία από αναπαραστάσεις όπως τα νευρωνικά δίκτυα, τα splines, τους πολυωνυμικούς εκτιμητές κ.λπ., αλλά υπάρχει μια μοναδική βέλτιστη λύση για κάθε επιλογή των SVM παραμέτρων. Αυτό είναι διαφορετικό σε άλλες μηχανές μάθησης όπως τα τυποποιημένα Νευρωνικά Δίκτυα που χρησιμοποιούν την προς τα πίσω διάδοση. Με λίγα λόγια η ανάπτυξη της SVM μεθόδου είναι εντελώς διαφορετική από τους συνήθεις αλγόριθμους που χρησιμοποιούνται για τη μάθηση και έτσι η SVM τεχνική παρέχει μία νέα άποψη μάθησης. Τα τέσσερα πιο σημαντικά χαρακτηριστικά της SVM τεχνικής είναι η δυαδικότητα**,** οι πυρήνες, η κυρτότητα και η σποραδικότητα.

Οι μηχανές διανυσματικής υποστήριξης λειτουργούν ως μία από τις καλύτερες προσεγγίσεις για τη μοντελοποίηση δεδομένων. Συνδυάζουν τον γενικευμένο έλεγχο ως μία τεχνική για τον έλεγχο των διαστάσεων. Η χαρτογράφηση του πυρήνα παρέχει μια κοινή βάση για τα περισσότερα από τα συνηθισμένα αρχιτεκτονικά μοντέλα και συγκρίσεις που πρέπει να εκτελεστούν. Στα προβλήματα ταξινόμησης επιτυγχάνεται γενικευμένος έλεγχος με τη μεγιστοποίηση του περιθωρίου κέρδους, το οποίο αντιστοιχεί στην ελαχιστοποίηση του διανύσματος σε ένα κανονικό πλαίσιο. Η ελαχιστοποίηση του διανύσματος βάρους μπορεί να χρησιμοποιηθεί ως κριτήριο και σε προβλήματα παλινδρόμησης με μία τροποποιημένη συνάρτηση απώλειας. Οι μελλοντικές κατευθύνσεις περιλαμβάνουν μία τεχνική για την επιλογή αυτής της συνάρτησης και έναν επιπλέον έλεγχο της ικανότητας.

Οι Μηχανές διανυσματικής υποστήριξης (SVMs) (V. Vapnik (2000), Cortes and Vapnik (1995), Boser et al. (1992), Cristianinio and Shawe-Taylor (2000), Scholkopf and Smola (2001), Burges (1998), Chang and Lin (2011), Veropoulos et al.(1999)) είναι μια δημοφιλής τεχνική της μηχανικής μάθησης, η οποία έχει εφαρμοστεί με επιτυχία σε πολλά πραγματικά προβλήματα ταξινόμησης από διάφορους τομείς. Λόγω των θεωρητικών και πρακτικών της πλεονεκτημάτων, όπως το στέρεο μαθηματικό υπόβαθρο, η ικανότητα γενίκευσης και η ικανότητα εύρεσης γενικών και μη-γραμμικών λύσεων σε προβλήματα ταξινόμησης. Οι SVMs είναι λοιπόν μία από τις πιο δημοφιλείς μεθόδους για τους ερευνητές ανάμεσα στα ποικίλα εργαλεία της μηχανικής μάθησης και της εξόρυξης δεδομένων.

[6] Οι αλγόριθμοι μηχανικής μάθησης έχουν στόχο τις αναπαραστάσεις απλών λειτουργιών. Ως εκ τούτου, στόχος της εκπαίδευσης είναι το αποτέλεσμα μιας υπόθεσης που πραγματοποιεί σωστή ταξινόμηση των δεδομένων εκπαίδευσης και οι αρχικοί αλγόριθμοι εκμάθησης έχουν σχεδιαστεί για να βρίσκουνε μία τέτοια λύση που να ταιριάζει με τα δεδομένα. Η SVM τεχνική αποδίδει καλύτερα σε όρους που δεν είναι πάνω στη γενίκευση, σε αντίθεση με τα νευρωνικά δίκτυα τα οποία καταλήγουν πιο εύκολα σε γενίκευση.

Αν και οι SVMs συχνά δουλεύουν αποδοτικά με ισορροπημένα σύνολα δεδομένων, θα μπορούσαν να παράξουν ανεπαρκή αποτελέσματα με μη ισορροπημένα σύνολα δεδομένων. Πιο συγκεκριμένα, ένας SVM ταξινομητής που εκπαιδεύτηκε σε ένα μη ισορροπημένο σύνολο δεδομένων παράγει συχνά μοντέλα που είναι μεροληπτικά προς την κλάση πλειοψηφίας και έχουν χαμηλή απόδοση στην κλάση μειοψηφίας. Υπάρχουν διάφορες τεχνικές προεπεξεργασίας των δεδομένων και αλγοριθμικές τεχνικές που προτείνονται για να ξεπεραστεί αυτό το πρόβλημα για τις SVMs. Σε αυτό το κεφάλαιο θα συζητήσουμε κάποιες από αυτές τις τεχνικές. Στην ενότητα 3.2 του παρόντος κεφαλαίου παρουσιάζουμε κάποιο υπόβαθρο για τον αλγόριθμο μάθησης SVM. Στην παράγραφο 3.3, συζητάμε γιατί οι SVMs είναι ευαίσθητες στην ανισορροπία των συνόλων δεδομένων. Οι ενότητες 3.4 και 3.5 παρουσιάζουν τις υπάρχουσες τεχνικές που προτείνονται στη βιβλιογραφία για να χειριστούμε το πρόβλημα της ανισορροπία των κλάσεων για τις SVMs.

3.2 Εισαγωγή στις Μηχανές Διανυσματικής Υποστήριξης

3.2.1 Η SVM μέθοδος για την δυαδική ταξινόμηση

Η SVM είναι μία χρήσιμη τεχνική για την ταξινόμηση των δεδομένων. Ακόμη και αν τα Νευρωνικά Δίκτυα θεωρούνται ότι είναι ευκολότερα στη χρήση, μερικές φορές λαμβάνονται μη ικανοποιητικά αποτελέσματα. Μια διαδικασία ταξινόμησης περιλαμβάνει συνήθως τα δεδομένα εκπαίδευσης και εξέτασης που αποτελούνται από κάποιες περιπτώσεις (στιγμιότυπα) δεδομένων. Κάθε στιγμιότυπο, στο σύνολο της εκαπαίδευσης, περιέχει μία τιμή-στόχο και διάφορα χαρακτηριστικά. Ο στόχος της SVM τεχνικής είναι να παράξει ένα μοντέλο το οποίο προβλέπει την τιμή-στόχο των δεδομένων στο σύνολο των δοκιμών.

Η SVM είναι μία μέθοδος μάθησης με πλήρη επίβλεψη[7]. Γνωστές ετικέτες βοηθάνε στην αναφορά εάν το σύστημα εκτελείται σε σωστό δρόμο ή όχι. Αυτή η πληροφορία παραπέμπει σε μια επιθυμητή απάντηση, είτε αφορά στην επικύρωση της ακρίβειας του συστήματος είτε στη χρησιμοποίηση του για να μάθει να ενεργεί σωστά. Ένα βήμα για την SVM ταξινόμηση περιλαμβάνει την αναγνώριση, κάτι το οποίο είναι άρρηκτα συνδε-δεμένο με τις γνωστές κατηγορίες. Αυτό ονομάζεται επιλογή χαρακτηριστικών ή εξαγωγή χαρακτηριστικών. Έτσι ο SVM ταξινομητής μπορεί να χρησιμοποιηθεί για να

[7]Μέθοδοι με επίβλεψη **(supervised methods)**: Οι 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 (input variables). Μερικές από αυτές τις supervised τεχνικές είναι και τα νευρωνικά δίκτυα (neural networks), τα δέντρα αποφάσεων (decision trees) και η λογιστική παλινδρόμηση (logistic regression).

προσδιορίσει τα βασικά σύνολα που εμπλέκονται στις διεργασίες για διάκριση μεταξύ των κλάσεων.

Η απλούστερη μορφή επίλυσης ενός προβλήματος πρόβλεψης είναι η δυαδική κατηγοριοποίηση (binary classification) ή δυαδική ταξινόμηση , όπου πρέπει να γίνει ένας διαχωρισμός σε αντικείμενα που ανήκουν σε μία από δύο κατηγορίες/κλάσεις οι οποίες συμβολίζονται με θετικό (+1) ή αρνητικό (-1) πρόσημο. Οι SVMs χρησιμοποιούν για την επίλυση αυτού του προβλήματος: α) διαχωρισμό δεδομένων με μεγάλο περιθώριο (large margin separation) και β) πράξεις στο επίπεδο των kernels (πυρήνων) (kernel functions).

Γραμμικά διαχωρίσιμα δεδομένα

Έχουμε n σημεία εκπαίδευσης, όπου κάθε είσοδος x_i έχει p χαρακτηριστικά (δηλαδή είναι διάστασης p) και ανήκει σε μία από τις δύο κατηγορίες y_i = -1 ή +1, δηλαδή τα δεδομένα εκπαίδευσης είναι της μορφής

$$\{x_i, y_i\}, \text{όπου } i = 1, \dots, n, y_i \in \{-1,1\}, \quad \mathbf{x} \in \Re^p$$

Εδώ υποθέτουμε ότι τα δεδομένα είναι γραμμικά διαχωρίσιμα, πράγμα που σημαίνει ότι μπορούμε να δημιουργήσουμε μια γραμμή επί του γραφήματος των x_1 vs x_2 που χωρίζει τις δύο κλάσεις όταν $p = 2$, και ένα υπερεπίπεδο σε γραφήματα $x_1, x_2, \dots, x_p$ όταν $p > 2$. Αυτό το διαχωριστικό υπερεπίπεδο στον χώρο των χαρακτηριστικών μπορεί να παρουσιαστεί ως

$$\boldsymbol{w} \cdot \Phi(\boldsymbol{x}) + b = 0$$

όπου **w** το διάνυσμα βάρους, κάθετο προς το υπερεπίπεδο και $\frac{b}{\|\boldsymbol{w}\|}$ είναι η κάθετη απόσταση από το υπερεπίπεδο προς την αρχή.

Αν το σύνολο δεδομένων είναι γραμμικά διαχωρίσιμο το διαχωριστικό υπερεπίπεδο με το μέγιστο περιθώριο μπορεί να βρεθεί από την λύση του προβλήματος βελτιστοποίησης που περιγράφουμε στη συνέχεια.

Τα διανύσματα υποστήριξης είναι τα παραδείγματα που βρίσκονται πλησιέστερα προς το διαχωριστικό υπερεπίπεδο και ο στόχος της SVM είναι να προσανατολίσει το υπερεπίπεδο κατά τέτοιο τρόπο ώστε να είναι όσο το δυνατόν μακρύτερα από τα πλησιέστερα μέλη των δύο τάξεων. Η υλοποίηση της SVM στηρίζεται στην επιλογή των μεταβλητών **w** και b, έτσι ώστε τα δεδομένα εκπαίδευσης να μπορούν να περιγραφούν με τις ακόλουθες ανισώσεις:

$$\begin{aligned} \boldsymbol{x}_i \cdot \boldsymbol{w} + b \geq +1 \quad & \gamma\iota\alpha \; y_i = +1 \\ \boldsymbol{x}_i \cdot \boldsymbol{w} + b \leq -1 \quad & \gamma\iota\alpha \; y_i = -1 \end{aligned}$$

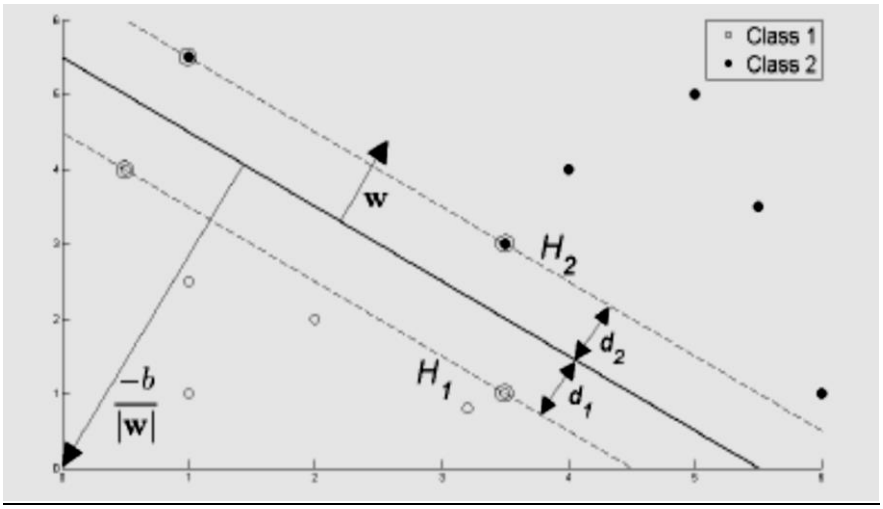
Αυτές οι εξισώσεις μπορούν να συνδυαστούν ως εξής :

$$y_i\,(\boldsymbol{x}_i \cdot \boldsymbol{w} + b)\; - 1 \;\geq\; 0 \qquad \forall i \tag{3.2.1}$$

Αν εξετάσουμε τώρα μόνο τα σημεία που κείτονται πλησιέστερα προς το διαχωριστικό υπερεπίπεδο, δηλαδή τα διανύσματα υποστήριξης (που φαίνονται σε κύκλους στο διάγραμμα, Σχήμα 3.1), τότε τα δύο επίπεδα H_1 και H_2 όπου βρίσκονται αυτά τα σημεία μπορούν να περιγραφούν με τις εξισώσεις:

$$\boldsymbol{x}_i \cdot \boldsymbol{w} + b \;=\; +1 \qquad για\;\; H_1$$

$$\boldsymbol{x}_i \cdot \boldsymbol{w} + b \;=\; -1 \qquad για\;\; H_2$$



Σχήμα 3.1: Υπερεπίπεδο ανάμεσα σε δύο γραμμικά διαχωρισμένες κλάσεις

Αναφερόμενοι στο Σχήμα 3.1, έχουμε ορίσει ως d_1 την απόσταση από το H_1 προς το υπερεπίπεδο και ως d_2 την απόσταση από το H_2 προς το υπερεπίπεδο. Η ίση απόσταση του υπερεπιπέδου από το H_1 και το H_2 σημαίνει ότι $d_1 = d_2$, και είναι μια ποσότητα γνωστή ως περιθώριο του SVM. Για να προσανατολιστεί το υπερεπίπεδο έτσι ώστε να είναι όσο το δυνατόν πιο μακριά γίνεται από τα διανύσματα υποστήριξης, θα πρέπει να μεγιστοποιηθεί αυτό το περιθώριο.

Με απλή γεωμετρία διανυσμάτων αντιλαμβανόμαστε ότι το περιθώριο είναι ίσο με $\frac{1}{\|\boldsymbol{w}\|}$ και η μεγιστοποίηση αυτού του περιθωρίου υπό τους περιορισμούς (3.2.1) είναι ισοδύναμη με την εύρεση:

$$min\;\|\boldsymbol{w}\| \qquad \text{s.t.} \qquad y_i\,(\boldsymbol{x}_i \cdot \boldsymbol{w} + b)\; - 1 \;\geq\; 0 \qquad \forall i$$

Ελαχιστοποιώντας το $\|\boldsymbol{w}\|$ είναι ισοδύναμο με την ελαχιστοποίηση του $\frac{1}{2}\|\boldsymbol{w}\|^2$ και η χρήση αυτού κάνει δυνατή την επίλυση καθώς έχουμε να χειριστούμε ένα πρόβλημα

βελτιστοποίησης τετραγωνικού προγραμματισμού (Quadratic programming optimization).

Εμείς χρειάζεται να υπολογίσουμε:

$$\min \frac{1}{2} \|\boldsymbol{w}\|^2 \quad \text{s.t.} \quad y_i \, (\boldsymbol{x}_i \cdot \boldsymbol{w} + b) \, - 1 \, \geq \, 0 \qquad \forall i \tag{3.2.2}$$

Προκειμένου να λάβουμε υπόψη τους περιορισμούς σε αυτό το πρόβλημα ελαχιστοποίησης, θα πρέπει να εφαμόσουμε τη μέθοδο των πολλαπλασιαστών Lagrange $\boldsymbol{a}$, όπου $a_i \geq 0, \forall i$ ως εξής:

$$L_P \equiv \frac{1}{2} \|\boldsymbol{w}\|^2 - \boldsymbol{a}[y_i \, (\boldsymbol{x}_i \cdot \boldsymbol{w} + b) \, - 1 \, \forall i]$$

$$\equiv \frac{1}{2} \|\boldsymbol{w}\|^2 - \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{a_i} \, [\, y_i \, (\boldsymbol{x}_i \cdot \boldsymbol{w} + b) \, - 1 \,]$$

$$\equiv \frac{1}{2} \|\boldsymbol{w}\|^2 - \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{a_i} \, y_i \, (\boldsymbol{x}_i \cdot \boldsymbol{w} + b) \, + \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{a_i} \tag{3.2.3}$$

Θέλουμε να βρούμε τα **w** και b που ελαχιστοποιούν και το $\boldsymbol{a}$ το οποίο μεγιστοποιεί την (3.2.2) (κρατώντας τα $a_i \geq 0, \forall i$).

Μπορούμε να το κάνουμε αυτό διαφορίζοντας την L_P ως προς το **w** και το b, και θέτοντας τις παραγώγους ίσες με το μηδέν :

$$\frac{\partial L_P}{\partial \mathbf{w}} = 0 \; \Rightarrow \; \mathbf{w} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{a_i} \, y_i \, \boldsymbol{x}_i \tag{3.2.4}$$

$$\frac{\partial L_P}{\partial \mathrm{b}} = 0 \; \Rightarrow \; \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{a_i} \, y_i = 0 \tag{3.2.5}$$

Αντικαθιστώντας την (3.2.4) και (3.2.5) στην (3.2.3) παίρνουμε μία άλλη μορφή η οποία εξαρτάται από το $\boldsymbol{a}$, και τότε πρέπει να μεγιστοποιήσουμε την συνάρτηση:

$$L_D \equiv \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{a_i} \; - \frac{\boldsymbol{1}}{\boldsymbol{2}} \sum_{\boldsymbol{i,j}} a_i a_j \, y_i y_j \, \boldsymbol{x}_i \boldsymbol{x}_j \quad s.t. \quad a_i \geq 0 \;\; \forall i, \;\; \sum_{i=1}^{n} a_i \, y_i = 0$$

$$\equiv \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{a_i} \; - \frac{\boldsymbol{1}}{\boldsymbol{2}} \sum_{\boldsymbol{i,j}} a_i H_{ij} a_j \qquad \text{όπου} \quad H_{ij} \equiv \; y_i y_j \, \boldsymbol{x}_i \boldsymbol{x}_j$$

$$\equiv \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{a_i} \; - \frac{\boldsymbol{1}}{\boldsymbol{2}} \, \boldsymbol{\alpha^T H \alpha} \qquad s.t. \quad a_i \geq 0 \;\; \forall i, \;\; \sum_{i=1}^{n} a_i \, y_i = 0 \tag{3.2.6}$$

Αυτή η νέα σύνθεση L_D αναφέρεται ως η διπλή μορφή της πρωτοβάθμιας L_P. Αξίζει να σημειωθεί ότι η διπλή μορφή απαιτεί μόνο να υπολογιστεί το γινόμενο όλων των διανυσμάτων εισόδου x_i. Αυτό είναι σημαντικό για το τέχνασμα του πυρήνα το οποίο περιγράφεται παρακάτω. Αφού το πρόβλημα έχει μετατοπιστεί από την ελαχιστοποίηση L_P στη μεγιστοποίηση της L_D, θα πρέπει να βρεθεί:

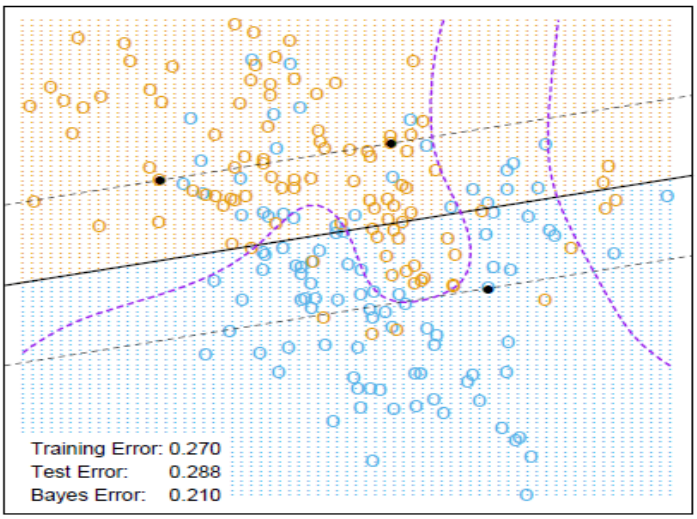
$$max \left[\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{a_i} \; - \frac{\boldsymbol{1}}{\boldsymbol{2}} \, \boldsymbol{\alpha^T H \alpha} \right] \qquad s.t. \qquad a_i \geq 0 \;\; \forall i \; , \qquad \sum_{i=1}^{n} a_i \, y_i = 0$$

Αυτό είναι ένα κυρτό τετραγωνικό πρόβλημα βελτιστοποίησης όπου χρειάζεται να πραγματοποιήσουμε μια QP επίλυση η οποία θα επιστρέψει το $\boldsymbol{\alpha}$ και από την (3.3.4) θα λάβουμε το **w**. Αυτό που απομένει είναι να υπολογιστεί το b. Αντικαθιστώντας στην (3.2.4) την (3.2.5), χρησιμοποιώντας την (3.2.1) και τέλος παίρνοντας το μέσο όρο όλων των x_s βρίσκουμε το b.

Έχουμε τώρα τις μεταβλητές w και b που ορίζουν το βέλτιστο διαχωριστικό προσανατολισμό του υπερεπιπέδου, και ως εκ τούτου τη Μηχανή Διανυσματικής Υποστήριξης. Υπενθυμίζουμε ότι η συνάρτηση απόφασης δίνεται από

$$f(x) = sign\,(w * \Phi(x) + b)$$

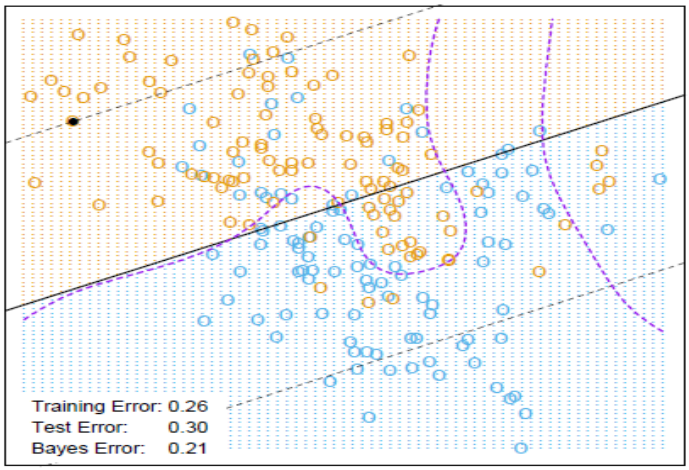
Παρακάτω παραθέτουμε δύο χαρακτηριστικά σχήματα για το γραμμικό όριο του διανύσματος υποστήριξης θεωρώντας τα δεδομένα του παραδείγματος mixture (Κεφάλαιο 2 Hastie et al. (2001), με δύο επικαλυπτόμενες κλάσεις. Παρουσιάζουμε τη συμπεριφορά των μηχανών διανυσματικής υποστήριξης για δύο διαφορετικές τιμές της παραμέτρου C.



$C = 10000$

Σχήμα 3.2: Για C=10000, το 62% των παρατηρήσεων είναι σημεία υποστήριξης.

Οι διακεκομμένες γραμμές καταδεικνύουν τα περιθώρια, όπου $f(x) = \pm 1$. Τα σημεία υποστήριξης ($\alpha_i > 0$) είναι όλα τα σημεία στη λάθος πλευρά του περιθωρίου τους. Οι μαύρες τελείες είναι εκείνα τα σημεία υποστήριξης που είναι ακριβώς στο περιθώριο ($\xi_i = 0, \alpha_i > 0$). Στο άνω σχήμα το 62% των παρατηρήσεων (Σχήμα 3.2) είναι σημεία υποστήριξης, ενώ στο κάτω σχήμα είναι το 85% (Σχήμα 3.3). Η διακκεκομμένη μοβ καμπύλη στο πίσω μέρος είναι το όριο απόφασης του Bayes.



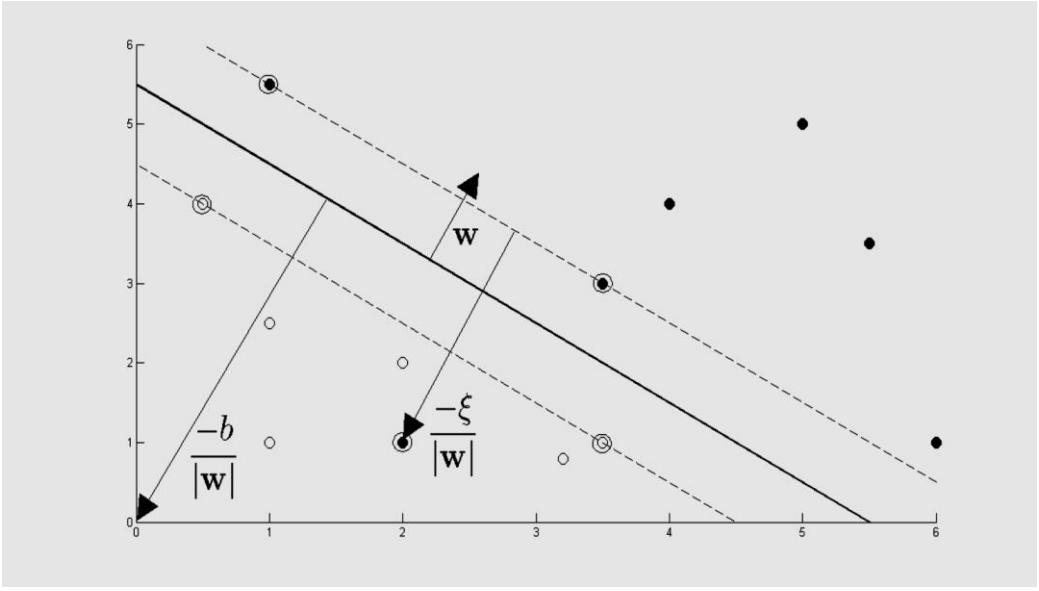
$C = 0.01$

Σχήμα 3.3: Για C=0.01 το 85% των παρατηρήσεων είναι σημεία υποστήριξης.

Μη γραμμικά διαχωρίσιμα δεδομένα

Προκειμένου να επεκταθεί η μεθοδολογία SVM για να διαχειριστεί τα δεδομένα που δεν είναι πλήρως γραμμικά διαχωρίσιμα, θα χαλαρώσουμε τους περιορισμούς (3.2.1) για να επιτρέπουν τα ελαφρώς μη ταξινομημένα σημεία. Αυτό γίνεται με την εισαγωγή μιας θετικής χαλαρής μεταβλητής $\xi_i, i = 1, \dots, n$

$$\begin{aligned} \boldsymbol{x}_i \cdot \boldsymbol{w} + b &\geq +1 - \xi_i \quad \text{για } y_i = +1 \\ \boldsymbol{x}_i \cdot \boldsymbol{w} + b &\leq -1 + \xi_i \quad \text{για } y_i = -1 \\ \xi_i &\geq 0 \; \forall i \end{aligned} \tag{3.2.7}$$



Σχήμα 3. 4 : Υπερεπίπεδο διαμέσου δύο μη γραμμικά διαχωρίσιμων κλάσεων

Αυτές οι ανισώσεις μπορούν να συνδυαστούν σε μία ως εξής :

$$y_i\ (\boldsymbol{x}_i \cdot \boldsymbol{w} + b)\ - 1 + \ \xi_i \geq\ 0\ \ \text{όπου}\ \ \ \xi_i\ \geq 0\ \forall i$$

Σε αυτό το «μαλακό» SVM περιθώριο (soft margin[8]), τα σημεία δεδομένων για την εσφαλμένη πλευρά του ορίου περιθωρίου έχουν μια ποινή που αυξάνει καθώς μεγαλώνει η απόσταση από αυτό.
Καθώς προσπαθούμε να μειώσουμε τον αριθμό των μη ταξινομημένων σημείων, ένας λογικός τρόπος για να προσαρμόσουμε την αντικειμενική μας συνάρτηση (3.2.2), είναι να βρούμε λύση στο πρόβλημα:

$$\min\left(\frac{1}{2}\|\boldsymbol{w}\|^2 + C\ \sum_{i=1}^{n}\xi_i\right) \qquad s.t. \qquad y_i\ (\boldsymbol{x}_i \cdot \boldsymbol{w} + b)\ - 1 + \ \xi_i \geq\ 0\ \ \forall i$$

όπου η παράμετρος C ελέγχει το trade-off μεταξύ της ποινής της χαλαρής μεταβλητής και του μεγέθους του περιθωρίου, w είναι το διάνυσμα των βαρών κάθετο προς το υπερεπίπεδο, ξ_i είναι οι μεταβλητές slack που κατέχουν τα παραδείγματα που δεν είναι σωστά ταξινομημένα και, ως εκ τούτου, ο όρος $\sum_{i=1}^{n}\xi_i$ μπορεί να θεωρηθεί ως ένα μέτρο του ποσού του συνόλου των εσφαλμένων ταξινομήσεων του μοντέλου (πιο συγκεκριμένα τα σφάλματα της εκπαίδευσης). Το trade-off μεταξύ της μεγιστοποίησης του περιθωρίου και της ελαχιστοποίησης των σφαλμάτων ελέγχεται από την παράμετρο του κόστους C. Η αναδιατύπωση του προβλήματος με τη βοήθεια των πολλαπλασιαστών Lagrange γίνεται ως

$$L_P\ \equiv\ \frac{1}{2}\|\boldsymbol{w}\|^2 +\ C\ \sum_{i=1}^{n}\xi_i - \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} a_i\ [\ y_i\ (\boldsymbol{x}_i \cdot \boldsymbol{w} + b) - 1 + \ \xi_i\ ] - \sum_{i=1}^{n}\mu_i\,\xi_i$$

η οποία όπως και πριν θα πρέπει να ελαχιστοποιηθεί σε σχέση με τα w, b και ξ_i και να μεγιστοποιηθεί ως προς $\boldsymbol{a}$ (όπου $a_i\ \geq 0,\ \mu_i\ \forall i$).

Διαφορίζοντας την L_P ως προς τα **w,** b και ξ_i και θέτοντας τις παραγώγους ίσες με το μηδέν έχουμε:

$$\frac{\partial L_P}{\partial \mathbf{w}} = 0\ \Rightarrow\ \mathbf{w} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{a_i}\ y_i\ \boldsymbol{x}_i$$

$$\frac{\partial L_P}{\partial \mathrm{b}} = 0\ \Rightarrow\ \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{a_i}\ y_i = 0$$

$$\frac{\partial L_P}{\partial \xi_i} = 0\ \Rightarrow\ C = \ a_i + \ \mu_i \qquad (3.2.8)$$

[8] Για να αποτραπεί η υπερβολική χρήση των λάθος τοποθετημένων σημείων, ορίζεται η σταθερά C η οποία θέτει τους όρους για τη μεγιστοποίηση του περιθωρίου και την ελαχιστοποίηση των λάθος κατηγοριοποιήσεων. Η μέθοδος αυτή ονομάζεται soft-margin SVM.

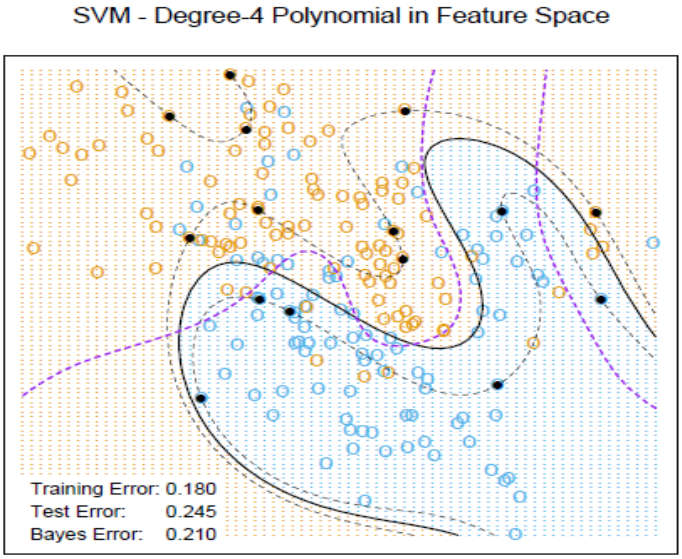
Αντικαθιστώντας, η L_P παίρνει την ίδια μορφή όπως η σχέση (3.2.6) προηγουμένως. Ωστόσο από την (3.2.8) μαζί με το ότι τα $\mu_i \geq 0$, συνεπάγεται ότι α ≤ C.

Επομένως χρειάζεται να βρούμε:

$$max\left[\sum_{i=1}^{n} \boldsymbol{a_i} - \frac{1}{2}\boldsymbol{\alpha^T H \alpha}\right] \; s.t. \;\; 0 \leq a_i \leq C \;\; \forall i, \;\; \sum_{i=1}^{n} a_i\, y_i = 0$$

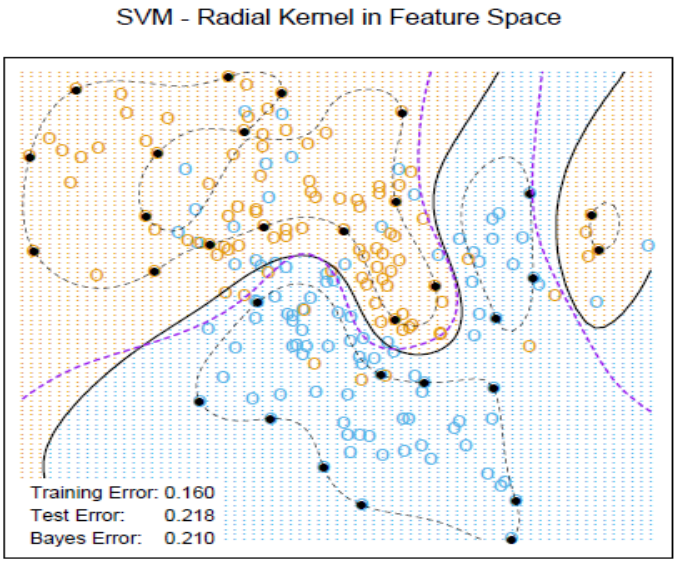
Στη συνέχεια το b υπολογίζεται με τον ίδιο τρόπο όπως προηγουμένως στην (3.2.2), αν και σε αυτή την περίπτωση το σύνολο των διανυσμάτων υποστήριξης χρησιμοποιείται για τον υπολογισμό του b που προσδιορίζεται με την εύρεση των δεικτών i, όπου $0 \leq a_i \leq C$.

Στα παρακάτω σχήματα (Σχήμα 3.5:, Σχήμα 3.6:) παρουσιάζουμε δύο μη γραμμικά SVMs για τα δεδομένα του παραδείγματος του κεφαλαίου 2 (mixture example) στο Hastie et al. (2001).



Σχήμα 3.5: SVM με πολυωνυμικό πυρήνα

Το άνω γράφημα χρησιμοποιεί 4^{ου} βαθμού πολυωνυμικό πυρήνα, το κάτω ένα Γκαουσιανό (radial basis) πυρήνα (με γ=1). Σε κάθε περίπτωση η παράμετρος C είναι συντονισμένη για την επίτευξη περίπου της καλύτερης δυνατής απόδοσης σφάλματος δοκιμής. Η τιμή C=1 λειτούργησε καλά και στις δύο περιπτώσεις. Ο radial basis πυρήνας αποδίδει καλύτερα (κοντά στο βέλτιστο Bayes), όπως θα ήταν αναμενόμενο δοθέντων των δεδομένων που προκύπτουν από το μείγμα των Gaussians. Η διακεκομμένη μοβ καμπύλη στο πίσω μέρος είναι το όριο απόφασης του Bayes.



Σχήμα 3.6: SVM με γκαουσιανό (radial basis) πυρήνα

Μία ελαφρώς διαφορετική διατύπωση του προβλήματος με τους πολλαπλασιαστές Lagrange για την εύρεση της παραμέτρου b και των συντελεστών $\boldsymbol{a_i}$ με χρήση των συναρτήσεων πυρήνα που παρουσιάζονται στη συνέχεια δίνεται από την

$$maximize \left[\sum_{i=1}^{n} a_i - \frac{1}{2}\sum_{i,j} a_i a_j y_i y_j K(x_i, x_j)\right]$$

$$subject\ to \quad 0 \leq a_i \leq C \quad \forall i, \qquad \sum_{i=1}^{n} a_i y_i = 0$$

Η οποία ικανοποιεί τις ΚΚΤ συνθήκες.

Σημειώνουμε εδώ το ότι η προσέγγιση του προβλήματος των μετρίως ισορροπημένων δεδομένων με τον ταξινομητή SVM αποτελεί μία αποτελεσματική λύση σε σύγκριση με άλλους ταξινομητές, οφείλεται στο γεγονός ότι ο SVM λαμβάνει υπόψη μόνο εκείνες τις περιπτώσεις που είναι κοντά στο όριο απόφασης, δηλαδή τα διανύσματα υποστήριξης, για την κατασκευή του μοντέλου (για περισσότερες λεπτομέρειες βλέπε Akbani et al. (2004)). Πιο συγκεκριμένα, οι Akbani et al. (2004) υποστήριξαν ότι, λόγω του περιορισμού $\sum_{i=1}^{n} a_i y_i = 0$, οι συντελεστές a_i από κάθε θετικό διάνυσμα υποστήριξης είναι λιγότερα από τα αρνητικά διανύσματα υποστήριξης, και ως εκ τούτου θα πρέπει να είναι μεγαλύτερα σε μέγεθος από τις τιμές των a_i που αντιστοιχούν στα αρνητικά διανύσματα υποστήριξης. Οι εν λόγω a_i, λειτουργούν σαν βάρη στον τελικό ταξινομητή και, κατά συνέπεια, λαμβάνουν ένα μεγαλύτερο βάρος από ότι τα αρνητικά, κάτι που λειτουργεί ως αντίβαρο σε κάποιο βαθμό, για την ανισορροπία των διανυσμάτων υποστήριξης.

3.2.2 Η SVM μέθοδος για την παλινδρόμηση

Οι SVMs μπορούν επίσης να εφαρμοστούν σε προβλήματα παλινδρόμησης με την εισαγωγή μιας εναλλακτικής λειτουργίας απώλειας . Η λειτουργία απώλειας πρέπει να τροποποιηθεί ώστε να συμπεριλάβει το μέτρο της απόστασης. Η παλινδρόμηση μπορεί

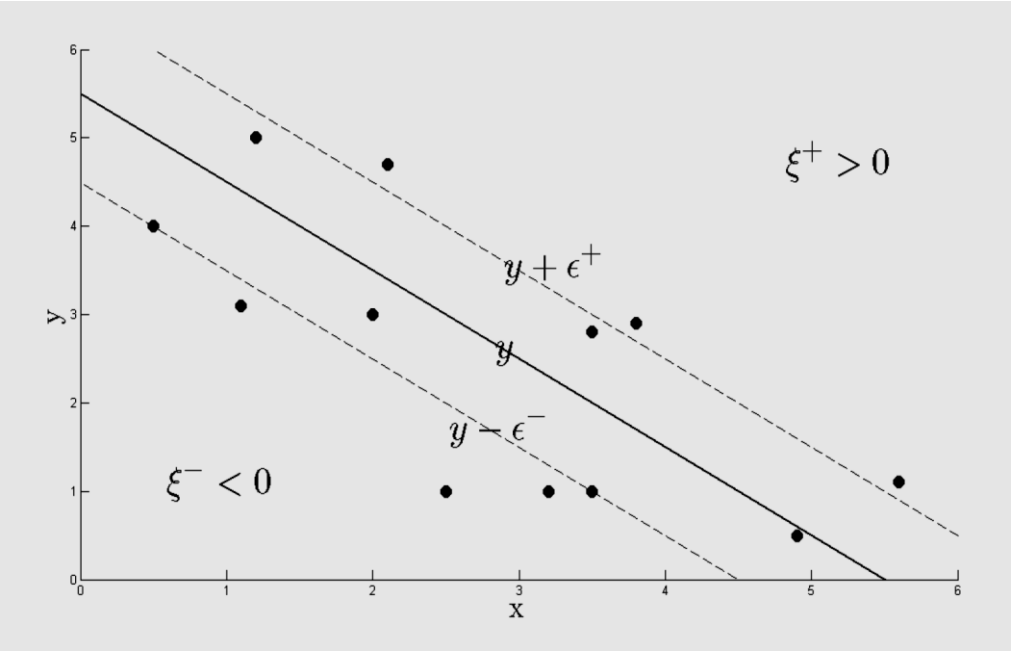
να είναι γραμμική και μη γραμμική. Τα γραμμικά μοντέλα αποτελούνται κυρίως από τις ακόλουθες λειτουργίες απώλειας, εντατικές λειτουργίες απώλειας, τετραγωνική και Huber λειτουργίες απώλειας.

Ομοίως με τα προβλήματα ταξινόμησης, ένα μη γραμμικό μοντέλο συνήθως απαιτεί επαρκή δεδομένα. Με τον ίδιο τρόπο, μια μη γραμμική χαρτογράφηση μπορεί να χρησιμοποιηθεί για να χαρτογραφήσει τα δεδομένα σε ένα υψηλό διαστάσεων χώρο χαρακτηριστικών, όπου η γραμμική παλινδρόμηση εκτελείται. Η προσέγγιση του πυρήνα και πάλι χρησιμοποιείται για την αντιμετώπιση της διάστασης. Στη μέθοδο παλινδρόμησης υπάρχουν εκτιμήσεις που βασίζονται σε προγενέστερη γνώση του προβλήματος και τη διανομή του θορύβου.

Αντί να προσπαθούμε να κατατάξουμε νέες άγνωστες μεταβλητές x΄ σε μία από τις δύο κατηγορίες $y' = \pm 1$, τώρα επιθυμούμε να προβλέψουμε μια πραγματική τιμή εξόδου για το y' και έτσι τα δεδομένα εκπαίδευσης μας είναι της μορφής:

$$\{x_i, y_i\}, \qquad \text{όπου} \qquad i = 1, \dots, n, \quad y_i \in \Re, \quad \mathbf{x} \in \Re^p$$

$$y_i = \boldsymbol{x}_i \cdot \boldsymbol{w} + b \tag{3.2.9}$$



Σχήμα 3.7 : Παλινδρόμηση με ε- insensitive σωλήνα (tube)

Η SVM παλινδρόμηση χρησιμοποιεί μια πιο εξελιγμένη λειτουργία ποινής από πριν, μη χορηγώντας ποινή εάν η προβλεπόμενη τιμή y_i είναι μικρότερη από απόσταση ε μακριά από την πραγματική τιμή t_i, δηλαδή αν $|t_i - y_i| < \varepsilon$. Αναφερόμενοι στο Σχήμα 3.7, η περιοχή που οριοθετείται από $y_i \pm \varepsilon$ λέγεται ε - insensitive σωλήνας (tube). Μια άλλη μορφοποίηση στη λειτουργία ποινής είναι ότι οι μεταβλητές εξόδου που είναι εκτός του σωλήνα δίνουν μία από τις δύο χαλαρές μεταβλητές, ποινές ανάλογα με το αν

βρίσκονται πάνω (ξ^+) ή κάτω από το σωλήνα (ξ^-) (όπου $\xi^+ > 0$, $\xi^- > 0, \forall i$):

$$t_i \leq y_i + \varepsilon + \xi^+$$

$$t_i \geq y_i - \varepsilon - \xi^+ \quad (3.2.10)$$

Η συνάρτηση σφάλματος για την SVM παλινδρόμηση (SVR) μπορεί να γραφεί ως εξής:

$$C \sum_{i=1}^{n} (\xi_i^+ + \xi_i^-) + \frac{1}{2}\|\boldsymbol{w}\|^2$$

Αυτή χρειάζεται να ελαχιστοποιηθεί υπό τους περιορισμούς $\xi^+ \geq 0$, $\xi^- \geq 0, \forall i$ και την (3.2.9) και (3.2.10). Για να το κάνουμε αυτό εισάγουμε τους πολλαπλασιαστές Lagrange

$$\alpha_i^+ \geq 0, \qquad \alpha_i^- \geq 0, \qquad \mu_i^+ \geq 0, \ \mu_i^- \geq 0 \ \ \forall i$$

Με την ίδια διαδικασία δημιουργούμε την L_P , διαφορίζουμε ως προς w, b, ξ^+ και ξ^- και θέτουμε τις παραγώγους ίσες με το μηδέν. Με ανάλογο τρόπο αντικαθιστούμε και βρίσκουμε το L_D το οποίο θέλουμε να μεγιστοποιήσουμε ως προς α_i^+ και α_i^- ($\alpha_i^+ \geq 0$, $\alpha_i^- \geq 0, \ \forall i$).

Με τα βήματα λοιπόν που περιγράφηκαν στα προηγούμενα κεφάλαια βρίσκουμε τις παραμέτρους που χρειαζόμαστε.

3.2.3 Το SVM ως ποινικοποιημένη μέθοδος

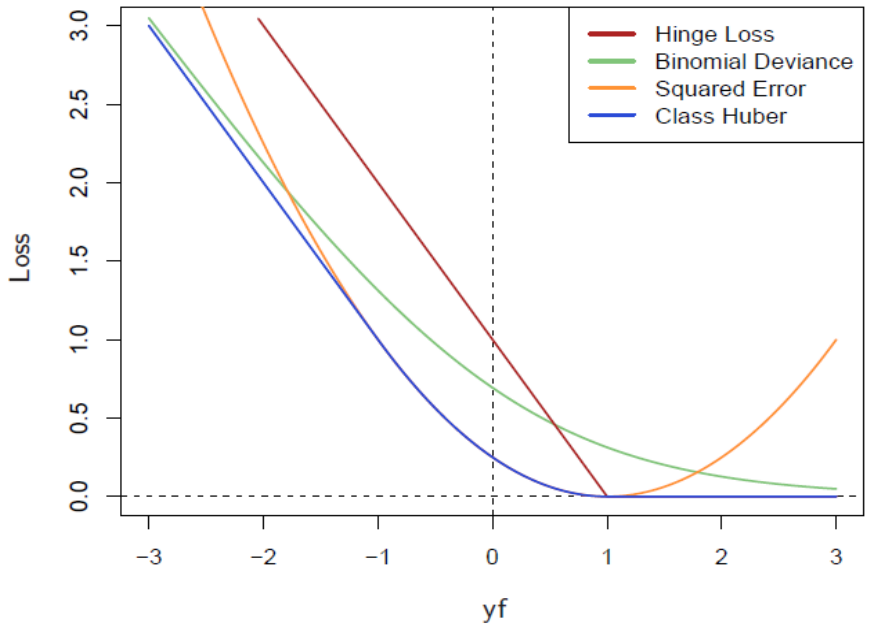
Με το $f(x) = h(x)^T\beta + \beta_0$, θεωρούμε το πρόβλημα βελτιστοποίησης

$$\min_{\beta_0,\beta} \sum_{i=1}^{N} [1 - y_i f(x_i)]_+ + \lambda\|\beta\|^2 \quad (3.2.11)$$

όπου ο δείκτης "+" υποδηλώνει θετικό τμήμα. Αυτό έχει τη μορφή *απώλεια + ποινή*, το οποίο είναι ένα γνώριμο παράδειγμα στην εκτίμηση συναρτήσεων. Είναι εύκολο να δούμε ότι η λύση για την (3.2.10) με $\lambda = \frac{1}{C}$, είναι ίδια με την (3.2.7).

Η εξέταση της «hinge» απώλειας $L(y, f) = [1 - yf]_+$ δείχνει ότι είναι λογική για την ταξινόμηση δύο τάξεων, όταν συγκρίνεται με άλλες πιο παραδοσιακές συναρτήσεις απώλειας.

Στο Σχήμα 3.8 παρουσιάζεται η συνάρτηση απώλειας των διανυσμάτων υποστήριξης (hinge loss), σε σύγκριση με την συνάρτηση απώλειας της αρνητικής λογαριθμοπιθανοφάνειας (διωνυμική απόκλιση) για τη λογιστική παλινδρόμηση, η απώλεια τετραγωνικού σφάλματος, και μία "Huberized" εκδοχή της τετραγωνικής συνάρτησης hinge loss. Όλα εμφανίζονται σαν συνάρτηση του yf αντί του f, λόγω της συμμετρίας μεταξύ της περίπτωσης $y = 1$ και της $y = -1$.



Σχήμα 3.8: Συναρτήσεις απώλειας

Μπορούμε να χαρακτηρίσουμε αυτές τις λειτουργίες απώλειας από την άποψη του τι εκτιμούν σε επίπεδο πληθυσμού. Θεωρούμε την ελαχιστοποίηση $EL(Y, f\,(X))$. Η απόκλιση και η Huber έχουν τις ίδιες ασύμπτωτες με την απώλεια του SVM, αλλά έχουν στρογγυλοποιηθεί στο εσωτερικό. Όλα κλιμακώνονται να έχουν τον περιορισμό της κλίσης της αριστερής-ουράς του -1.

Η (αρνητική) λογαριθμοπιθανοφάνεια ή διωνυμική απόκλιση έχει παρόμοιες ουρές, όπως η SVM απώλεια (hinge loss[9]), δίνοντας μηδενική ποινή σε σημεία που είναι καλά μέσα στο περιθώριο τους, και μια γραμμική ποινή στα σημεία που είναι στη λάθος πλευρά και πολύ μακριά. Το τετραγωνικό σφάλμα, από την άλλη πλευρά δίνει μια τετραγωνική ποινή, καθώς και τα σημεία μέσα στο ίδιο τους το περιθώριο έχουν επίσης μια ισχυρή επιρροή στο μοντέλο. Η τετραγωνική hinge απώλεια $L(y, f) = [1 - yf]^2{}_+$ είναι σαν την τετραγωνική, εκτός του ότι είναι μηδέν για τα σημεία μέσα στο περιθώριο τους. Αυξάνεται ακόμα τετραγωνικά στην αριστερή ουρά, και θα είναι λιγότερο εύρωστη από ότι η hinge ή η απόκλιση στο να ταξινομηθεί εσφαλμένα παρατηρήσεις. Πρόσφατα οι Rosset και Zhu (2007) πρότειναν μια "Huberized" έκδοση της τετραγωνικής hinge απώλειας, η οποία μετατρέπει ομαλά σε μία γραμμική απώλεια στην $yf = -1$.

[9]Στη μηχανική μάθηση, η «hinge loss» είναι μια συνάρτηση που χρησιμοποιείται για την εκπαίδευση των ταξινομητών . Η «hinge loss» χρησιμοποιείται για «μεγιστοποίηση του περιθωρίου" ταξινόμησης, κυρίως για τις μηχανές διανυσματικής υποστήριξης (SVMs).

Πίνακας 3.1: Συναρτήσεις ελαχιστοποίησης για τις διάφορες συναρτήσεις απώλειας

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	$-4yf(x)$, $yf(x) < -1$ $[1 - yf(x)]_+^2$ otherwise	$f(x) = 2\Pr(Y = +1\|x) - 1$

Ο Πίνακας 3.1 συνοψίζει τα αποτελέσματα. Ενώ η hinge απώλεια εκτιμά τον ταξινομητή $G(x)$ η ίδια, όλες οι άλλες εκτιμούν ένα μετασχηματισμό της τάξης εκ των υστέρων πιθανότητας. Η "Huberized" τετραγωνική hinge απώλεια δίνει ελκυστικές ιδιότητες της λογιστικής παλινδρόμησης (ομαλή συνάρτηση απώλειας, εκτιμήσεις πιθανοτήτων), όπως και η SVM hinge απώλεια (σημεία υποστήριξης).

Στον Πίνακας 3.1 παρουσιάζονται οι συναρτήσεις που ελαχιστοποιούνται για τις διαφορετικές συναρτήσεις απώλειας που παρουσιάστηκαν στο Σχήμα 3.8.

Η λογιστική παλινδρόμηση χρησιμοποιεί τη διωνυμική λογαριθμοπιθανοφάνεια ή απόκλιση. Η γραμμική διακριτή ανάλυση χρησιμοποιεί την απώλεια του τετραγωνικού σφάλματος. Η hinge απώλεια του SVM εκτιμά τη εκ των υστέρων συνάρτηση πιθανότητας, ενώ οι άλλες εκτιμούν ένα γραμμικό μετασχηματισμό αυτών των πιθανοτήτων.

Ο τύπος (3.2.10) ρίχνει το SVM ως τακτοποιημένο πρόβλημα εκτίμησης συνάρτησης, όπου οι συντελεστές της γραμμικής επέκτασης $f(x) = \beta_0 + h(x)^T \beta$ έχουν συρρικνωθεί προς το μηδέν (εκτός από τη σταθερά). Αν το h (x) παριστάνει μια ιεραρχική βάση έχοντας κάποια διατεταγμένη δομή (όπως η διάταξη στην τραχύτητα), τότε η ομοιόμορφη συρρίκνωση γίνεται πιο λογική αν το σκληρότερο h_j στο διάνυσμα h έχει μικρότερη νόρμα. Όλες οι συναρτήσεις απώλειας τετραγωνικού-σφάλματος είναι οι λεγόμενες "μεγιστοποιημένες συναρτήσεις απώλειας περιθωρίου" (Rosset et al., 2004b).

Αυτό σημαίνει ότι εάν τα δεδομένα είναι διαχωρίσιμα, τότε το όριο του β_λ στην (3.2.10), καθώς $\lambda \to 0$ καθορίζει το βέλτιστο διαχωριστικό υπερεπίπεδο.

Γενική L1-νόρμα στις μηχανές διανυσματικής υποστήριξης για την επιλογή χαρακτηριστικών

Έχει αποδειχθεί από τους Nguyen et al.(2011) ότι η παραδοσιακή L1-νόρμα SVM που προτάθηκε από τους Bradley και Mangasarian (1998) μπορεί να γενικευθεί σε μια γενική L1-νόρμα SVM (GL1-SVM).

Επιπλέον, έχει αποδειχθεί ότι η επίλυση του νέου προτεινόμενου προβλήματος βελτιστοποίησης (GL1-SVM) δίνει μικρότερο σφάλμα ποινής και διευρύνει το περιθώριο μεταξύ των δύο υπερεπιπέδων των διανυσμάτων υποστήριξης, έτσι δίνει ίσως την καλύτερη ικανότητα γενίκευσης των SVM από την επίλυση της παραδοσιακής L1 νόρμας SVM. Η GL1-SVM μπορεί επίσης να αντιμετωπιστεί ως μια ειδική περίπτωση ορισμένων γενικών επιλογών χαρακτηριστικών (Nguyen et al. (2010)).

3.2.4 Πυρήνες

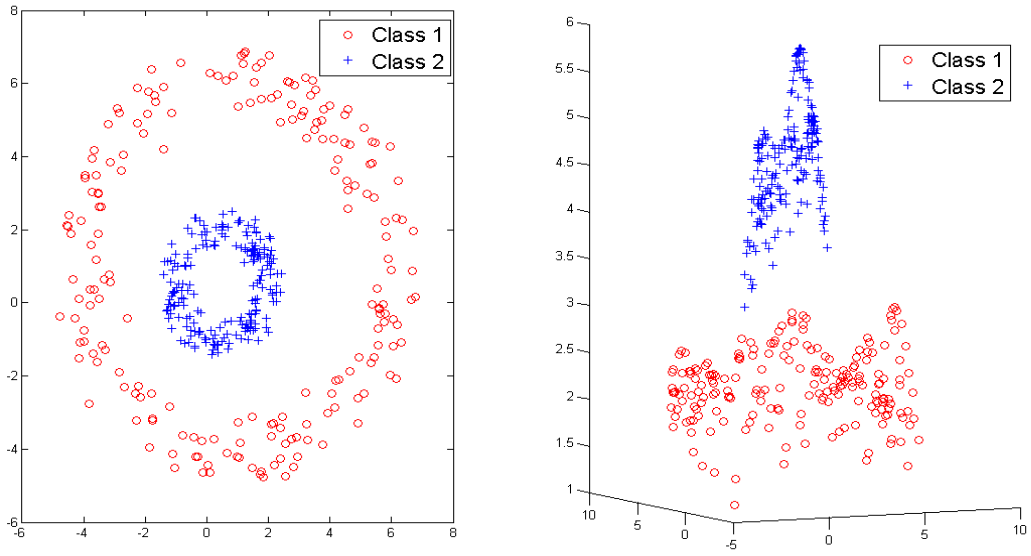
Οι μέθοδοι των πυρήνων είναι μία πολύ δημοφιλής και επιτυχημένη περιοχή της μηχανικής μάθησης. Η κοινή βάση τους είναι το αποκαλούμενο κόλπο του πυρήνα (kernel trick), το οποίο μπορεί να εφαρμοστεί σε οποιονδήποτε γραμμικό αλγόριθμο ο οποίος βασίζεται μόνο στα δεδομένα από την άποψη των εσωτερικών γινομένων μεταξύ δύο παραδειγμάτων.

Κατά την εφαρμογή της SVM τεχνικής για γραμμικά διαχωρίσιμα δεδομένα είχαμε ξεκινήσει δημιουργώντας ένα πίνακα Η από το γινόμενο των μεταβλητών εισόδου:

$$H_{ij} \equiv y_i y_j \, k\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) = \boldsymbol{x}_i \cdot \boldsymbol{x}_j = \boldsymbol{x}_i{}^{\boldsymbol{T}} \, \boldsymbol{x}_j \tag{3.2.12}$$

Η $k\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)$ είναι ένα παράδειγμα μιας οικογένειας συναρτήσεων που ονομάζονται συναρτήσεις πυρήνα (kernel functions) (ο $k\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) = \boldsymbol{x}_i{}^{\boldsymbol{T}} \, \boldsymbol{x}_j$ είναι γνωστός ως γραμμικός πυρήνας). Το σύνολο των συναρτήσεων του πυρήνα αποτελείται από παραλλαγές του (3.2.13) με την έννοια ότι όλα βασίζονται στον υπολογισμό εσωτερικών γινομένων των δύο διανυσμάτων.

Αυτό σημαίνει ότι αν οι συναρτήσεις μπορούν να αναδιατυπωθούν σε ένα χώρο υψηλότερης διάστασης από κάποια πιθανά μη-γραμμικά χαρακτηριστικά χαρτογράφησης της συνάρτησης $x \to \varphi(x)$, μόνο τα εσωτερικά γινόμενα της αντίστοιχης εισόδου στο χώρο των χαρακτηριστικών χρειάζεται να καθοριστούν, χωρίς να χρειάζεται να υπολογιστεί ρητά η φ. Ο λόγος που αυτό το τέχνασμα του πυρήνα είναι χρήσιμο είναι ότι υπάρχουν πολλά προβλήματα ταξινόμησης/ παλινδρόμησης που δεν είναι γραμμικά διαχωρίσιμα στο χώρο των εισόδων x, τα οποία μπορεί να είναι σε ένα υψηλότερης διάστασης χώρο χαρακτηριστικών δεδομένης μιας κατάλληλης χαρτογράφησης $x \to \varphi(x)$. Για περισσότερες λεπτομέρειες σχετικά με τις συναρτήσεις πυρήνων στην ταξινόμηση παραπέμπουμε στον Herbrich (2002).



Σχήμα 3.9: Διχοτόμηση δεδομένων, ανασχηματισμός με τη χρήση του πυρήνα RBF**.**

Αναφερόμενοι στο Σχήμα 3.9, αν ορίσουμε τον πυρήνα να είναι:

$$k\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) = e^{-\left(\frac{\left\|x_i - x_j\right\|^2}{2\,\sigma^2}\right)} \tag{3.2.13}$$

τότε ένα σύνολο δεδομένων το οποία δεν είναι γραμμικά διαχωρίσιμο σε ένα δισδιάστατο χώρο δεδομένων **x** (όπως στην αριστερή πλευρά του Σχήμα 3.9) μπορεί να διαχωριστεί στο μη γραμμικό χώρο των χαρακτηριστικών (δεξιά πλευρά του Σχήμα 3.9) έμμεσα από αυτή τη μη- γραμμική συνάρτηση πυρήνα γνωστή ως Ακτινική Βάση Πυρήνα (Radial Basis Kernel).

Άλλοι δημοφιλείς πυρήνες για ταξινόμηση και παλινδρόμηση είναι ο πολυωνυμικός πυρήνας (Polynomial Kernel)

$$k\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) = \left(\boldsymbol{x}_i \cdot \boldsymbol{x}_j + a\right)^{\boldsymbol{b}}$$

και ο σιγμοειδής πυρήνας

$$k\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) = \tanh(a\boldsymbol{x}_i \cdot \boldsymbol{x}_j - b)$$

όπου a και b είναι οι παράμετροι που καθορίζουν τη συμπεριφορά του πυρήνα.

Υπάρχουν συγκεκριμένες απαιτήσεις που πρέπει να ικανοποιεί μία συνάρτηση έτσι ώστε να μπορεί να θεωρηθεί ως συνάρτηση πυρήνα, που βρίσκονται πέρα από το πεδίο εφαρμογής της παρούσας πολύ σύντομης εισαγωγής στην περιοχή.

Με τη χρήση των συναρτήσεων πυρήνα και την εφαρμογή τους στην συνάρτηση που θέλουμε να βελτιστοποιήσουμε παίρνουμε την ακόλουθη μορφή για τη μεταβλητή w

$$\boldsymbol{w} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} a_i \, y_i \varphi(x_i)$$

Η b μπορεί να καθοριστεί από τις ΚΚΤ συνθήκες. Τα σημεία των δεδομένων που έχουν μη μηδενικά a_i καλούνται διανύσματα υποστήριξης. Τέλος η συνάρτηση απόφασης για το SVM δίνεται από την συνάρτηση:

$$f(x) = sign\,(w * \Phi(x) + b) = sign\,(\textstyle\sum_{i=1}^{n} a_i y_i \, K\,(x, x_i) + b) \qquad (3.2.14)$$

3.2.5 Επιλογή παραμέτρων για τις SVM

Η απόδοση των μηχανών διανυσματικής υποστήριξης (SVM) επηρεάζεται σημαντικά από τις παραμέτρους του μοντέλου. Μία κοινώς χρησιμοποιούμενη μέθοδος επιλογής παραμέτρων SVM, είναι το πλέγμα αναζήτησης (GS), η οποία είναι πολύ χρονοβόρα. Έχουν γίνει αρκετές προσπάθειες από διάφορους ερευνητές για την μείωση του υπολογιστικού κόστους. Οι Ou et al. (2003) πρότειναν ένα μηχανισμό για τη μείωση των δεδομένων με σκοπό την επίσπευση της διαδικασίας επιλογής μοντέλου στην SVM μέθοδο. Τα πειραματικά αποτελέσματα δείχνουν ότι ο προτεινόμενος μηχανισμός είναι σε θέση να μειώσει σημαντικά το χρόνο για να πραγματοποιηθεί η επιλογή μοντέλου με το ελάχιστο κόστος. Τον επόμενο χρόνο, οι Zhu et al. (2004), μέσω της εισαγωγής ενός ενιαίου σχεδιασμού (UD, uniform design) αλλά και με τη χρήση της μεθόδου παλινδρόμησης των μηχανών διανυσματικής υποστήριξης (SVR) κατάφεραν να μειώσουν το κόστος υπολογισμού της παραδοσιακής μεθόδου GS. Μία άλλη προσέγγιση του ίδιου προβλήματος έγινε από τους Lebrun et al. (2006) όπου προτείνεται μια νέα μέθοδος μάθησης για την κατασκευή μιας δίτιμης συνάρτησης αποφάσης (Binary Decision function (BDF)) στις μηχανές διανυσματικής υποστήριξης (SVMs) μειώνοντας την πολυπλοκότητα και καθιστώντας αποτελεσματική τη γενίκευση. Στόχος είναι η κατασκευή ενός γρήγορου και αποτελεσματικού SVM ταξινομητή.

3.3 Προσεγγιστικές Μηχανές Διανυσματικής Υποστήριξης (PSVM)

3.3.1 Γραμμικό PSVM

Ο κλασικός αλγόριθμος SVM έχει ως στόχο να βρεθεί το βέλτιστο διαχωριστικό υπερεπίπεδο που διαχωρίζει αποτελεσματικά τα σημεία δεδομένων στις επισημασμένες κλάσεις, οι οποίες στην περίπτωση των δυαδικών δεδομένων είναι προφανώς δύο. Ας θεωρήσουμε, λοιπόν, ότι έχουμε ένα δυαδικό πρόβλημα ταξινόμησης και θέλουμε να ταξινομήσουμε n σημεία στον p −διάστατο χώρο $\boldsymbol{R^p}$ αντιπροσωπεύεται από ένα πίνακα A, $n \times p$ σύμφωνα με την ένταξη του κάθε σημείου $\boldsymbol{A_i}$ σε μία από τις δύο κατηγορίες, όπως καθορίζεται από ένα διαγώνιο πίνακα D.

Κάνοντας μια απλή αναδιατύπωση του κλασικού SVM αλγορίθμου με περιθώριο, οι Fung και Mangasarian (2001) διαμόρφωσαν το λεγόμενο PSVM. Το PSVM κατατάσσει τις παρατηρήσεις σε δύο κατηγορίες με την ανάθεσή τους σε ένα από τα δύο παράλληλα υπερεπίπεδα που ωθούνται να απέχουν όσο το δυνατόν περισσότερο το ένα από το άλλο. Για να δημιουργήσουν το PSVM, οι Fung και Mangasarian χρησιμοποίησαν μια τροποποιημένη έκδοση του αλγορίθμου SVM γνωστή ως L2SVM διαμορφώνοντας το εξής πρόβλημα:

$$\min_{(w,\gamma,y)R^{p+1+n}} \frac{1}{2}(w'w + \gamma^2) + \frac{C}{2}(y'y) \tag{3.3.1}$$

$$s.t. \quad D(Aw - e\gamma) + y \geq e$$

όπου η θετική σταθερά C καθορίζει το trade-off μεταξύ του εμπειρικού σφάλματος και και του όρου πολυπλοκότητας και τα y_i είναι μη αρνητικοί περιορισμοί αφού αν κάθε περιορισμός είναι αρνητικός τότε η αντικειμενική συνάρτηση μπορεί να μειωθεί θέτοντας ότι το $y_i = 0$ ενώ ικανοποιείται παράλληλα και ο παραπάνω ανισοτικός περιορισμός. Αλλάζοντας την ανισότητα σε ισότητα στους περιορισμούς στο παραπάνω πρόβλημα κάποιος μπορεί να αποκτήσει το PSVM ως ακολούθως:

$$\min_{(\boldsymbol{w},\boldsymbol{\gamma},\boldsymbol{y})\boldsymbol{R^{p+1+n}}} \frac{\boldsymbol{1}}{\boldsymbol{2}}\left(\boldsymbol{w'w} + \boldsymbol{\gamma^2}\right) + \frac{\boldsymbol{C}}{\boldsymbol{2}}(\boldsymbol{y'y}) \tag{3.3.2}$$

$$s.t.\ D(Aw - e\gamma) + y = e$$

Η αντικειμενική συνάρτηση έχει τροποποιηθεί για να ελαχιστοποιεί τη σταθμισμένη νόρμα 2 αντί της νόρμας 1 του προβλήματος των μεταβλητών (w, γ). Σημειώνουμε ότι το w είναι ο προσανατολισμός και γ είναι η σχετική θέση προς την αρχή. Η διατύπωση

αυτή, η οποία μπορεί επίσης να ερμηνευθεί ως κανονικοποιημένα ελάχιστα τετράγωνα (Tikhonov και Arsenin (1977)) του συστήματος των γραμμικών εξισώσεων, αν και πολύ απλό αλλάζει τη φύση του προβλήματος βελτιστοποίησης. Η συνδυαστική φύση της εξίσωσης (3.3.1), καθιστά απαγορευτικό να γράψουμε τη ρητή και ακριβή λύση στο πρόβλημα το οποίο μπορεί να επιτευχθεί χρησιμοποιώντας τη μορφοποίηση του SVM (3.3.2). Η κύρια διαφορά είναι ότι τα διαχωριστικά υπερεπίπεδα μετατρέπονται σε «προσεγγιστικά» υπερπίπεδα τα οποία ωθούνται όσο το δυνατόν πιο μακριά.

3.3.2 Μη γραμμικό Proximal Support vector machine (NPSVM)

Για την επέκταση στην περίπτωση του μη γραμμικού ταξινομητή, οι Fung και Mangasarian (2001), απλά τροποποίησαν τους περιορισμούς στο παραπάνω πρόβλημα βελτιστοποίησης ως εξής

$$\min_{(w,\gamma,y)R^{p+1+n}} \frac{1}{2}\left(\boldsymbol{u'u} + \boldsymbol{\gamma^2}\right) + \frac{C}{2}(\boldsymbol{y'y}) \tag{3.3.3}$$

$$s.t. \quad D(K(A,A')Du - e\gamma) + y = e$$

όπου $K(A,A')$ είναι μια συνάρτηση πυρήνα. Ειδικότερα ο $K(A,A')$ είναι ένας πίνακας $\boldsymbol{nxn}$. Στην εξίσωση (3), που αντικατέστησαν τις πρωταρχικές μεταβλητές w με το ισοδύναμο $w = A'Du$. Για τη γραμμική περίπτωση που περιγράφεται παραπάνω στην εξίσωση (3.3.2) ο $K(A,A') = AA'$. Ωστόσο, όταν έχουμε να χειριστούμε προβλήματα ταξινόμησης μεγάλης κλίμακας, μπορεί να βρεθούμε αντιμέτωποι με το πρόβλημα υπολογισμού του αντιστρόφου ενός πίνακα nxn. Ο πίνακας του πυρήνα είναι διάστασης nxn και η αντιστροφή του δεν είναι εφικτή είτε και δεν μπορεί να αποθηκευτεί στις περισσότερες περιπτώσεις. Το γεγονός αυτό έχει οδηγήσει πολλούς ερευνητές να προτείνουν γρήγορους αλγορίθμους που να επιτρέπουν μια εφικτή λύση (για παράδειγμα Xu et al., (2008), Liu et al., (2007)).

3.4 SVM και μη ισορροπημένα δεδομένα

Παρόλο που οι SVMs παράγουν συχνά αποτελεσματικές λύσεις σε ισορροπημένα σύνολα δεδομένων, είναι ιδιαίτερα ευαίσθητες σε μη ισορροπημένα σύνολα και παράγουν υπό-βέλτιστα μοντέλα. Οι Veropoulos et al.(1999), οι Wu και Chang (2003) και οι Akbani et al.(2004) μελέτησαν προσεκτικά αυτό το πρόβλημα και έχουν προτείνει πολλούς πιθανούς λόγους στους οποίους μπορεί να οφείλεται αυτή η ευαισθησία των ταξινομητών SVM, οι οποίοι συζητούνται στη συνέχεια.

3.4.1 Αδυναμία του προβλήματος βελτιστοποίησης του μαλακού περιθωρίου (Weakness of the soft margin optimization problem

Έχει διαπιστωθεί ότι το διαχωριστικό υπερεπίπεδο ενός μοντέλου SVM που αναπτύσσεται με ένα μη ισορροπημένο σύνολο δεδομένων μπορεί να έχει μία κλίση προς την κλάση μειοψηφίας (Veropoulos et al. (1999)), και αυτή η κλίση μπορεί να υποβαθμίσει την απόδοση αυτού του μοντέλου σε σχέση με την κλάση μειοψηφίας. Αυτό το φαινόμενο μπορεί να εξηγηθεί ως ακολούθως.

Ας θυμηθούμε την αντικειμενική συνάρτηση του προβλήματος βελτιστοποίησης μαλακού-περιθωρίου SVM, η οποία δόθηκε σε προηγούμενο κεφάλαιο ως εξής

$$\min\left(\frac{1}{2}\|\boldsymbol{w}\|^2 + C\ \sum_{i=1}^{n}\xi_i\right) \qquad s.t. \qquad y_i\ (w * \Phi(x_i) + b) \geq\ -1 +\ \xi_i\ \ \forall i \qquad (3.3.1)$$

$$\xi_i \geq 0\ i = 1, \ldots, n$$

Το πρώτο μέρος της αντικειμενικής συνάρτησης επικεντρώνεται στη μεγιστοποίηση του περιθωρίου, ενώ το δεύτερο μέρος επιχειρεί να ελαχιστοποιήσει τον όρο της ποινής που συνδέεται με τα μη σωστά ταξινομημένα σημεία, όπου η παράμετρος κανονικοποίησης C μπορεί επίσης να θεωρηθεί ως το αυτό το κόστος των μη σωστά ταξινομημένων σημείων. Επειδή θεωρούμε το ίδιο κόστος για όλα τα παραδείγματα εκπαίδευσης (δηλαδή, την ίδια τιμή C τόσο για τα θετικά όσο και για τα αρνητικά παραδείγματα), προκειμένου να μειωθεί ο όρος ποινής, θα πρέπει να μειωθεί και ο συνολικός αριθμός των μη σωστά ταξινομημένων σημείων. Όταν το σύνολο δεδομένων είναι ισορροπημένο, η πυκνότητα των παραδειγμάτων της κλάσης πλειοψηφίας θα είναι υψηλότερη από την πυκνότητα των παραδειγμάτων της κλάσης μειοψηφίας ακόμα και γύρω από την περιοχή του συνόρου μεταξύ των κλάσεων, όπου το ιδανικό υπερεπίπεδο θα περνάει από μέσα (σε όλο αυτό το κεφάλαιο θεωρούμε την κατηγορία πλειοψηφίας ως την αρνητική κλάση και την κλάση μειοψηφίας ως τη θετική κλάση). Αυτό επισημαίνεται επίσης από το (Wu και Chang (2003)), ότι η χαμηλή παρουσία θετικών παραδειγμάτων τα κάνει να φαίνονται πιο μακριά από το ιδανικό όριο της κλάσης από ότι τα αρνητικά

παραδείγματα. Κατά συνέπεια, προκειμένου να μειωθεί ο συνολικός αριθμός των μη σωστά ταξινομημένων σημείων στη μάθηση με τις SVMs, το υπερεπίπεδο διαχωρισμού μπορεί να μετατοπιστεί προς την κατηγορία της μειοψηφίας. Αυτή η μετατόπιση /ασυμμετρία μπορεί να προκαλέσει την παραγωγή περισσότερων ψευδώς αρνητικών προβλέψεων, γεγονός που μειώνει την απόδοση του μοντέλου για την κλάση μειοψηφίας που είναι η θετική κλάση. Όταν η μη ισορροπημένη κλάση είναι ακραία, οι SVMs μπορούν να παράγουν μοντέλα που έχουν σε μεγάλο βαθμό ασύμμετρα υπερεπίπεδα, που θα μπορούσαν να αναγνωρίζουν ακόμη και όλα τα παραδείγματα ως αρνητικά (Akbani et al. (2004)).

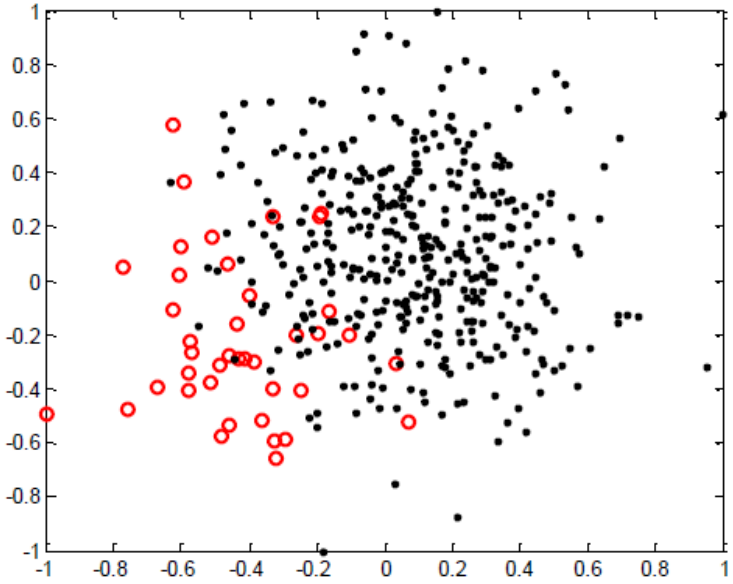
3.4.2 Το μη ισορροπημένο ποσοστό των διανυσμάτων υποστήριξης (The imbalanced support-vector ratio)

Οι Wu και Chang (2003) έχουν εξακριβώσει με πειραματικά αποτελέσματα ότι καθώς τα δεδομένα εκπαίδευσης γίνονται περισσότερο μη ισορροπημένα τόσο η αναλογία μεταξύ θετικών και αρνητικών διανυσμάτων υποστήριξης γίνονται όλο και περισσότερο μη ισορροπημένη. Αυτοί υπέθεσαν ότι ως αποτέλεσμα αυτής της ανισορροπίας η γειτονιά ενός παραδείγματος του συνόλου δοκιμής κοντά στο όριο απόφασης είναι περισσότερο πιθανό να κυριαρχείται από αρνητικά διανύσματα υποστήριξης, και ως εκ τούτου, η συνάρτηση απόφασης είναι πιο πιθανό να ταξινομήσει ένα σημείο που βρίσκεται οντά στο όριο ως αρνητικό. Ωστόσο, οι Akbani et al. (2004) διαφώνησαν με αυτή την ιδέα, επισημαίνοντας ότι, λόγω του περιορισμού $\sum_{i=1}^{n} y_i \alpha_i = 0$, τα α_i από κάθε θετικό διάνυσμα υποστήριξης, τα οποία είναι λιγότερα σε αριθμό από τα αρνητικά διανύσματα υποστήριξης, πρέπει να είναι μεγαλύτερα σε μέγεθος από τις τιμές των α_i που συνδέονται με τα αρνητικά διανύσματα υποστήριξης. Αυτά τα α_i δρουν σαν βάρος στην τελική συνάρτηση απόφασης, και ως εκ τούτου, μεγαλύτερα α_i στα θετικά διανύσματα υποστήριξης λαμβάνουν υψηλότερα βάρη από τα αρνητικά διανύσματα υποστήριξης, κάτι που μπορεί να μειώσει σε κάποιο βαθμό την επίδραση της ανισορροπίας στα διανύσματα υποστήριξης. Οι Akbani et al. (2004) υποστήριξαν περαιτέρω ότι αυτό θα μπορούσε να είναι ο λόγος για τον οποίο οι SVMs δεν λειτουργούν τόσο άσχημα σε σύγκριση με άλλους με άλλους αλγορίθμους της μηχανικής μάθησης για μέτρια μη ισορροπημένα σύνολα δεδομένων.

3.4.3 Αποτελεσματικότητα της εξισορρόπησης των κλάσεων

Για τα μη ισορροπημένα και σε μεγάλο βαθμό επικαλυπτόμενα δεδομένα κλάσης, οι μέθοδοι δειγματοληψίας, όπως η υπερδειγματοληψία ή η υποδειγματοληψία είναι πολύ αποτελεσματική από την άποψη της διαδικασίας βελτιστοποίησης σε ένα SVM με μαλακό περιθώριο. Προκειμένου να απεικονίσουμε το αποτέλεσμα των μεθόδων δειγματοληψίας στην εξισορρόπηση της κατανομής μεταξύ των κλάσεων, εξετάζουμε

την κίνηση του ορίου απόφασης με δύο κοινές μεθόδους, τη SMOTE υπερδειγματοληψία και την τυχαία υποδειγματοληψία. Σε αυτό το παράδειγμα, για την ταξινόμηση με SVM χρησιμοποιήσαμε έναν Γκαουσιανό πυρήνα.



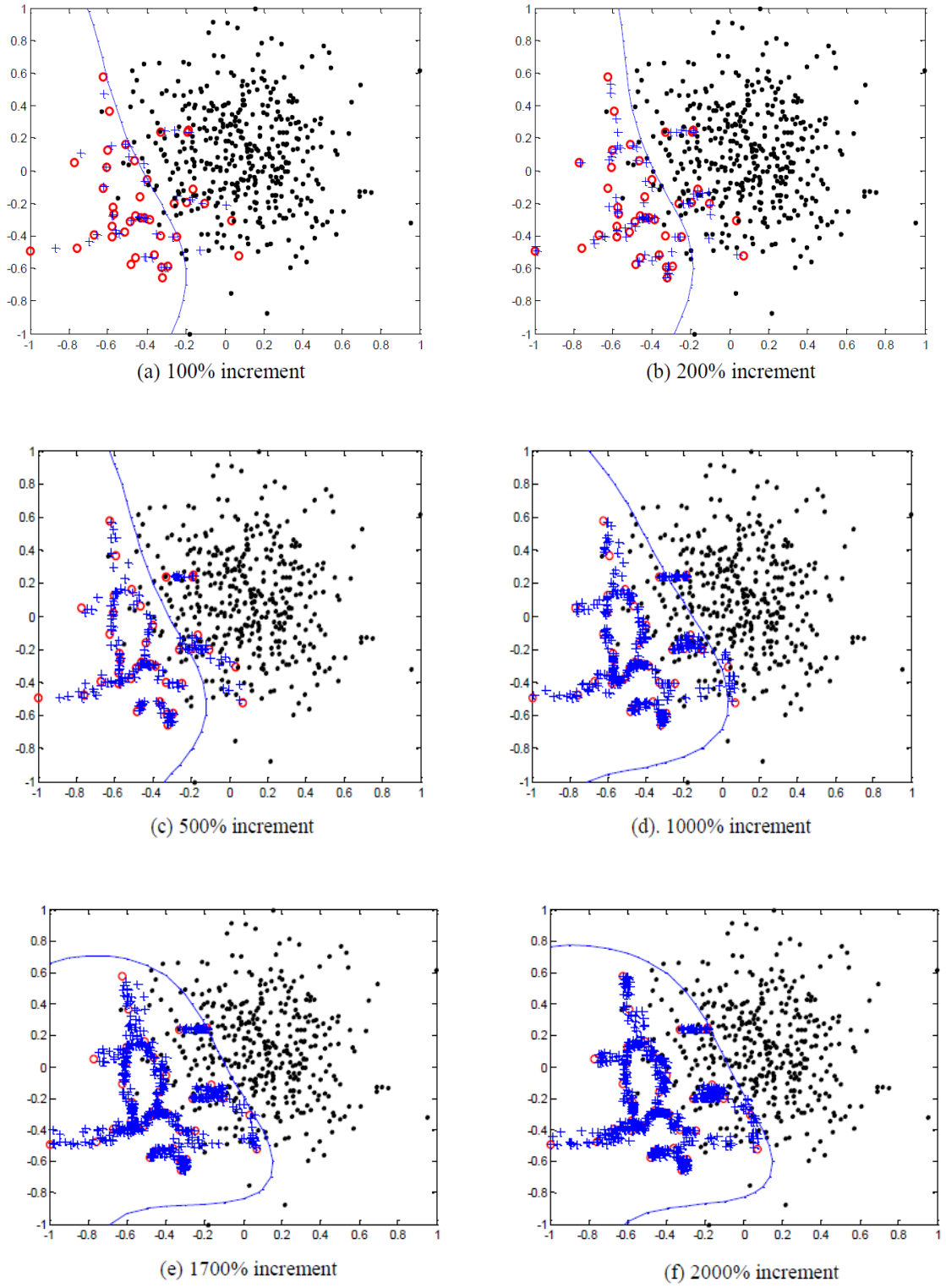
Σχήμα 3.10: Παράδειγμα του προβλήματος ανισορροπίας στις SVM

Πρώτα, δημιουργήσαμε μια απλή δομή ενός συνθετικού συνόλου δεδομένων που δείχνει ένα τυπικό πρόβλημα ανισορροπίας μεταξύ των κλάσεων, η οποία θα μπορούσε να αναπαρασταθεί σε 2-διάστατο χώρο, όπως φαίνεται στο Σχήμα 3.10. Η κλάση μειοψηφίας αποτελείται από 40 περιπτώσεις που σημειώνονται με "o" και η κλάση πλειοψηφίας από 400 περιπτώσεις με "•" (ο λόγος μεταξύ των κλάσεων είναι 1:10).

Μετά την ταξινόμηση, σχεδόν όλες οι περιπτώσεις πλειοψηφίας έχουν ταξινομηθεί σωστά, ενώ πολλές περιπτώσεις της κλάσεως μειοψηφίας ταξινομήθηκαν σαν να ανήκουν στην κλάση πλειοψηφίας. Αυτό το παράδειγμα απεικονίζει ένα τυπικό πρόβλημα ανισορροπίας που προκαλείται από τους αλγορίθμους SVM με μαλακό περιθώριο. Με άλλα λόγια, ένα βέλτιστο υπερεπίπεδο απορέει από το trade-off μεταξύ της μεγιστοποίησης του περιθωρίου της κλάσης μειοψηφίας και της κλάσης πλειοψηφίας και την ελαχιστοποίηση του κόστους εσφαλμένης ταξινόμησης στο χώρο των χαρακτηριστικών. Για να βελτιωθεί η ακρίβεια για την κατηγορία της κλάσης μειοψηφίας, θα πρέπει να μετακινήσουμε το όριο προς την πλευρά της κλάσης πλειοψηφίας. Για να το απεικονίσουμε αυτό, θα χρησιμοποιήσουμε δύο μεθόδους δειγματοληψίας για την εξισορρόπηση, τη SMOTE και την τυχαία υποδειγματοληψία, που συνήθως αναφέρονται ως SVM-SMOTE και SVM-RU, αντίστοιχα.

Χρησιμοποιώντας SVM-SMOTE, ο αριθμός των συνθετικών περιπτώσεων για να επιτευχθεί η επιθυμητή ισορροπία μεταξύ των κλάσεων είναι άγνωστος και γι αυτό το

λόγο πρέπει να διεξάγονται εμπειρικές μελέτες. Στις περιπτώσεις μειοψηφίας εκτελείται σταδιακά υπερδειγματοληψία με 100%, 300%, 500% και 1000% αυξήσεις στις περιπτώσεις μειοψηφίας.



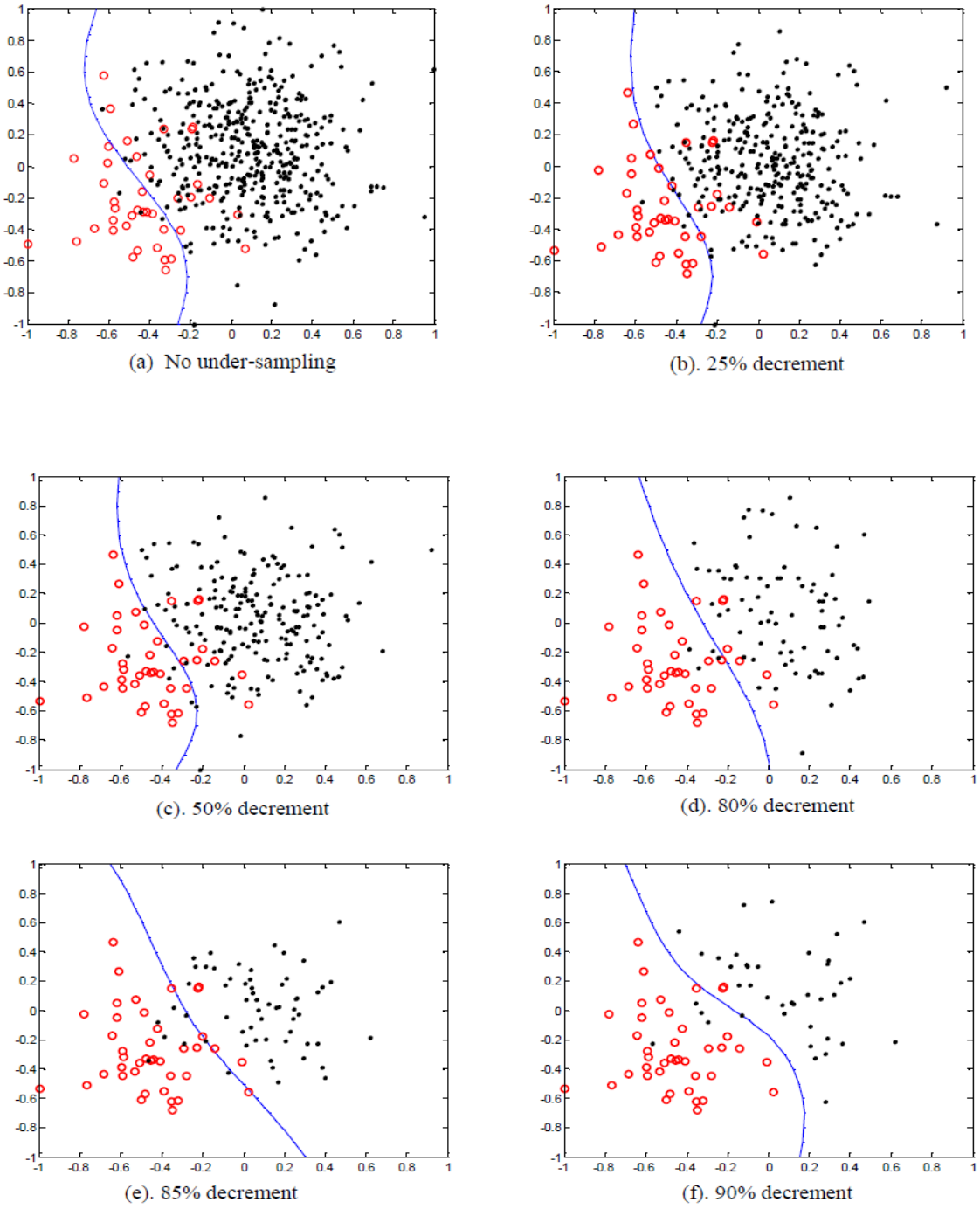
Σχήμα 3.11: Κινήσεις του ορίου απόφασης από τον αλγόριθμο SMOTE

Μετά την εξισορρόπηση με τη μέθοδο SMOTE, παρατηρήσαμε ότι το όριο μετατοπίζεται σταδιακά προς την κλάση πλειοψηφίας καθώς οι περιπτώσεις μειοψηφίας αυξάνονται όπως φαίνεται στο Σχήμα 3.11 (α) έως (στ).

Παρόλο που το SVM-SMOTE μετατοπίζει το όριο απόφασης, έχουμε μια ποινή που αφορά στην αύξηση του μεγέθους του συνόλου δεδομένων, όπως αναφέραμε προηγουμένως. Ας υποθέσουμε ότι Np είναι ο αριθμός των θετικών (μειοψηφία) περιπτώσεων και Nn ο αριθμός των αρνητικών (πλειοψηφία) περιπτώσεων, συνήθως το SVM παίρνει $O((Np+Nn)^3)$ χρόνο για την εκμάθηση στη χειρότερη περίπτωση (Burges, 1998). Για την μάθηση με μη ισορροπημένα δεδομένα, το SMOTE-SVM θα παίρνει $O((Np \times (1+R_{smote})+Nn)^3)$ όπου το R_{smote} είναι το βέλτιστο ποσοστό του μεγέθους των περιπτώσεων που αυξάνονται. Εδώ το R_{smote} καθορίστηκε εμπειρικά. Αυτό που είναι χειρότερο, είναι ότι η υπερδειγματοληψία αυξάνει επίσης τις περιπτώσεις στην περιοχή μεταξύ των τάξεων. Με τη δημιουργία περιπτώσεων κοντά ή σε επικαλυπτόμενες περιοχές που ενδέχεται να ταξινομηθούν εσφαλμένα, η ταξινόμηση είναι πιο δύσκολη. Κατά την επίλυση του προβλήματος βελτιστοποίησης στον αλγόριθμο SVM, πολλές περιπτώσεις μπορούν να παραβιάζουν τις συνθήκες ΚΚΤ. Ως εκ τούτου, αυτό θα απαιτεί πολύ περισσότερο χρόνο για τη σύγκλιση της βελτιστοποίησης σε αλγόριθμους SVM, παρά τις καλές επιδόσεις τους. Έτσι, εάν ένα σύνολο δεδομένων είναι εξαιρετικά μη ισορροπημένο και οι κλάσεις επικαλύπτονται σε μεγάλο βαθμό, η υπερδειγματοληψία μέσω του αλγορίθμου SMOTE δεν θα είναι αποτελεσματική. Όσον αφορά στα προβλήματα που εισήχθησαν με την προσέγγιση της υπερδειγματοληψίας, συνήθως προτιμάται και χρησιμοποιείται η μέθοδος της υπο-δειγματοληψίας αντί των μεθόδων υπερδειγματοληψία.

Η SVM-RU χρησιμοποιεί την τυχαία υπόδειγματοληψία για να εξισορροπήσει την κατανομή των κλάσεων. Αντί να αυξήσει τις περιπτώσεις μειοψηφίας, μειώνοντας τον αριθμό των περιπτώσεων πλειοψηφίας θα κατασκευάσει ένα βέλτιστο υπερεπίπεδο από την άποψη της απόστασης (trade-off) μεταξύ της μεγιστοποίησης του περιθωρίου και της ελαχιστοποίησης του κόστους εσφαλμένης ταξινόμησης. Παρακάτω, στο Σχήμα 3.5 φαίνεται η κίνηση του ορίου καθώς ο αριθμός των περιπτώσεων πλειοψηφίας που αφαιρούνται τυχαία, αυξάνεται.

Παρόμοια με το SVM-SMOTE, η τυχαία υποδειγματοληψία προκαλεί μια μετατόπιση στο όριο απόφασης προς την κλάση πλειοψηφίας. Λαμβάνοντας υπόψη την πολυπλοκότητα του χρόνου για την εκμάθηση του συνόλου δεδομένων, η SVM-RU παίρνει $O\left((Np+Nn\times R_u)^3\right)$ η οποία είναι ταχύτερη από ότι ο SVM με το αρχικό σύνολο εκπαίδευσης αφού το $Nn\times R_u$ είναι περίπου ίσο με Nn. Επειδή οι περιπτώσεις πλειοψηφίας εξαλείφθησαν τυχαία, μπορεί να μην είναι εύκολο να προσδιοριστεί το βέλτιστο μέγεθος του συνόλου εκπαίδευσης για την εξισορρόπηση. Επίσης, παρόμοια με SVM-SMOTE, η επιθυμητή βέλτιστη κατανομή μεταξύ της κατανομής των κλάσεων για τη μη ισορροπημένη μάθηση των δεδομένων είναι άγνωστη και πρέπει να προσδιοριστεί εμπειρικά.



Σχήμα 3.12: Κινήσεις του ορίου απόφασης με τυχαία υποδειγματοληψία

3.4.4 Υποθέσεις

Δεδομένης της φύσης των μη ισορροπημένων συνόλων δεδομένων, δύο υποθέσεις διαμορφώθηκαν για να αντιμετωπίσουν τα ζητήματα της αποτελεσματικότητας και της αποδοτικότητας στη μάθηση με μη ισορροπημένα δεδομένα για SVMs.

Υπόθεση 1

Ένας σχετικά μικρός αριθμός των περιπτώσεων από ένα μη ισορροπημένο σύνολο εκπαίδευσης είναι αναγκαίος για την απόκτηση καλών επιδόσεων στην επίλυση του προβλήματος της ανισορροπίας μεταξύ των κλάσεων χρησιμοποιώντας ένα SVM.

Υπόθεση 2

Ένα μικρότερο υποσύνολο μέσα στο σύνολο των διανυσμάτων υποστήριξης, μπορεί να βρεθεί, που να παράγει ένα καλύτερο όριο χρησιμοποιώντας ένα SVM.

Η υπόθεση 1 σχετίζεται με την προϋπόθεση ότι το όριο μεταξύ των κλάσεων μειοψηφίας και πλειοψηφίας είναι έντονα επηρεασμένο από ένα σχετικά μικρό αριθμό περιπτώσεων. Το συμπέρασμα είναι ότι μία πιθανή μείωση του μεγέθους των συνόλων μάθησης θα οδηγήσει σε σημαντική βελτίωση στην απόδοση της ταξινόμησης. Η υπόθεση 2 βασίζεται στην προσδοκία ότι υπάρχει μια μικρότερη ομάδα διανυσμάτων υποστήριξης, που θα παρέχει ένα κοντινό βέλτιστο στην ταξινόμηση των κλάσεων και βελτιώνει την αποτελεσματικότητα της μάθησης. Αν ένα τέτοιο σύνολο διανυσμάτων υποστήριξης υπάρχει, τότε μια metaheuristic προσέγγιση δειγματοληψίας μπορεί να χρησιμοποιηθεί για την επιλογή των διανυσμάτων αυτών.

Παρακάτω παραθέτουμε κάποιες από τις τεχνικές που χρησιμοποιούνται για το πρόβλημα τις ανισσοροπίας τόσο σε επίπεδο δεδομένων όσο και σε αλγοριθμικό επίπεδο και κάνουν χρήση των Μηχανών Διανυσματικής Υποστήριξης.

3.5 Εξωτερική μάθηση μη ισορροπημένων δεδομένων για τις SVMs: Μέθοδοι προεπεξεργασίας των δεδομένων

(External imbalance learning methods for SVMs:Data preprocessing methods)

3.5.1 Resampling methods

Όλες οι μέθοδοι προεπεξεργασίας των δεδομένων που συζητήθηκαν στο Κεφάλαιο 2 της παρούσας εργασίας μπορεί να χρησιμοποιηθούν για την εξισορρόπηση των συνόλων δεδομένων πριν την εκπαίδευση ενός μοντέλου SVM. Αυτές οι μέθοδοι περιλαμβάνουν τυχαίες και εστιασμένες μεθόδους υπό/υπέρ-δειγματοληψίας καθώς και μεθόδους παραγωγής συνθετικών δεδομένων, όπως η μέθοδος SMOTE (Chawla et al.(2002)). Οι μέθοδοι επαναδειγματοληψίας έχουν εφαρμοστεί με επιτυχία στην εκπαίδευση των SVMs με μη ισορροπημένα σύνολα δεδομένων σε διαφορετικά πεδία (Akbani et al. (2004), Chawla et al. (2002), Chen et al. (2004), Fu et al. (2004), Lessmann (2004)).

Ειδικότερα, οι Batuwita και Palade (2010), παρουσιάζουν μια αποτελεσματική εστιασμένη μέθοδο υπερδειγματοληψίας για τις SVMs. Σε αυτή τη μέθοδο, αρχικά το διαχωριστικό υπερεπίπεδο που βρέθηκε από την εκπαίδευση ενός μοντέλου SVM στα αρχικά μη ισορροπημένα δεδομένα χρησιμοποιήθηκε για την επιλογή των πιο πληροφοριακών παραδειγμάτων για ένα δεδομένο πρόβλημα ταξινόμησης, τα οποία είναι τα σημεία δεδομένων που βρίσκονται γύρω από την περιοχή του ορίου της κλάσης. Στη συνέχεια, μόνο αυτά τα επιλεγμένα παραδείγματα εξισορροπούνται με υπερδειγματοληψία αντί να πραγματοποιηθεί υπερδειγματοληψία σε ολόκληρο το σύνολο δεδομένων. Αυτή η μέθοδος μειώνει σημαντικά το χρόνο εκπαίδευσης της SVM, ενώ παράλληλα λαμβάνει συγκρίσιμα αποτελέσματα ταξινόμησης με την αρχική μέθοδο υπερδειγματοληψίας.

Η μέθοδος των μηχανών διανυσματικής συσταδοποίησης (Support Cluster Machines, SCMs) που παρουσιάζεται από τους Yuan et al. (2006) μπορεί να θεωρηθεί ως μία άλλη εστιασμένη μέθοδος επαναδειγματοληψίας για τις SVMs. Αυτή η μέθοδος πρώτα χωρίζει τα αρνητικά παραδείγματα σε ασυνεχείς συστάδες χρησιμοποιώντας τη μέθοδο συσταδοποίησης kernel-k-means. Στη συνέχεια, εκπαιδεύει ένα αρχικό μοντέλο SVM χρησιμοποιώντας τα θετικά παραδείγματα και τους εκπροσώπους των αρνητικών συστάδων, δηλαδή, τα παραδείγματα που αντιπροσωπεύουν τα κέντρα των συστάδων. Με την συνολική εικόνα των αρχικών SVMs, η μέθοδος προσδιορίζει περίπου τα διανύσματα υποστήριξης και μη-υποστήριξης. Τότε χρησιμοποιείται μια τεχνική συρρίκνωσης για την αφαίρεση των δειγμάτων που πιθανότατα δεν είναι διανύσματα υποστήριξης. Αυτή η διαδικασία της ομαδοποίησης και της συρρίκνωσης εκτελείται επαναληπτικά αρκετές φορές μέχρι τη σύγκλιση.

3.5.2 Ensemble learning methods

Οι Ensemble μέθοδοι μάθησης έχουν εφαρμοστεί ως λύση για την εκπαίδευση των SVMs με μη ισορροπημένα σύνολα δεδομένων (Lin et al. (2009), Kang and Cho (2006), Liu et al. (2006), Wang and Japkowicz (2010)). Γενικά, σε αυτές τις μεθόδους, η κλάση πλειοψηφίας διαχωρίζεται σε πολλαπλά υποσύνολα δεδομένων τέτοια ώστε κάθε ένα από αυτά τα υποσύνολα δεδομένων έχει ένα παρόμοιο αριθμό παραδειγμάτων ως το σύνολο δεδομένων της κλάσης μειοψηφίας. Αυτό μπορεί να γίνει με τυχαία δειγματοληψία με επανάθεση ή χωρίς (bootstrapping), είτε μέσω μεθόδων συσταδοποίησης. Στη συνέχεια, αναπτύσσεται ένα σύνολο από ταξινομητές SVM έτσι ώστε ο καθένας να έχει εκπαιδευτεί με το ίδιο θετικό σύνολο δεδομένων και ένα διαφορετικό αρνητικό υποσύνολο δεδομένων. Τέλος, οι αποφάσεις λαμβάνονται από το σύνολο των ταξινομητών που συνδυάζονται με τη χρήση μιας μεθόδου όπως είναι η μέθοδος ψηφοφορίας της πλειοψηφίας. Επιπρόσθετα, ειδικοί boosting αλγόριθμοι, όπως ο Adacost (Fan et al.(1999)), ο RareBoost (Joshi et al.(2001)) και ο SMOTEBoost (Chawla et al. (2004)), οι οποίοι έχουν χρησιμοποιηθεί στην μάθηση με μη ισορροπημένα δεδομένα, θα μπορούσαν επίσης να εφαρμοστούν με τις SVMs.

3.6 Εσωτερική μάθηση μη ισορροπημένων δεδομένων για τις SVMs: Αλγοριθμικές Μέθοδοι
(Internal imbalance learning methoods for SVMs: Algorithmic methods)

Στην ενότητα αυτή θα παρουσιάσουμε τις αλγοριθμικές τροποποιήσεις που προτείνονται στην βιβλιογραφία για να κάνουν τον αλγόριθμο SVM λιγότερο ευαίσθητο στην ανισορροπία μεταξύ των κλάσεων.

3.6.1 Different Error Costs (DEC)
Cost sensitivity SVM (TCSVM) for imbalanced data

Όπως έχουμε επισημάνει στο κεφάλαιο 3.3, ο κύριος λόγος που κάνει τον αλγόριθμο SVM ευαίσθητο στο πρόβλημα της ανισορροπίας μεταξύ των κλάσεων είναι ότι η αντικειμενική συνάρτηση με το μαλακό περιθώριο που δίνεται στην εξίσωση (3.3.1) αποδίδει το ίδιο κόστος (δηλαδή, το ίδιο C) τόσο για τα θετικά όσο και και αρνητικά εσφαλμένα ταξινομημένα παραδείγματα στον όρο ποινής. Αυτό θα μπορούσε να προκαλέσει μία κλίση του διαχωριστικού υπερεπιπέδου προς την κλάση μειοψηφίας, η οποία θα αποδώσει τελικά ένα υποβέλτιστο μοντέλο. Η μέθοδος DEC είναι μια μέθοδος μάθησης ευαίσθητου κόστους που προτείνεται από τους (Veropoulos et al.(1999)), για να ξεπεραστεί αυτό το πρόβλημα στις SVMs. Σε αυτή τη μέθοδο, η αντικειμενική συνάρτηση ενός SVM με μαλακό περιθώριο έχει τροποποιηθεί αποδίδοντας δύο κόστη εσφαλμένης ταξινόμησης έτσι ώστε C^+ να είναι το κόστος εσφαλμένης ταξινόμησης για την κλάση των θετικών παραδειγμάτων, ενώ C^- είναι το κόστος εσφαλμένης ταξινόμησης για την κλάση των αρνητικών παραδειγμάτων, όπως αναφέρονται στην εξίσωση (3.5.1). Εδώ υποθέτουμε επίσης ότι η θετική κλάση είναι η κλάση της μειοψηφίας και η αρνητική κλάση είναι η τάξη πλειοψηφίας.

$$minimize\ \{\frac{1}{2}\|\boldsymbol{w}\|^2 + C^+ \sum_{\{i|\ y_i=+1\}}^{n} \xi_i + C^- \sum_{\{i|\ y_i=-1\}}^{n} \xi_i$$

$$subject\ to \qquad [\ y_i\ (\boldsymbol{x_i} \cdot \boldsymbol{w} + b) - 1 + \xi_i\ ] \geq 0 \qquad (3.5.1)$$

$$\xi_i \geq 0, \quad i = 1, \ldots, n$$

Με την ανάθεση υψηλότερου κόστος εσφαλμένης ταξινόμησης για τα παραδείγματα της κλάσης μειοψηφίας από ότι τα παραδείγματα της κλάσης πλειοψηφίας (δηλ. $C^+ > C^-$), μπορούν να μειωθούν οι επιπτώσεις που επιφέρει η ανισορροπία των κλάσεων. Δηλαδή, ο τροποποιημένος αλγόριθμος SVM δεν έχει την τάση να παραποιήσει το διαχωριστικό υπερεπίπεδο προς τα παραδείγματα της κλάσης μειοψηφίας για να μειώσει το σύνολο

των εσφαλμένων ταξινομήσεων, καθώς στα παραδείγματα της κλάσης μειοψηφίας τώρα ανατήθεται υψηλότερο κόστος εσφαλμένης ταξινόμησης.

Το Lagrangian πρόβλημα αυτής της τροποποιημένης αντικειμενικής συνάρτησης μπορεί να παρασταθεί ως εξής:

$$L_P \equiv \frac{1}{2}\|\boldsymbol{w}\|^2 + C^+ \sum_{\{i|\, y_i=+1\}}^{n} \xi_i + C^- \sum_{\{i|\, y_i=-1\}}^{n} \xi_i$$
$$- \sum_{i=1}^{n} a_i \left[y_i \left(\boldsymbol{x_i} \cdot \boldsymbol{w} + b\right) - 1 + \xi_i \right] - \sum_{i=1}^{n} \mu_i \xi_i ,$$

όπου $\mu_i \geq 0$ and $\alpha_i \geq 0$.

Η δυϊκή τυποποίηση δίνει την Langrangian

$$L_D \equiv \sum_{i=1}^{n} a_i - \frac{1}{2}\sum_{i,j} a_i a_j y_i y_j K(x_i, x_j) \tag{3.5.2}$$

Υπό τους περιορισμούς

$$0 \leq a_i^+ \leq C^+, \quad \alpha\nu\, y_i = +1$$

και

$$0 \leq a_i^- \leq C^-, \quad \alpha\nu\, y_i = -1$$

Όπου a_i^+ και a_i^- αντιπροσωπεύουν τους πολλαπλασιαστές Lagrangian των θετικών και αρνητικών παραδειγμάτων, αντίστοιχα.

Αυτό το δυϊκό πρόβλημα βελτιστοποίησης μπορεί να λυθεί κατά τον ίδιο τρόπο όπως και στην επίλυση του προβλήματος βελτιστοποίησης των κανονικών SVM. Οι Akbani et al.(2004), έχουν αναφέρει ότι αρκετά καλά αποτελέσματα ταξινόμησης μπορούν να προκύψουν θέτοντας $\frac{C^-}{C^+}$ ίσο με την αναλογία μεταξύ της κλάσης μειοψηφίας και της κλάσης πλειοψηφίας.

3.6.2 One class learning

Οι Raskutti and Kowalczyk (2004), Kowalczyk and Raskutti (2002) παρουσίασαν δύο ακραίες μεθόδους εξισορρόπησης για την εκπαίδευση των SVMs με εξαιρετικά μη ισορροπημένα σύνολα δεδομένων. Στην πρώτη μέθοδο εκπαιδεύεται ένα μοντέλο SVM μόνο με τα παραδείγματα της κλάσης μειοψηφίας. Στη δεύτερη μέθοδο, έχει επεκταθεί η μέθοδος DEC αναθέτοντας ένα κόστος $C^- = 0$ για τα παραδείγματα της κλάσης πλειοψηφίας και $C^+ = \frac{1}{N_+}$ για την κλάση της μειοψηφίας, όπου N_+ είναι ο αριθμός των παραδειγμάτων της κλάσης μειοψηφίας.

Από τα πειραματικά αποτελέσματα που ελήφθησαν σε διάφορα μη ισορροπημένα συνθετικά αλλά και πραγματικά σύνολα δεδομένων, παρατηρήθηκε ότι οι μέθοδοι αυτές είναι πιο αποτελεσματικές από τις γενικές μεθόδους αναπροσαρμογής των δεδομένων.

3.6.3 z-SVM

Το zSVM είναι μία άλλη αλγοριθμική τροποποίηση για τις SVMs που προτάθηκε από τους Imam et al. (2006) για την μάθηση από μη ισορροπημένα σύνολα δεδομένων. Σε αυτή τη μέθοδο, πρώτα αναπτύσσεται ένα μοντέλο SVM χρησιμοποιώντας το αρχικό μη ισορροπημένο σύνολο εκπαίδευσης. Στη συνέχεια, τροποποιείται το όριο απόφασης του μοντέλου που προέκυψε έτσι ώστε να αφαιρεθεί η μεροληψία του προς την κλάση πλειοψηφίας (αρνητική) κλάση. Ας σκεφτούμε την τυπική συνάρτηση απόφασης ενός SVM η οποία δίνεται στην εξίσωση (3.2.14), η οποία μπορεί να ξαναγραφτεί ως εξής:

$$f(x) = sign\,(w * \Phi(x) + b) = sign\,(\sum_{i=1}^{n} a_i y_i\, K\,(x, x_i) + b)$$

$$= sign\left(\sum_{i=1}^{n_1} a_i^+ y_i\, K\,(x, x_i) + \sum_{j=1}^{n_2} a_j^- y_j\, K\left(x, x_j\right) + b\right) \quad (3.5.3)$$

όπου a_i^+ είναι οι συντελεστές των θετικών διανυσμάτων υποστήριξης a_j^- είναι οι συντελεστές των αρνητικών διανυσμάτων υποστήριξης, και τα n_1 και n_2 αντιπροσωπεύουν τον αριθμό των θετικών και αρνητικών παραδειγμάτων εκπαίδευσης, αντίστοιχα. Στη μέθοδο zSVM, το μέγεθος των a_i^+ τιμών των θετικών διανυσμάτων υποστήριξης αυξάνεται πολλαπλασιάζοντας όλα αυτά με μία συγκεκριμένη μικρή z θετική τιμή. Τότε, η τροποποιημένη συνάρτηση απόφασης του ταξινομητή SVM μπορεί να παρασταθεί ως εξής:

$$f(x) = sign\left(z * \sum_{i=1}^{n_1} a_i^+ y_i\, K\,(x, x_i) + \sum_{j=1}^{n_2} a_j^- y_j\, K\left(x, x_j\right) + b\right)$$

Αυτή η τροποποίηση a_i^+ θα αυξήσει τα βάρη των θετικών διανυσμάτων υποστήριξης στη συνάρτηση απόφασης, και ως εκ τούτου θα μειώσει την μεροληψία της προς την αρνητική πλειοψηφική κλάση. Στους Imam et al. (2006), επιλέχθηκε ως βέλτιστη η τιμή του z που δίνει το καλύτερο αποτελέσματα της ταξινόμησης για το σύνολο δεδομένων εκπαίδευσης. Τέλος αξίζει να αναφέρουμε τη μορφή που λαμβάνει η μεταβλητή w στην περίπτωση του zSVM. Είναι γνωστό ότι η εξίσωση για το διάνυσμα w εξαρτάται από τα διανύσματα υποστήριξης. Επομένως έχουμε:

$$w = \sum a_i y_i x_i$$

όπου a_i είναι οι θετικοί πολλαπλασιαστές Langrange. Για το z-svm η προηγούμενη σχέση παίρνει τη μορφή:

$$\boldsymbol{w} = \sum_{i|y_i=1}^{n_1} a_i y_i x_i + z * \sum_{i|y_i=-1}^{n_2} a_i y_i x_i$$

Οι συγγραφείς πρότειναν ότι η βέλτιστη τιμή για το z είναι εκείνη που μεγιστοποιεί την τιμή του γεωμετρικού μέσου του ποσοστού των ορθώς ταξινομημένων παρατηρήσεων στις δύο κλάσεις.

3.6.4 Modified proximal SVM

Ένα μειονέκτημα των τυποποιημένων αλγορίθμων SVM, συμπεριλαμβανομένου και του PSVM, είναι η αναποτελεσματικότητα χειρισμού μη ισορροπημένων συνόλων δεδομένων όπου μία τάξη έχει περισσότερες παρατηρήσεις από την άλλη υποθέτοντας ότι έχουμε ένα δυαδικό πρόβλημα ταξινόμησης. Σε τέτοιες περιπτώσεις, τα αποτελέσματα είναι συχνά μεροληπτικά υπερ της μίας κλάσης, όπως πολλάκις έχουμε αναφέρει σε προηγούμενα κεφάλαια, με την απόδοση να επηρεάζεται σημαντικά από την κλάση μειοψηφίας. Ως εκ τούτου, αναθέτοντας την ίδια ποινή εσφαλμένης ταξινόμησης και για τις δύο κατηγορίες, παράγουμε μοντέλα που είναι μεροληπτικά προς την κατηγορίας πλειοψηφίας κάτι που σημαίνει ότι οι SVM τείνουν να εξουδετερώνουν την πλειοψηφική κλάση. Το πρόβλημα των μη ισορροπημένων δεδομένων έχει μελετηθεί από πολλούς ερευνητές οι οποίοι πρότειναν διάφορες λύσεις (για λεπτομέρειες δείτε Chawla et al., (2004)). Στη μελέτη αυτή παρουσιάζουμε μία τροποποίηση της PSVM που πρότειναν οι Tao et al. (2007), η οποία συνδυάζει την ιδέα του γραμμικού PSVM και την ιδέα του TCSVM που πρότειναν οι Veropoulos et al (1999). Οι Tao et al. τροποποίησαν το PSVM με την προσθήκη ενός νέου διαγώνιου πίνακα στο αρχικό πρόβλημα βελτιστοποίησης αναθέτοντας διαφορετικούς συντελεστές ποινής για τα θετικά και τα αρνητικά δείγματα αντίστοιχα, ως εξής

$$\frac{1}{2}(y'Vy) + \frac{1}{2}(w'w + \gamma^2)$$

$$s.t. D(Aw - e\gamma) + y = e$$

όπου, V είναι ένας διαγώνιος πίνακας με:

$$\begin{cases} C, & D_{ii} = 1 \\ \delta C, & D_{ii} = -1 \end{cases}$$

όπου δ είναι το ποσοστό μεταξύ των δύο μεγεθών των κλάσεων και C είναι το κόστος που σχετίζεται με τη θετική κλάση. Αρχικά αυτή είναι η ιδέα των Veropoulos et al (1999) για τη χρήση δύο κόστων στο SVM στην περίπτωση των μη ισορροπημένων δεδομένων συνδυαζόμενη με την ιδέα των Fung and Mangasarian (2002) για το σταδικό αλγόριθμο SVM.

3.6.5 Μέθοδοι Τροποποίησης του πυρήνα (Kernel modification methods)

Υπάρχουν αρκετές τεχνικές που προτείνονται στη βιβλιογραφία για να κάνουν τον αλγόριθμο SVM λιγότερο ευαίσθητο στην ανισορροπία μεταξύ των κλάσεων τροποποιώντας τη συνάρτηση του πυρήνα.

3.6.5.1 Όριο ευθυγράμμισης των κλάσεων (Class boundary alignment)

Οι Wu και Chang (2003a) πρότειναν μια παραλλαγή της μεθόδου SVM, όπου η συνάρτηση του πυρήνα είναι τροποποιημένη ώστε να μεγιστοποιεί το περιθώριο γύρω από την περιοχή του συνόρου της κλάσης στο μετασηματισμένο υψηλότερης διάστασης χώρο των χαρακτηριστικών ώστε να έχουν βελτιωμένη απόδοση. Οι Wu και Chang (2003a) βελτίωσαν αυτή τη μέθοδο για τα μη ισορροπημένα σύνολα δεδομένων, μεγαλώνοντας περισσότερο το όριο γύρω από την κλάση μειοψηφίας σε σύγκριση με το όριο γύρω από την κλάση πλειοψηφίας. Αυτή η μέθοδος ονομάζεται ευθυγράμμιση του ορίου των κλάσεων (class boundary alignment (CBA)) και μπορεί να χρησιμοποιηθεί μόνο με το διανυσματικό χώρο αναπαράστασης των επεξηγηματικών μεταβλητών. Οι Wu και Chang (2005) πρότειναν μία περαιτέρω παραλλαγή της μεθόδου CBA για την αναπαράσταση της αλληλουχίας των μη ισορροπημένων δεδομένων εισόδου, τροποποιώντας τον πίνακα του πυρήνα ώστε να έχει παρόμοια επίδραση και η οποία ονομάζεται ευθυγράμμιση του ορίου του πυρήνα (Kernel Boundary Alignment (KBA)).

3.6.5.2 Ευθυγράμμιση στόχου πυρήνα (Kernel target alignment)

Στο πλαίσιο της SVM μάθησης, ένα ποσοτικό μέτρο της συμφωνίας μεταξύ της συνάρτησης του πυρήνα που χρησιμοποιείται και του έργου της μάθησης είναι ιδιαίτερα σημαντικό τόσο από θεωρητικής όσο και πρακτικής άποψης. Η μέθοδος Ευθυγράμμισης του στόχου του πυρήνα έχει προταθεί ως μία μέθοδος για τη μέτρηση της συμφωνίας μεταξύ του πυρήνα που χρησιμοποιείται και της περίπτωσης ταξινόμησης που αναλύεται στους Cristianini et al. (2002). Αυτή η μέθοδος έχει βελτιωθεί για τη μάθηση μη ισορροπημένων συνόλων δεδομένων από τους Kandola και Shawe-taylor (2003).

3.6.5.3 Βαθμονόμηση του περιθωρίου (Margin calibration)

Η μέθοδος DEC (TCSVM) που περιγράφηκε προηγουμένως τροποποιεί την αντικειμενική συνάρτηση SVM με την ανάθεση υψηλότερου κόστος εσφαλμένης ταξινόμησης για τα θετικά απ’ ότι για τα αρνητικά παραδείγματα ώστε να αλλάξει ο

όρος ποινής. Οι Yang et al. (2009) επέκτειναν αυτή τη μέθοδο για να τροποποιήσουν την αντικειμενική συνάρτηση SVM όχι μόνο από την άποψη του όρου ποινής, αλλά επίσης και όσον αφορά το περιθώριο ώστε να διορθώσουν την μεροληπτική απόφαση του ορίου. Όπως προτείνεται σε αυτή τη μέθοδο, η τροποποίηση πρώτα υιοθετεί μία αντίστροφη κανονικοποιημένη ποινή ώστε αν εξισορροπηθούν οι κλάσεις. Στη συνέχεια, χρησιμοποιείται τροποποιημένο περιθώριο να οδηγήσει το περιθώριο έτσι ώστε να είναι μονόπλευρο, το οποίο επιτρέπει το όριο απόφασης να μετατοπιστεί.

3.6.5.4 Other kernel-modification methods

Υπάρχουν αρκετές τεχνικές μάθησης για μη ισορροπημένα δεδομένα που προτείνονται στη βιβλιογραφία για άλλους ταξινομητές που βασίζονται στην τεχνική του πυρήνα. Αυτές οι μέθοδοι περιλαμβάνουν τον αλγόριθμο κατασκευής ταξινομητή με πυρήνα που προτείνεται από τους Hong et al. (2007) που βασίζεται στην ορθογώνια προς τα εμπρός επιλογή (orthogonal forward selection OFS) και στον εκτιμητή κανονικοποιημένων ορθογώνιων σταθμισμένων ελαχίστων τετραγώνων (ROWLSs), ο αλγόριθμος KNG για την μη ισορροπημένη ομαδοποίηση Qin και Suganthan (2004), ο αλγόριθμος P2PKNNC με βάση τον ταξινομητή των k-πλησιέστερων γειτόνων και το παράδειγμα της επικοινωνίας, P2P, των Yu και Yu (2007), τον Adaboost (relevance vector machine (RVM)) Tashk και Faez (2007), μεταξύ άλλων.

ΚΕΦΑΛΑΙΟ 4

Μέτρα Απόδοσης σε Μη Ισορροπημένα Πεδία
(Evaluation Criteria in Imbalanced Domains)

4.1 Εισαγωγή

Η γενικευμένη απόδοση της μεθόδου εκμάθησης σχετίζεται με την ικανότητα της πρόβλεψης σε ανεξάρτητα δεδομένα δοκιμών (test data). Η αξιολόγηση της απόδοσης είναι εξαιρετικά σημαντική στην πράξη, δεδομένου ότι κατευθύνει την επιλογή της μεθόδου μάθησης ή το μοντέλο που θα χρησιμοποιήσουμε και μας δίνει ένα μέτρο της ποιότητας του τελικώς επιλεγμένου μοντέλου. Σε αυτό το κεφάλαιο θα περιγράψουμε κάποιες μεθόδους για την αξιολόγηση της απόδοσης.

4.2 Μεροληψία, Διασπορά και περιπλοκότητα μοντέλου

Το Σχήμα 4.1 απεικονίζει ένα σημαντικό ζήτημα που δεν είναι άλλο από την εκτίμηση της δυνατότητας της μεθόδου μάθησης να γενικευθεί. Θεωρούμε την περίπτωση μιας ποσοτικής μεταβλητής απόκρισης. Έχουμε μια μεταβλητή Y-στόχο, ένα διάνυσμα από εισόδους X, και μία πρόβλεψη του μοντέλου, $\hat{f}(X)$, που έχει υπολογιστεί από ένα σύνολο δεδομένων, το σύνολο εκπαίδευσης T.

Η λειτουργία απώλειας για τη μέτρηση των σφαλμάτων μεταξύ της Y και της πρόβλεψης $\hat{f}(X)$ συμβολίζεται με $L\ (Y, \hat{f}(X))$.

Τυπικές επιλογές αποτελούν οι:

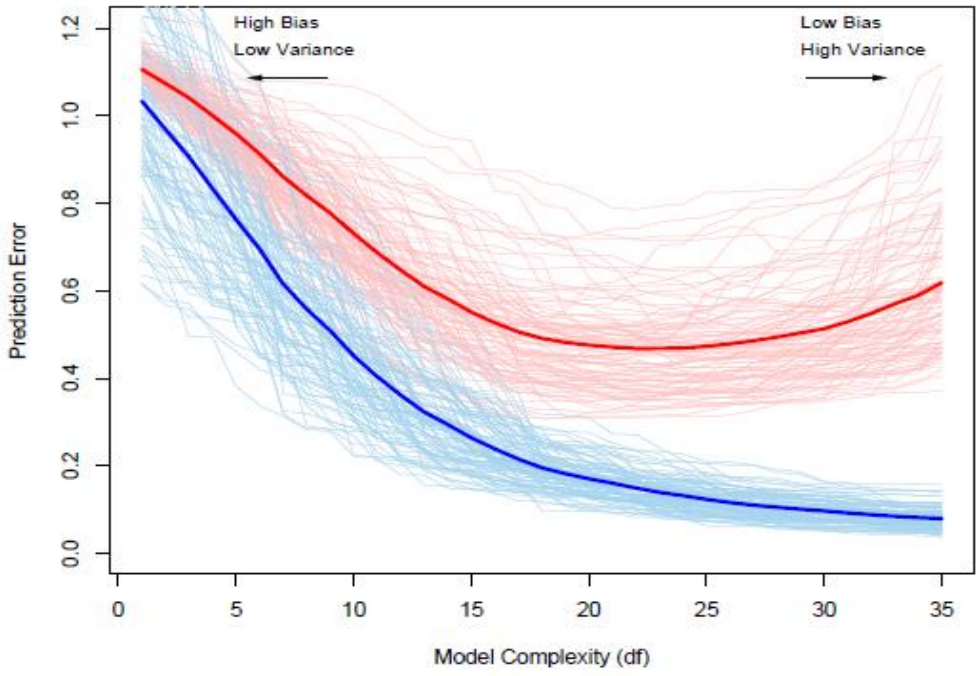
$$L\left(Y,\hat{f}(X)\right)=\begin{cases}(Y-\hat{f}(X))^2 & squared\ error\\ \left|Y-\hat{f}(X)\right| & absolute\ error\end{cases}$$

Το σφάλμα δοκιμών (test error), που αναφέρεται επίσης ως σφάλμα γενίκευσης (generalization error), είναι το σφάλμα πρόβλεψης πάνω σ' ένα ανεξάρτητο σύνολο δοκιμών (test sample):

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

όπου και οι δύο X και Y επιλέγονται τυχαία από την κοινή τους κατανομή (πληθυσμό). Εδώ το T σύνολο εκπαίδευσης είναι σταθερό, και το σφάλμα δοκιμής αναφέρεται στο σφάλμα για αυτό το συγκεκριμένο σύνολο εκπαίδευσης. Μία σχετική ποσότητα είναι το αναμενόμενο σφάλμα πρόβλεψης ή αλλιώς το αναμενόμενο σφάλμα δοκιμής (expected prediction error / expected test error):

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



Σχήμα 4.1: Συμπεριφορά του σφάλματος του συνόλου δοκιμών και του συνόλου εκπαίδευσης καθώς ποικίλει η περιπλοκότητα του μοντέλου. Οι ανοιχτόχρωμες μπλε καμπύλες δείχνουν το σφάλμα της εκπαίδευσης err, ενώ οι ανοιχτόχρωμες κόκκινες καμπύλες δείχνουν το υποθετικό σφάλμα δοκιμών Err_T για 100 σετ εκπαίδευσης μεγέθους 50 το καθένα, καθώς η πολυπλοκότητα του μοντέλου μεγαλώνει. Οι συμπαγείς καμπύλες δείχνουν το αναμενόμενο σφάλμα δοκιμών Err και το αναμενόμενο σφάλμα εκπαίδευσης E(err).

Σημειώνουμε ότι αυτός ο αναμενόμενος μέσος όρος είναι πάνω σε ότι είναι τυχαίο, συμπεριλαμβανομένης και της τυχαιότητας στο σύνολο εκπαίδευσης που παρήγαγε την $\hat{f}(X)$. Το Σχήμα 4.1 παρουσιάζει το σφάλμα πρόβλεψης (ανοιχτόχρωμες κόκκινες καμπύλες) Err_T για 100 προσομοιωμένα σύνολα εκπαίδευσης το καθένα μεγέθους 50. Η μέθοδος Lasso[10] είναι εκείνη που χρησιμοποιήθηκε για να παράγει την ακολουθία που ταιριάζει. Η συμπαγής κόκκινη καμπύλη είναι ο μέσος όρος, και ως εκ τούτου, η εκτίμηση του Err. Στόχος μας είναι η εκτίμηση του Err_T, αν και θα δούμε ότι το Err είναι πιο επιδεκτικό σε στατιστική ανάλυση, και οι περισσότερες μέθοδοι εκτιμούν αποτελεσματικά το αναμενόμενο σφάλμα. Επίσης, δεν φαίνεται να είναι δυνατό να εκτιμηθεί αποτελεσματικά το υποθετικό σφάλμα (conditional error), αφού δίνεται μόνο η πληροφορία για το ίδιο σύνολο εκπαίδευσης.

Το σφάλμα εκπαίδευσης είναι η μέση απώλεια πάνω στο δείγμα εκπαίδευσης

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

Θα θέλαμε να γνωρίζουμε το αναμενόμενο σφάλμα (test error) της μέτρησης του εκτιμώμενου μοντέλου μας $\hat{f}$. Καθώς το μοντέλο γίνεται όλο και πιο πολύπλοκο, χρησιμοποιεί τα δεδομένα εκπαίδευσης περισσότερο και είναι σε θέση να προσαρμοστούν σε περισσότερο πολύπλοκες υποκείμενες δομές. Ως εκ τούτου, υπάρχει μια μείωση στην μεροληψία αλλά αύξηση στη διασπορά (διακύμανση). Υπάρχει κάποια ενδιάμεση πολυπλοκότητα ενός μοντέλου που δίνει το ελάχιστο αναμενόμενο σφάλμα δοκιμών (test error). Δυστυχώς το σφάλμα εκπαίδευσης δεν είναι μια καλή εκτίμηση του σφάλματος δοκιμής, όπως φαίνεται στο Σχήμα 4.1. Το σφάλμα εκπαίδευσης μειώνεται σταθερά με την πολυπλοκότητα του μοντέλου, συνήθως πέφτει στο μηδέν εάν αυξηθεί αρκετά η πολυπλοκότητα του μοντέλου. Ωστόσο, ένα μοντέλο με μηδενικό σφάλμα εκπαίδευσης είναι «υπερπροσαρμοσμένο» στα δεδομένα εκπαίδευσης και τυπικά δεν θα γενικευθεί καλά. Η διαδικασία είναι παρόμοια για μια ποιοτική ή κατηγορηματική μεταβλητή απόκρισης.

Στο κεφάλαιο αυτό περιγράφουμε ένα αριθμό μεθόδων για την εκτίμηση του αναμενόμενου σφάλματος δοκιμών (test error) για ένα μοντέλο. Συνήθως το μοντέλο μας θα έχει μια ρυθμιστική παράμετρο ή παραμέτρους και έτσι μπορούμε να γράψουμε τις προβλέψεις μας ως $f_a(x)$. Η παράμετρος ρύθμισης διαφέρει ανάλογα με την πολυπλοκότητα του μοντέλου μας, και θέλουμε να βρούμε την τιμή αυτού που ελαχιστοποιεί σφάλμα, δηλαδή αυτού που παράγει την ελάχιστη καμπύλη του σφάλματος δοκιμών (test error) στο Σχήμα 4.1. Είναι σημαντικό να σημειωθεί ότι υπάρχουν στην πραγματικότητα δύο ξεχωριστοί στόχοι που θα μπορούσαμε να έχουμε κατά νου:

[10] Η Lasso είναι μια μέθοδος συρρίκνωσης όπως η μέθοδος κορυφογραμμής, με λεπτές αλλά σημαντικές διαφορές.

Επιλογή Μοντέλου (Model selection): εκτίμηση της απόδοσης των διαφορετικών μοντέλων για να επιλέξουμε το καλύτερο.

Εκτίμηση Μοντέλου (Model assessment): έχοντας επιλέξει ένα τελικό μοντέλο, εκτίμηση του σφάλματος πρόβλεψης του μοντέλου αυτού (σφάλμα γενίκευσης) για τα νέα δεδομένα.

Αν έχουμε ένα αρκετά μεγάλο σύνολο δεδομένων, η καλύτερη προσέγγιση για τα δύο προβλήματα είναι να διαιρέσουμε το σύνολο δεδομένων τυχαία σε τρία μέρη: ένα σύνολο εκπαίδευσης (*training set*), ένα σύνολο επικύρωσης (*validation set*), και ένα σύνολο δοκιμών (test set). Το σύνολο εκπαίδευσης χρησιμοποιείται για να προσαρμόσει τα μοντέλα, το σύνολο επικύρωσης χρησιμοποιείται για την εκτίμηση του σφάλματος πρόβλεψης (prediction error) για την επιλογή μοντέλου και τέλος, το σύνολο δοκιμής χρησιμοποιείται για την εκτίμηση του σφάλματος γενίκευσης (generalization error) του τελικού επιλεχθέντος μοντέλου. Ιδανικά, το σετ δοκιμής θα πρέπει να κρατείται σε απομονωμένο και να το φέρνουμε στην επιφάνεια μόνο στο τέλος της ανάλυσης των δεδομένων. Ας υποθέσουμε ότι αντί να χρησιμοποιούμε το test-set επανειλημμένα, επιλέγουμε το μοντέλο με το μικρότερο σφάλμα δοκιμών (test error). Τότε το σφάλμα δοκιμής του τελικά επιλεγμένου μοντέλου θα υποτιμήσει το πραγματικό σφάλμα της δοκιμής (test error), μερικές φορές σε σημαντικό βαθμό. Είναι δύσκολο να δοθεί ένας γενικός κανόνας για το πώς να επιλέξουμε τον αριθμό των παρατηρήσεων σε κάθε ένα από τα τρία μέρη, καθώς αυτό εξαρτάται από την αναλογία signal-to-noise στα δεδομένα και το μέγεθος του δείγματος εκπαίδευσης. Μια τυπική διάσπαση θα μπορούσε να είναι 50% για την εκπαίδευση, και 25% για κάθε ένα από τα σύνολα δοκιμής και επικύρωσης:

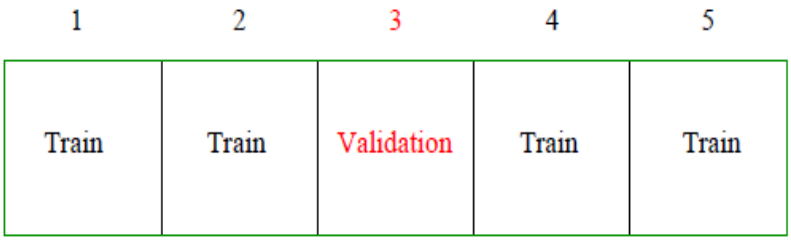
Συνήθως έχουμε περιπτώσεις όπου υπάρχουν ανεπαρκή στοιχεία για να χωριστούν τα δεδομένα σε τρία μέρη. Και πάλι είναι πάρα πολύ δύσκολο να δοθεί ένας γενικός κανόνας σχετικά με κατά πόσο ο όγκος των δεδομένων εκπαίδευσης είναι αρκετός; μεταξύ άλλων, αυτό εξαρτάται από την αναλογία signal-to-noise ratio σήματος προς θόρυβο της υποκεί-μενης λειτουργίας, και η πολυπλοκότητα των μοντέλων που ταιριάζουν με τα δεδομένα.

Η προσέγγιση στο στάδιο της επικύρωσης γίνεται είτε αναλυτικά (AIC, BIC, MDL, SRM) ή από την αποτελεσματική επαναχρησιμοποίηση του δείγματος (cross-validation και η bootstrap). Εκτός από τη χρήση αυτών των μεθόδων στην επιλογή μοντέλου, μπορούμε επίσης να εξετάσουμε σε ποιο βαθμό κάθε μέθοδος παρέχει μια αξιόπιστη εκτίμηση του σφάλματος δοκιμής του τελικά επιλεχθέντος μοντέλου.

4.3 Διασταυρωμένη επικύρωση (Cross-validation)

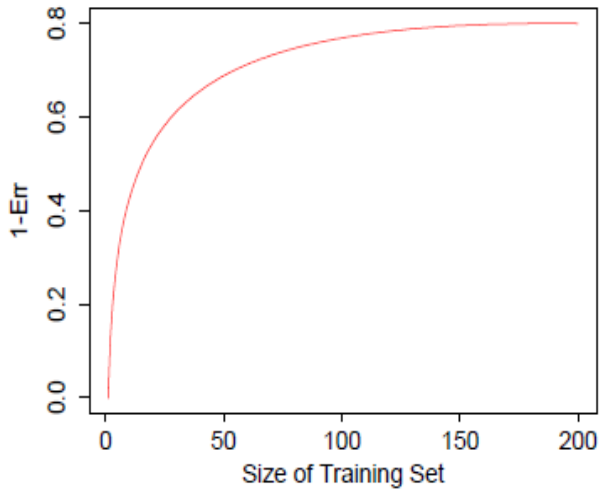
Πιθανώς η απλούστερη και πιο ευρέως χρησιμοποιούμενη μέθοδος για την εκτίμηση του σφάλματος πρόβλεψης (prediction error) είναι η διασταυρωμένη επικύρωση. Αυτή η μέθοδος υπολογίζει άμεσα το αναμενόμενοο εξω-δείγματος σφάλμα $Err = E\left[L\left(Y, \hat{f}(X)\right)\right]$ το μέσο σφάλμα γενίκευσης (generalization error) όταν η μέθοδος $\hat{f}(X)$ εφαρμόζεται σε ένα ανεξάρτητο δείγμα δοκιμής από την κοινή κατανομή των X και Y. Όπως αναφέρθηκε προηγουμένως, μπορούμε να ελπίζουμε ότι η διασταυρωμένη επικύρωση εκτιμά το υπό όρους σφάλμα (conditional error), με το σύνολο εκπαίδευσης T που έχει καθοριστεί. Αλλά, η διασταυρωμένη επικύρωση συνήθως είναι καλή υπολογιστικά μόνο για το αναμενόμενο σφάλμα πρόβλεψης (prediction error).

Στην ιδανική περίπτωση, αν είχαμε αρκετά δεδομένα, θα αναιρέσουμε το σύνολο επικύρωσης και θα το χρησιμοποιήσουμε για να αξιολογήσουμε την απόδοση του μοντέλου που προβλέψαμε. Δεδομένου ότι τα δεδομένα συχνά σπανίζουν, αυτό δεν είναι συνήθως δυνατό. Για την φινέτσα του προβλήματος, η K-φορές διασταυρωμένη επικύρωση χρησιμοποιεί μέρος των διαθέσιμων δεδομένων για να προσαρμόσει το μοντέλο, και ένα διαφορετικό μέρος για να το δοκιμάσει. Έχουμε χωρίσει τα δεδομένα σε K τμήματα, περίπου ίσου μεγέθους, για παράδειγμα, όταν K = 5, το σενάριο μοιάζει με το ακόλουθο:



Για το k-οστό τμήμα (τρίτο στο παραπάνω σχήμα), προσαρμόζουμε το μοντέλο με τα άλλα K-1 μέρη των δεδομένων, και υπολογίζουμε το σφάλμα πρόβλεψης (prediction error) του προσαρμοσμένου μοντέλου όταν προβλέπουμε το κ-οστό τμήμα των δεδομένων. Το κάνουμε αυτό για k = 1, 2,. . . , K και συνδυάζουμε τις K εκτιμήσεις του σφάλματος πρόβλεψης (prediction error).

Τι τιμή θα πρέπει να επιλέξουμε για το K; Με K = N, ο εκτιμητής της διασταυρωμένης επικύρωσης είναι περίπου αμερόληπτος για το αληθές (αναμενόμενο) σφάλμα πρόβλεψης (prediction error), αλλά μπορεί να έχει μεγάλη διασπορά, επειδή τα N "σύνολα εκπαίδευσης" είναι τόσο όμοια το ένα στο άλλο. Η υπολογιστική επιβάρυνση είναι επίσης σημαντική, απαιτώντας N εφαρμογές της μεθόδου εκμάθησης.



Σχήμα 4.2: Υποθετική καμπύλη μάθησης για έναν ταξινομητή σε ένα συγκεκριμένο έργο: ένα γράφημα 1 - Err σε σχέση με το μέγεθος του συνόλου εκπαίδευσης N. Με ένα σύνολο δεδομένων από 200 παρατηρήσεις, μία 5-fold διασταυρωμένη επικύρωση θα χρησιμοποιεί σύνολα εκπαίδευσης μεγέθους 160, τα οποία θα συμπεριφέρονται σαν το πλήρες σύνολο. Ωστόσο, με ένα σύνολο δεδομένων των 50 παρατηρήσεων η 5-fold διασταυρωμένη επικύρωση θα χρησιμοποιεί σύνολα εκπαίδευσης μεγέθους 40, και αυτό θα είχε ως αποτέλεσμα μία σημαντική υπερεκτίμηση του σφάλματος πρόβλεψης.

4.4 Πίνακας Συνάφειας

Συγκρίνοντας τις μεθοδολογίες μεταξύ τους μπορούμε να αναγνωρίσουμε δύο πτυχές: της αποτελεσματικότητας (effectiveness) και της αποδοτικότητας (efficiency).

Ορισμός 1: Αποτελεσματικότητα σημαίνει η ικανότητα του μοντέλου να ταξινομήσει με ακρίβεια άγνωστα δείγματα, σε σχέση με κάποιο μέτρο απόδοσης.

Ορισμός 2: Αποδοτικότητα σημαίνει την ταχύτητα που χρησιμοποιεί ένα μοντέλο για να ταξινομήσει άγνωστα δείγματα.

Πολλά μέτρα έχουν χρησιμοποιηθεί για την αξιολόγηση της αποτελεσματικότητας ενός ταξινομητή στο πρόβλημα της μάθησης από μη ισορροπημένα δεδομένα. Τα περισσότερα από αυτά βασίζονται στον πίνακα συνάφειας (Πίνακας 4.1). Δεδομένου ενός ταξινομητή και ενός παραδείγματος, υπάρχουν τέσσερα πιθανά αποτελέσματα.

TP: Αν η περίπτωση είναι *θετική* και είναι ταξινομημένη ως *θετική*, υπολογίζεται ως μια αληθώς θετική.
FN: Αν η περίπτωση είναι *θετική* και έχει ταξινομηθεί ως *αρνητική*, αυτό υπολογίζεται ως ψευδώς αρνητική.
TN: Αν η περίπτωση είναι *αρνητική* και έχει ταξινομηθεί ως *αρνητική*, αυτή υπολογίζεται ως μια αληθώς αρνητική.
FP: Αν η περίπτωση είναι *αρνητική* και έχει ταξινομηθεί ως *θετική*, προσμετράται ως ψευδώς θετική.

Δεδομένου ενός ταξινομητή και μια σειράς από περιπτώσεις (στο σύνολο δοκιμής), μπορεί να κατασκευαστεί ένας 2×2 πίνακας συνάφειας (ονομάζεται επίσης πίνακας έκτακτης ανάγκης) όπου αντιπροσωπεύονται οι διατάξεις του συνόλου των περιπτώσεων.

Πίνακας 4.1 Πίνακας Συνάφειας

Πραγματική/ προβλεπόμενη		**Προβλεπόμενη**	
		Θετικά	Αρνητικά
Πραγματική	Θετικά (P)	Αληθώς Θετικά (TP)	Ψευδώς Αρνητικά (FN) (Type II error)
	Αρνητικά (N)	Ψευδώς Θετικά (FP) (Type I error)	Αληθώς Αρνητικά (TN)

Αυτός ο πίνακας, όπως προείπαμε, αποτελεί τη βάση για πολλές μετρήσεις. Οι αριθμοί κατά μήκος των κυρίων διαγωνίων αντιπροσωπεύουν τις σωστές αποφάσεις, και οι αριθμοί εκτός της διαγωνίου αντιπροσωπεύουν τα λάθη – τη σύγχυση – μεταξύ των διαφόρων κατηγοριών. Ακρίβεια (***accuracy***) είναι η αναλογία των πραγματικών αποτελεσμάτων (και τα δύο, αληθώς θετικά (TP) και αληθώς αρνητικά (TN)) στον πληθυσμό. Είναι μια παράμετρος της δοκιμής/τεστ.

$$accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

Ακρίβεια 100% σημαίνει ότι οι μετρούμενες τιμές είναι ακριβώς ίδιες με τις τιμές που δίνονται. Στην περίπτωση που υπάρχει μεγάλη ανισορροπία στην κατανομή των δεδόμενων, η συνολική ακρίβεια (accuracy) παύει να αποτελεί ένα επαρκές μέτρο. Υποθέτοντας και πάλι ότι έχουμε δύο κλάσεις και ότι η αρνητική κατηγορία αποτελεί την κλάση πλειοψηφίας, τότε ο αριθμός των αληθώς αρνητικών (True Negative) μπορεί να είναι ασυνήθιστα υψηλός σε ένα κανονικά κατανεμημένο συνόλο δεδομένων (δηλαδή με δύο κλάσεις που είναι κανονικά κατανεμημένες), βασιζόμενοι στο ποσοστό της ανισορροπίας (IR: Imbanced Ratio). Η αύξηση των TN θα αυξήσει την ακρίβεια της ταξινόμησης, αλλά πολλά από τα δείγματα της κλάσης μειοψηφίας θα ταξινομούνται λανθασμένα. Κατά συνέπεια, η ακρίβεια μπορεί να είναι ένα αναξιόπιστο μέτρο για την ταξινόμηση του συνόλου δεδομένων. Για την καλύτερη κατανόηση, εξετάζουμε μια περίπτωση από 100 δείγματα σε ένα σύνολο δεδομένων το οποίο το 97% είναι αρνητικά και το 3% θετικά παραδείγματα αντίστοιχα. Σε γενικές γραμμές, η πρόβλεψη της κλάσης πλειοψηφίας θα μας δώσει ακρίβεια 97%. Ωστόσο, αυτή η στρατηγική δεν είναι αρκετά ακριβείς για να χαρακτηρίσει την κλάση των θετικών περιπτώσεων.

Αντιλαμβανόμαστε λοιπόν ότι η συνολική ακρίβεια μπορεί να κυριαρχείται από την ακρίβεια ταξινόμησης της κλάσης πλειοψηφίας.

Για να διαχειριστούμε αυτό το πρόβλημα υπάρχουν δύο είδη μέτρων απόδοσης που λαμβάνουμε από τον πίνακα συνάφειας και μας παρέχουν μία ανάλυση ευαίσθητη σε τέτοιου είδους καταστάσεις. Για να λάβουμε την επιθυμητή ικάνοτητα ταξινόμησης και να ελέγξουμε την επίδοση της ταξινόμησης ξεχωριστά σε κάθε κλάση χρησιμοποιούμε δύο μέτρα απόδοσης, την ειδικότητα και την ευαισθησία. Βασιζόμενοι σε αυτά τα δύο, προτάθηκαν νέα μέτρα για τα μη ισορροπημένα δεδομένα όπως είναι για παράδειγμα ο γεωμετρικός μέσος και η περιοχή κάτω από την καμπυλη ROC τα οποία θα συζητήσουμε στη συνέχεια.

Γενικά η απόδοση των ταξινομητών στην εκμάθηση από μη ισορροπημένα δεδομένα μπορούν να αξιολογηθούν με βάση τέσσερα κριτήρια. Αυτά είναι (1) το κριτήριο Ελάχιστου Κόστους (MC), (2) το κριτήριο του μέγιστου Γεωμετρικού Μέσου (Maximum Geometric Mean, MGM) της ακρίβειας στην κλάση πλειοψηφίας και στην κλάση μειοψηφίας (3) το κριτήριο του Μέγιστου Αθροίσματος (MS) της ακρίβειας για την κλάση πλειοψηφίας και την κλάση μειοψηφίας, και (4) το κριτήριο ανάλυσης των ROC καμπυλών.Τα διάφορα είδη μέτρων απόδοσης για τα μη ισορροπημένα σύνολα δεδομένων φαίνονται στο ακόλουθο σχήμα (Σχήμα 4.3).

Σχήμα 4.3 Μέτρα απόδοσης για τα μη ισορροπημένα δεδομένα

Όπως προείπαμε, η ευαισθησία και η ειδικότητα είναι βασικά μέτρα για τη δημιουργία νέων, πιο αξιόπιστων στην περίπτωση των μη ισορροπημένων κλάσεων. Η ευαισθησία (sensitivity) και η ειδικότητα (specificity) είναι στατιστικά μέτρα της απόδοσης ενός τεστ δυαδικής ταξινόμησης, γνωστές στη στατιστική ως συναρτήσεις ταξινόμησης. Αυτά τα δύο μέτρα συνδέονται στενά με τις έννοιες των σφαλμάτων τύπου I και τύπου II. Ένας τέλειος εκτιμητής θα πρέπει να περιγράφεται με 100% ευαίσθησία και 100%

ειδικότητα. Οι συχνότερα, λοιπόν, χρησιμοποιούμενες συνιστώσες της διαγνωστικής ποιότητας μιας δοκιμής, που καθορίζουν τη διακριτική της ικανότητα είναι:

Το *ποσοστό των αληθώς θετικών αποτελεσμάτων* είναι το ποσοστό των θετικών ενδείξεων στον πληθυσμό (true positive rate, **TPR**) ή η **ευαισθησία** (sensitivity) ενός ταξινομητή, δηλαδή η πιθανότητα το τεστ να είναι θετικό δεδομένου ότι κάποιος έχει το χαρακτηριστικό που εξετάζουμε και δίνεται από τον τύπο:

$$SE = TPR = \frac{\alpha\lambda\eta\theta\acute{\omega}\varsigma\ \theta\varepsilon\tau\iota\kappa\acute{\alpha}}{\sigma\acute{\upsilon}\nu o\lambda o\ \theta\varepsilon\tau\iota\kappa\acute{\omega}\nu} = \frac{TP}{P} = \text{Recall} = \frac{TP}{TP + FN} = 1 - \text{Type II error}$$

Η ευαισθησία σχετίζεται με την ικανότητα του τεστ να προσδιορίσει θετικά αποτελέσματα. Μια δοκιμή με υψηλή ευαισθησία έχει χαμηλό ποσοστό σφάλματος τύπου ΙΙ.

Το *ποσοστό των αληθώς αρνητικών αποτελεσμάτων* είναι το ποσοστό των αρνητικών ενδείξεων στον πληθυσμό (true negative rate, **TNR**) ή η **ειδικότητα** (specificity) της δοκιμασίας, δηλαδή η πιθανότητα το τεστ να είναι αρνητικό δεδομένου ότι κάποιος δεν έχει το χαρακτηριστικό που εξετάζουμε και υπολογίζεται ως εξής:

$$SPC = TNR = \frac{\alpha\lambda\eta\theta\acute{\omega}\varsigma\ \alpha\rho\nu\eta\tau\iota\kappa\acute{\omega}\nu}{\sigma\acute{\upsilon}\nu o\lambda o\ \alpha\rho\nu\eta\tau\iota\kappa\acute{\omega}\nu} = \frac{TN}{N} = \frac{TN}{FP + TN} = 1 - \text{Type I error}$$

Η ειδικότητα σχετίζεται με την ικανότητα του τεστ να εντοπίσει αρνητικά αποτελέσματα. Μια δοκιμή με υψηλή εξειδίκευση έχει χαμηλό ποσοστό σφάλματος τύπου Ι.

Πίνακας 4.2 Συγκεντρωτική εικόνα των μέτρων απόδοσης

		Πραγματική τιμή		
		Αληθές (T)	Ψευδές (F)	
Αποτέλεσμα του τεστ	Θετικό (P)	**TP**	**FP**	→ Θετική προγνωστική τιμή ($PPV/precision$)
	Αρνητικό (N)	**FN**	**TN**	→Αρνητική προγνωστική τιμή (NPV)
		↓ Ευαισθησία Sensitivity	↓ Ειδικότητα Specificity	Ακρίβεια Accuracy

Τα ποσοστά αυτά, καθώς και τα συμπληρωματικά τους (ποσοστό ψευδώς αρνητικών (**FNR**) και ψευδώς θετικών αποτελεσμάτων (**FPR**), αντίστοιχα) ονομάζονται *πιθανοφάνειες* (likelihood) ή, αλλιώς, λειτουργικά χαρακτηριστικά (operating characteristics) της διαγνωστικής δοκιμασίας.

Προφανώς ισχύει ότι το

$$TPR = 1 - FNR$$

όπου $FNR = \frac{FN}{P} = \frac{FN}{TP+FN}$.

Τα PPV και NPV είναι αντίστοιχα με τα σφάλματα τύπου I και II στους αντίστοιχους ελέγχους υποθέσεων (Πίνακας 4.2).

Το μειονέκτημα της χρήσης μέτρων αξιολόγησης με βάση τον πίνακα συνάφειας είναι ότι κοιτάζουμε μόνο την απόδοση σε ένα σημείο, που σημαίνει ότι δεν μπορούμε να πούμε πως η διαφορετική κατανομή των κλάσεων ή το διαφορετικό κόστος θα μεταβάλλουν την απόδοση. Έτσι οι ερευνητές μπορεί να προτιμήσουν να δουν οπτικά την απόδοση σε μια ποικιλία καταστάσεων, χρησιμοποιώντας ένα από τα γραφικά εργαλεία αξιολόγησης, όπως μια καμπύλη ROC.

4.5 Καμπύλες ROC

Η πραγματοποίηση προβλέψεων αποτελεί σημαντικό μέλημα σε κάθε επιστημονικό πεδίο. Είναι λοιπόν αναγκαία η πραγματοποίηση προβλέψεων και η εξασφάλιση προγνωστικής ακρίβειας στον σχεδιασμό και την σύγκριση μοντέλων, αλγορίθμων και τεχνολογιών που παράγουν προβλέψεις. Οι ROC καμπύλες συμβάλλουν στην εξασφάλιση της επιθυμητής ακρίβειας στις προβλέψεις και αποτελούν μία χρήσιμη τεχνική για την απεικόνιση, την οργάνωση και την επιλογή ταξινομητών με βάση την απόδοση τους. Οι καμπύλες Λειτουργικού χαρακτηριστικού δέκτη (Receiver Operating Characteristic, ROC) είναι μια τυποποιημένη μέθοδος που συνοψίζει την απόδοση ενός ταξινομητή σε σχέση με μια σειρά από «ανταλλαγές» μεταξύ αληθώς θετικών (TP) και ψευδώς θετικών (FP) ποσοστών σφάλματος. Η καμπύλη ROC ορίζεται ως το μοναδιαίο τετράγωνο $[0,1] \times [0,1]$ και ξεκινά από το σημείο (0,0) (όταν το σημείο απόφασης είναι μεγαλύτερο από όλες τις μετρήσεις θορύβου και σήματος) για να καταλήξει στο (1,1) (για την περίπτωση που το σημείο απόφασης είναι μικρότερο από όλες τις μετρήσεις).

Η περιοχή κάτω από την ROC καμπύλη (AUC) είναι ένα αποδεκτό σύστημα μέτρησης απόδοσης για μια καμπύλη ROC. Οι ROC καμπύλες μπορούν να αποτελέσουν ένα μέτρο παρουσίασης της οικογένεια των καλύτερων ορίων απόφασης για τα σχετικό κόστος των TP και FP. Το εμβαδόν που ορίζεται κάτω από την καμπύλη αποτελεί ένα μέτρο της ποιότητας διαχωρισμού θορύβου – σήματος και χρησιμοποιείται συχνά στη στατιστική συμπερασματολογία των καμπυλών ROC.

4.5.1. Γραφήματα και Ερμηνεία

Η σχέση του ποσοστού των αληθώς θετικών (***TPR***) και ψευδώς θετικών (***FPR***) αποτελεσμάτων της διαγνωστικής δοκιμασίας, καθώς μεταβάλλεται προοδευτικά προς μια κατεύθυνση το διαχωριστικό όριο, παριστάνεται γραφικά με την **καμπύλη ROC** (*Receiver Operating Characteristic Curve*) ή καμπύλη λειτουργικών χαρακτηριστικών.

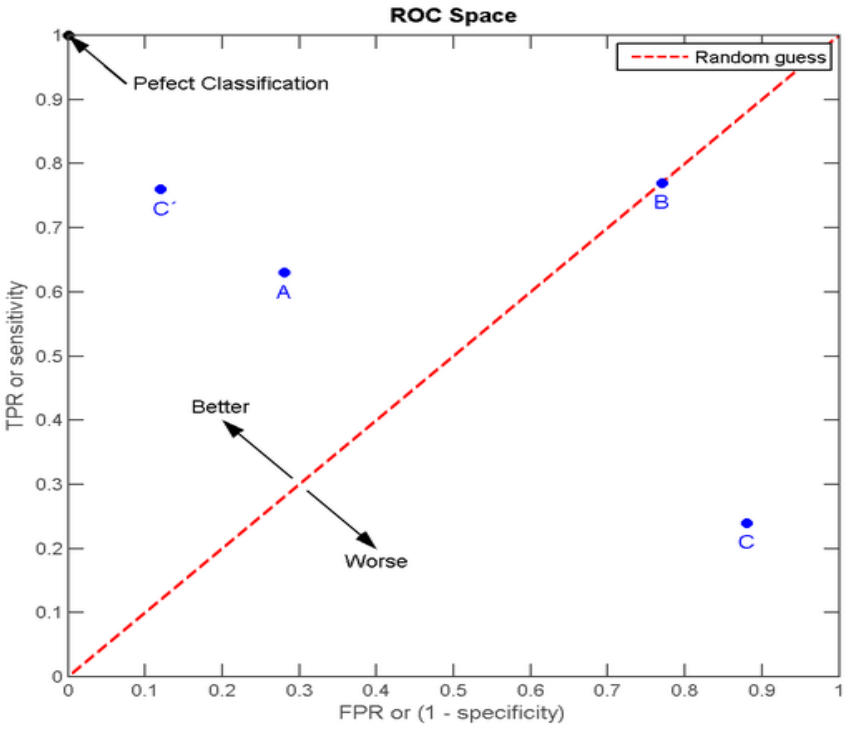
Ας θεωρήσουμε ένα πρόβλημα πρόβλεψης δύο κατηγοριών (δυαδική ταξινόμηση). Τα ROC γραφήματα είναι δισδιάστατα διαγράμματα στα οποία

Ο $y-$ άξονας αντιπροσωπεύει το $\%TP = TP\ /\ (TP\ +\ FN)$ ποσοστό

και

ο $x-$ άξονας αντιπροσωπεύει $\%FP\ =\ FP/(TN + FP)$ ποσοστό

Θα μπορούσαμε λοιπόν να πούμε ότι οι καμπύλες ROC απεικονίζουν τα σχετικά trade-offs ανάμεσα στα αληθώς θετικά (TP/οφέλη) και ψευδώς θετικά (FP/κόστοι). Επομένως θα μπορούσαμε να πούμε ότι ένα γράφημα ROC απεικονίζει τη σχετική μεταβολή μεταξύ του κέρδους (αληθώς θετικά) και του κόστους (ψευδώς θετικά). Η καμπύλη αυτή εγγράφεται μέσα σε ένα τετράγωνο, στις τέσσερις γωνίες του οποίου αντιστοιχούν οι ακραίες τιμές (0 και 1) του %TP και του %FP αποτελεσμάτων, καθώς και των συμπληρωματικών αυτών ποσοστών (%FN και %TN).



Σχήμα 4.4 Ο χώρος της ROC καμπύλης και το γράφημα 4 παραδειγμάτων πρόβλεψης

Λόγω του ότι το $TPR = sensitivity$ και το $FPR = 1 - specificity$ το ROC γράφημα καλείται κάποιες φορές και $(sensitivity)\ vs\ (1 - specificity)$ διάγραμμα. Κάθε πρόβλεψη ή περίπτωση του πίνακα συνάφειας αντιπροσωπεύει ένα σημείο στο χώρο ROC. Η καλύτερη δυνατή μέθοδος πρόβλεψης θα αποδώσει ένα σημείο στην επάνω αριστερή γωνία ή στη συντεταγμένη (0,1) του χώρου ROC, που αντιπροσωπεύει την περίπτωση που έχουμε 100% ευαισθησία (όχι ψευδώς αρνητικά) και την 100% ειδικότητα (δεν υπάρχουν ψευδώς θετικά) (perfect classification). Μια εντελώς τυχαία εικασία θα δώσει ένα σημείο κατά μήκος μιας διαγώνιας γραμμής από κάτω αριστερά προς τα πάνω δεξιά γωνία.

4.5.2. Η περιοχή κάτω από την ROC καμπύλη (AUC)

Η καμπύλη ROC είναι μια δισδιάστατη απεικόνιση της απόδοσης ταξινόμησης. Για να συγκρίνουμε τους ταξινομητές μπορεί να χρειαστεί να μειώσουμε την απόδοση ROC σε μία ενιαία βαθμωτή τιμή που αντιπροσωπεύει την αναμενόμενη απόδοση. Μια κοινή μέθοδος είναι να υπολογίσουμε το εμβαδόν κάτω από την καμπύλη(AUC) ROC (Hanley & McNeil, 1982). Εφόσον η AUC είναι μέρος της περιοχής της μονάδας, η τιμή της θα είναι πάντα μεταξύ 0 και 1. Ωστόσο, επειδή μια τυχαία εικασία παράγει τη διαγώνια γραμμή μεταξύ $(0, 0)$ και $(1, 1)$, η οποία έχει έκταση 0,5, κανένας ρεαλιστικός ταξινομητής δεν θα πρέπει να έχει AUC λιγότερο από 0,5.

Η AUC είναι επίσης στενά συνδεδεμένη με το δείκτη Gini (Breiman et al. (1984)), ο οποίος είναι διπλάσιος από το χώρο ανάμεσα στην διαγώνιο και στην καμπύλη ROC. Οι Hand και Till (2001) επισημαίνουν ότι

$$Gini\ = 2 \times AUC - 1$$

4.6 Precision and Recall

Για την καλύτερη κατανόηση θα ξεκινήσουμε με δύο παραδείγματα. Ας υποθέσουμε ότι ένα πρόγραμμα για την αναγνώριση σκύλων σε σκηνές από ένα βίντεο προσδιορίζει 7 σκυλιά σε μια σκηνή που περιέχει 9 σκυλιά και μερικές γάτες. Αν 4 από τις ταυτίσεις είναι σωστές, αλλά 3 είναι στην πραγματικότητα οι γάτες, η precision του προγράμματος είναι $\frac{4}{7}$ ενώ η recall είναι $\frac{4}{9}$. Όταν μια μηχανή αναζήτησης επιστρέφει 30 σελίδες, μόνο 20 από τις οποίες ήταν σχετικές, παραλείποντας να επιστρέψει 40 επιπρόσθετες σχετικές σελίδες, η precision του είναι $\frac{20}{30} = \frac{2}{3}$ ενώ η recall του είναι $\frac{20}{60} = \frac{1}{3}$.

Όσον αφορά στη στατιστική, η απουσία των σφαλμάτων τύπου I και τύπου II αντιστοιχεί αντίστοιχα στη μέγιστη precision (δεν υπάρχουν ψευδώς θετικά) και μέγιστη recall (όχι ψευδώς αρνητικά). Το παραπάνω παράδειγμα αναγνώρισης προτύπων περιείχε $7 - 4 = 3$ σφάλματα τύπου I και $9 - 4 = 5$ σφάλματα τύπου II. Η precision μπορεί να θεωρηθεί ως ένα μέτρο της ποιότητας, λαμβάνοντας υπόψη ότι η recall είναι ένα μέτρο της ποσότητα.

Σε θέματα ταξινόμησης, η precision για μια κλάση είναι ο αριθμός των αληθώς θετικών διαιρούμενο με τον συνολικό αριθμό των στοιχείων που επισημάνθηκε ότι ανήκουν στην θετική κλάση (δηλαδή το άθροισμα των αληθώς θετικών και των ψευδώς θετικών. Η Recall ορίζεται ως ο αριθμός των αληθώς θετικών διαιρούμενος με το συνολικό αριθμό των στοιχείων τα οποία ανήκουν πράγματι στη θετική κλάση (δηλαδή το άθροισμα των αληθώς θετικών και των ψευδώς αρνητικά, τα οποία είναι στοιχεία τα οποία δεν είχαν επισημανθεί ότι ανήκουν στην θετική κλάση αλλά θα έπρεπε να είχαν).

Η ακρίβεια ή *θετική προγνωστική τιμή* (*positive predictive value*) ορίζεται ως το ποσοστό των αληθώς θετικών έναντι όλων των θετικών αποτελεσμάτων (τόσο αληθώς θετικά όσο και ψευδώς θετικά).

$$\text{Precision} = PPV = \frac{TP}{TP + \text{FP}}$$

Παρόμοια ορίζεται και η αρνητική προγνωστική (ή διαγνωστική ή προβλεπόμενη) τιμή (negative predictive value) που συμβολίζεται με **NPV** :

$$NPV = \frac{TN}{TN + FN}$$

Maximum sum (*MS*)

$$MS = \frac{TP}{TP + \text{FN}} + \frac{TN}{TN + \text{FP}}$$

Στη δυαδική ταξινόμηση, recall ονομάζεται η ευαισθησία. Επομένως, όπως αναφέραμε και προηγουμένως

$$\text{Recall} = \frac{TP}{TP + FN} = sensitivity$$

Κύριος στόχος της μάθησης από μη ισορροπημένα σύνολα δεδομένων είναι η βελτίωση της recall, χωρίς επιπτώσεις στην τιμή της precision. Από την άλλη πλευρά, οι στόχοι της *recall* και της *precision* μπορεί να είναι συχνά αντιφατικοί, δεδομένου ότι αυξάνοντας τις αληθώς θετικές περιπτώσεις για την κλάση μειοψηφίας, ο αριθμός των ψευδώς θετικών μπορεί επίσης να αυξηθεί γεγονός που θα μειώσει την precision (Chawla (2010)).

Συνήθως, το σκορ της precision και της recall δεν συζητούνται ανεξάρτητα το ένα από το άλλο. Είναι κρατώντας σταθερό το ένα βλέπουμε τις τιμές που λαμβάνει το άλλο, είναι και τα δύο συνδυάζονται σε ένα ενιαίο μέτρο. Παράδειγμα αποτελεί το F-measure που είναι ο σταθμισμένος αρμονικός μέσος της precision και της recall. Το μέτρο F δίνει μία τιμή που αντικατοπτρίζει την καλή συμπεριφορά του ταξινομητή υπό την παρουσία σπάνιων κλάσεων. Ενώ οι ROC καμπύλες αντιπροσωπεύουν το trade-off μεταξύ των τιμών του TP και FP, η F-τιμή αντιπροσωπεύει το trade-off μεταξύ των διαφορετικών τιμών των TP, FP και FN (Buckland και Gey, 1994).

Η έκφραση για το F-τιμή είναι

$$F_{\beta} = \frac{(1+\beta^2) * \text{Recall} * \text{Precision}}{\beta^2 * \text{Recall} + \text{Precision}}$$

όπου το β αντιστοιχεί στη σχετική σημασία ανάμεσα στην precision και την recall. Συνήθως ορίζεται ίσο με τη μονάδα.

4.7 Γεωμετρικός Μέσος (GMean)

Ο γεωμετρικός μέσος είναι ένα από τα μέτρα απόδοσης που χρησιμοποιούνται σε ταξινομητές που χειρίζονται μη ισορροπημένα σύνολα δεδομένων. Ο λόγος που χρησιμοποιούμε τον GMean είναι για την εξισορρόπηση της αναλογίας της πρόβλεψης μεταξύ της κλάσης πλειοψηφίας και μειοψηφίας. Το ποσοστό του GMean δείχνει πόσο καλά ένας ταξινομητής με μη ισορροπημένα δεδομένα προβλέπει τις κλάσεις. Το GMean δίνεται ως εξής:

$$Geometric\ mean = \sqrt{Sensitivity * Specificity}$$

4.8 Μέτρα Ευαίσθητου Κόστους (Cost sensitive measures)

4.8.1. Cost curve

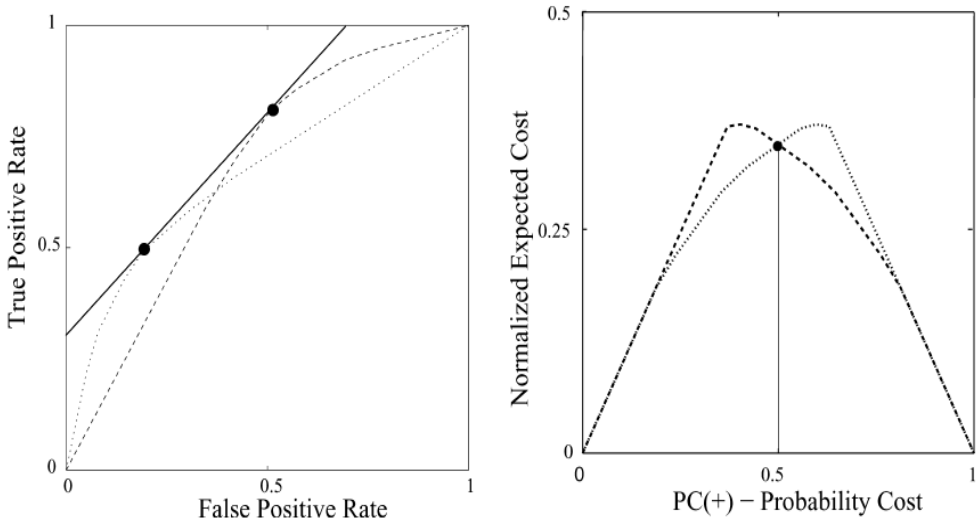
Οι Drummond και Holte (2006) πρότειναν ως μέτρο αξιολόγησης των ταξινομητών την καμπύλη κόστους. Ο άξονας x αντιπροσωπεύει την αναλογία της θετικής κλάσης στο σύνολο εκπαίδευσης και ο άξονας y αντιπροσωπεύει το ποσοστό του σφάλματος που δημιουργείται από τα σύνολα εκπαίδευσης. Τα σύνολα εκπαίδευσης για ένα σύνολο δεδομένων παράγονται από υπό (ή υπέρ) δειγματοληψία. Τα ποσοστά σφάλματος για την κατανομή των κλάσεων που δεν εκπροσωπούνται, ερμηνεύονται με παρεμβολή. Για έναν αλγόριθμο μηχανικής μάθησης καθορίζονται δύο συστατικά ευαίσθητου-κόστους

1) την παραγωγή μιας ποικιλίας ταξινομητών που ισχύουν για διάφορες κατανομές και
2) την επιλογή του κατάλληλου ταξινομητή για τη σωστή κατανομή.

Ωστόσο, όταν είναι γνωστά τα κόστη εσφαλμένης ταξινόμησης, ο x-άξονας μπορεί να αντιπροσωπεύει τη «συνάρτηση κόστους πιθανότητας», η οποία είναι το κανονικοποιημένο γινόμενο του

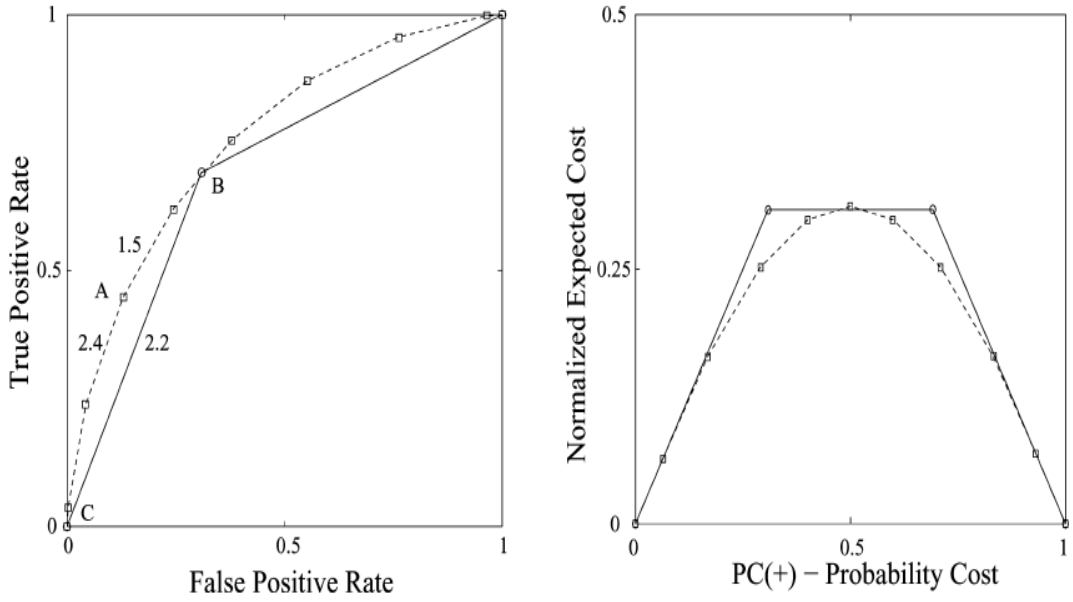
$$C\ (-\ |\ +) *\ P(+)$$

Ο y-άξονας αντιπροσωπεύει το αναμενόμενο κόστος.



Σχήμα 4.5: (a) Δύο καμπύλες ROC που διασταυρώνονται – (b) Αντίστοιχες καμπύλες κόστους

Βασικά, η καμπύλη κόστους εξετάζει πως οι ταξινομητές εκτελούνται σε ένα ευρύ φάσμα από διαφορετικά κόστη εσφαλμένης ταξινόμησης. Αυτό μπορεί να γίνει φανερό ως διαφορετική κλίση της γραμμής της εφαπτομένης στην καμπύλη ROC, ως εκ τούτου, κάθε καμπύλη ROC έχει μια αντίστοιχη καμπύλη του κόστους Drummond και Holte (2006). Ακολουθεί άλλο ένα παράδειγμα.



Σχήμα 4.6 Δύο καμπύλες ROC – (b) Αντίστοιχες καμπύλες κόστους

4.8.2. Cost matrix

Πολλάκις, όπως ήδη έχουμε επισημάνει, το μέτρο απόδοσης ενός αλγορίθμου μηχανικής μάθησης βασίζεται στην ακρίβεια της ταξινόμησης ενός συνόλου δεδομένων. Ενώ αυτό είναι ένα χρήσιμο μέτρο, μπορεί να είναι σημαντικό να ληφθούν επιπλέον παράγοντες υπόψη. Στην ταξινόμηση των δεδομένων, ορισμένοι τύποι εσφαλμένων ταξινομήσεων μπορεί να είναι χειρότεροι από κάποιους άλλους. Για παράδειγμα, η απόρριψη μιας έγκυρης συναλλαγής με πιστωτική κάρτα μπορεί να προκαλέσει ενόχληση, ενώ για η έγκριση μίας τεράστιας δόλιας συναλλαγής μπορεί να έχει πολύ αρνητικές συνέπειες. Σε καταστάσεις όπως αυτή, είναι σημαντικό να ληφθεί υπόψη το κόστος του κάθε τύπου σφάλματος, έτσι ώστε να αποφευχθούν τα δαπανηρότερα των σφαλμάτων.

Για τα μέτρα απόδοσης που ανήκουν στην κατηγορία των μέτρων ευαίσθητου κόστους συνήθως υποθέτουμε ότι τα κόστη δημιουργίας ενός σφάλματος είναι γνωστά (Turney, 2000, Domingos, 1999, Elkan, 2001). Ο πίνακας κόστους, περιέχει τα κόστη που είναι γνωστά για το πρόβλημα της ταξινόμησης, δηλαδή τα κόστη εσφαλμένης ταξινόμησης ενός θετικού ή αρνητικού παραδείγματος. Αυτό που απεικονίζει είναι το κόστος της εσφαλμένης ταξινόμησης. Κάθε παράδειγμα, x, μπορεί να συνδέεται με ένα κόστος $C\ (i, j, x)$, το οποίο καθορίζει το κόστος της προβλεπόμενης κλάσης i για το x, όταν η "αληθής" τάξη είναι η τάξη j. Στόχος είναι να λάβουμε μια απόφαση για την ελαχιστοποίηση του αναμενόμενου κόστους. Η βέλτιστη πρόβλεψη για το x μπορεί να οριστεί ως

$$\sum_j P(j|x)C(i,j,x)$$

Η παραπάνω εξίσωση απαιτεί υπολογισμό των δεσμευμένων (conditional) πιθανοτήτων της κατηγορίας j δοθέντος του διανύσματος των χαρακτηριστικών ή του παραδείγματος x. Ενώ η εξίσωση του κόστους είναι απλή, δεν έχουμε πάντα ένα κόστος που συνδέεται με την πραγματοποίηση ενός σφάλματος. Το κόστος μπορεί να είναι διαφορετικό για κάθε παράδειγμα και όχι μόνο για κάθε τύπο σφάλματος. Έτσι, το $C\ (i, j)$ δεν είναι πάντοτε ίσο με το $C\ (i, j, x)$.

Ο πίνακας Κόστους

Να παρουσίαση των διαφορετικών τύπων του κόστους κάθε τύπου εσφαλμένης ταξινόμησης, μπορεί να χρησιμοποιηθεί ένας πίνακας κόστους (Elkan (2001)). Τυπικά, κάθε σειρά του πίνακα χρησιμοποιείται για να αντιπροσωπεύσει την ετικέτα που έχει προβλεφθεί (predicted) και κάθε στήλη αντιστοιχεί στην πραγματική (actual) ετικέτα. Η είσοδος του πίνακα $C_{\{ij\}}$ είναι το κόστος της πρόβλεψης της i-οστής ετικέτας όταν η j-οστή ετικέτα είναι πράγματι σωστή. Γενικά, $C_{\{ij\}} > C_{\{jj\}}$ όταν $i \neq j$, δηλαδή μια σωστή

πρόβλεψη κοστίζει λιγότερο από ότι μια λανθασμένη πρόβλεψη. Συχνά οι καταχωρήσεις $C_{\{jj\}}$ κατά μήκος της κυρίας διαγωνίου θα είναι όλες μηδέν.

Αποτέλεσμα του τεστ	Πραγματική τιμή: Actual negative	Πραγματική τιμή: Actual positive
Predict negative	$\boldsymbol{C(0,0) = c_{00}}$	$\boldsymbol{C(0,1) = c_{01}}$
predict positive	$\boldsymbol{C(0,0) = c_{10}}$	$\boldsymbol{C(1,1) = c_{11}}$

Λαμβάνοντας βέλτιστες αποφάσεις

Οταν λαμβάνουμε υπόψη το κόστος εσφαλμένης ταξινόμησης, η βέλτιστη ετικέτα είναι η i-οστή ετικέτα που θα μπορούσε να ελαχιστοποιήσει την ακόλουθη συνάρτηση (Elkan (2001)):

$$\text{L (x, i)} = \sum_j P(j|x)C(i,j)$$

Με τον πολλαπλασιασμό του κόστους μίας πραγματικής ετικέτας δοθείσας μιας προβλεπόμενης ετικέτας, επί την πιθανότητα πραγματοποίησης αυτής της πραγματικής ετικέτας και αθροίζοντας μπορούμε να καθορίσουμε το αναμενόμενο κόστος για κάθε προβλεπόμενη ετικέτα. Τα υψηλά κόστη, υπό ορισμένες λανθασμένες ταξινομήσεις μπορεί να καθιστούν βέλτιστο να αποδόσουμε μια ετικέτα που δεν είναι η πιο πιθανή. Σε ορισμένες περιπτώσεις, μια ετικέτα δεν μπορεί να ανατεθεί κάτω από οποιεσδήποτε συνθήκες. Η mth γραμμή του πίνακα κόστους κυριαρχεί της $n - οστής$ γραμμής, αν για κάθε j, $C_{mj} \geq C_{nj}$. Σε αυτή την περίπτωση, η mth ετικέτα δεν θα εκχωρηθεί ποτέ, επειδή η $n - οστή$ ετικέτα θα είναι πάντα φθηνότερη.

Συνδυάζοντας τον πίνακα κόστους με ταξινομητές

Οι περισσότεροι ταξινομητές μηχανκής μάθησης είναι κατασκευασμένοι να μεγιστοποιούν την ακρίβεια (accuracy) ταξινόμησης. Προκειμένου να ληφθούν υπόψη τα κόστη της εσφαλμένης ταξινόμησης, πρέπει να γίνουν προσαρμογές στη διαδικασία της μάθησης. Υπάρχουν τρεις κύριες μέθοδοι για την ενσωμάτωση της ευαισθησίας του κόστους στη διαδικασία της μάθησης (Margineantu, (2002)).

1. Να χρησιμοποιήσουμε ένα ταξινομητή που παρέχει εκτιμήσεις πιθανότητας της κλάσης και υπολογίζει το αναμενόμενο κόστος της κάθε ετικέτας.
2. Να χειριστούμε τα δεδομένα εκπαίδευσης έτσι ώστε να έχουν μια κατανομή κλάσης που ταιριάζει με την κατανομή του κόστους από την υπερδειγματοληψία και την υποδειγματοληψία από κάποιες κλάσεις.

3. Να αλλάξουμε τις εσωτερικές λειτουργίες ενός ταξινομητή, έτσι ώστε ο πίνακας του κόστους να χρησιμοποιείται μέσα στον ίδιο τον αλγόριθμο.

Παρακάτω παραθέτουμε ένα αντιπροσωπευτικό παράδειγμα για το πώς η ευαισθησία του κόστους μπορεί να εφαρμοστεί στα δέντρα αποφάσεων που αποτελούν έναν από τους πιο συνηθισμένους ταξινομητές μηχανικής μάθησης.

Δέντρα αποφάσεων

Ένας ταξινομητής που βασίζεται στα δέντρα απόφασης μπορεί να επηρεάζεται από την τροποποίηση της κατανομής της κλάσης του συνόλου εκπαίδευσης. Ένα δέντρο απόφασης κατασκευάζεται συνήθως σε δύο περάσματα: την δημιουργία/ανάπτυξη και το κλάδεμα. Η φάση ανάπτυξης είναι σχετικά ανεπηρέαστη από μεταβολές στις αναλογίες των κλάσεων. Η φάση του κλαδέματος, ωστόσο, μπορεί να επηρεαστεί σημαντικά [1]. Οι αναλογίες των κλάσεων είναι σημαντικές για πολλούς αλγορίθμους κλαδέματος, όπως ο C4.5, και μπορούν ακόμη και να κόψουν το δέντρο κάτω σε ένα μόνο κόμβο, εάν όλες οι κλάσεις, εκτός από μία είναι σπάνιες.

Είναι επίσης δυνατόν να χρησιμοποιήσουμε ένα δέντρο απόφασης για την απόκτηση πρόχειρων εκτιμήσεων πιθανοτήτων (Margineantu, D. (2002)). Όταν χρησιμοποιείται με αυτόν τον τρόπο, η κατανομή των κλάσεων δεν πρέπει να τροποποιηθεί. Ο τύπος για να λάβουμε μία πιθανότητα είναι:

$$P\ (j|x) = \frac{N_j(D_x)}{N\ (D_x)}$$

όπου j είναι η κλάση για να εκτιμηθεί η πιθανότητα, το x είναι το παράδειγμα που προαναφέραμε, και D_x είναι το φύλλο που επιτεύχθηκε από το παράδειγμα x στο δέντρο. Το N_j επιστρέφει τον αριθμό των παραδειγμάτων εκπαίδευσης της ετικέτας j στο φύλο D_x. Το N επιστρέφει το συνολικό αριθμό των παραδειγμάτων εκπαίδευσης στο φύλλο D_x. Αυτό μπορεί να είναι πολύ ανακριβές καθώς οι κόμβοι του φύλλου συνήθως έχουν μικρά σύνολα δεδομένων και είναι πολύ μονόπλευροι στις κατανομές των κλάσεων. Για να αντισταθμιστεί αυτή την ανακρίβεια, μπορεί να εφαρμοστεί μία διόρθωση Laplace ως εξής:

$$P\ (j\mid x)\ = \frac{N_j(D_x) + \ \lambda_j}{N\ (D_x) + \ \sum_{i=1}^{K} \lambda_i}$$

ΚΕΦΑΛΑΙΟ 5

Πειραματικά Αποτελέσματα
(Experimental Results)

Στην πέμπτη και τελευταία ενότητα διεξάγουμε μία μελέτη σε 5 διαφορετικά σύνολα δεδομένων με στόχο τη σύγκριση των διαφορετικών μεθόδων ταξινόμησης. Πρόκειται για σύνολα δεδομένων τα οποία είναι μη ισορροπημένα και ως εκ τούτου η ανάγκη για χρήση μεθόδων «εξισορρόπησης» είναι επιτακτική.

Στην πρώτη παράγραφο πραγματοποιούμε την ανάλυση τεσσάρων συνόλων δεδομένων τα οποία είναι σχετικά μικρών διαστάσεων, δηλαδή αποτελούνται από λίγες επεξηγηματικές μεταβλητές, καθώς περιλαμβάνουν και λίγες πειραματικές εκτελέσεις. Μόνη εξαίρεση αποτελεί το τέταρτο σύνολο δεδομένων το οποίο αποτελεί υποσύνολο ενός μεγαλύτερου συνόλου αλλά περιλαμβάνει ένα αρκετά μεγάλο αριθμό πειραματικών εκτελέσεων. Όλα τα τέσσερα σύνολα περιλαμβάνουν δεδομένα από τον ιατρικό κλάδο.

Στην δεύτερη παράγραφο πραγματοποιούμε την ανάλυση ενός συνόλου δεδομένων μεγάλων διαστάσεων. Πρόκειται για ένα σύνολο πραγματικών, ιατρικών δεδομένων με 40 επεξηγηματικές μεταβλητές και μία δίτιμη μεταβλητή απόκρισης που αποτελείται δηλαδή από δύο κλάσεις. Η ανισορροπία μεταξύ των κλάσεων είναι διακριτή και για το λόγο αυτό είναι απαραίτητη η χρήση μεθόδων για το χειρισμό τέτοιων δεδομένων. Επιπρόσθετα, λόγω της φύσης των δεδομένων (ιατρικά δεδομένα), αναλογιζόμενοι την σημαντικότητα μιας ασφαλούς πρόβλεψης, η ανάγκη για τη χρήση μεθόδων εξισορρόπησης καθώς και εύρωστων, πιο ευαίσθητων μέτρων διάκρισης, γίνεται ακόμη μεγαλύτερη. Η ανάλυση των δεδομένων είναι ελαφρώς διαφορετική από τα προηγούμενα σύνολα αφού εδώ καλούμαστε να αντιμετωπίσουμε δύο προβλήματα. Αφενός την ανισορροπία μεταξύ των κλάσεων και αφετέρου τη μεγάλη διάσταση των δεδομένων.

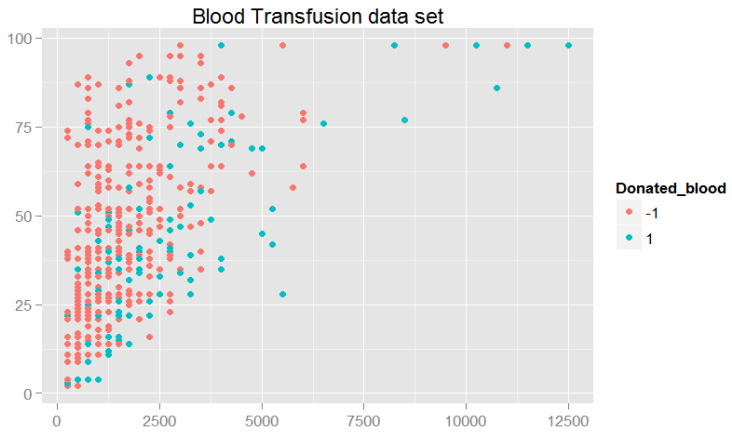
Στη συνέχεια παρουσιάζουμε τα αποτελέσματα της διαδικασίας μοντελοποίησης. Για να παρέχουμε μία αμερόληπτη εκτίμηση για την ποιότητα ταξινόμησης του κάθε μοντέλου υπολογίζουμε τις τιμές των κριτηρίων απόδοσης σ' ένα σύνολο δεδομένων που δεν χρησιμοποιήθηκε στη διαδικασία μοντελοποίησης. Για το σκοπό αυτό χρησιμοποιήσαμε από το πραγματικό σύνολο δεδομένων, ένα μέρος (το σύνολο δοκιμής) το οποίο χρησιμοποιήθηκε αργότερα για αυτό το σκοπό. Θα συγκρίνουμε τα αποτελέσματα στηριζόμενοι όχι μόνο στην συνολική ακρίβεια (ACC), την ευαισθησία (sensitivity) και την ειδικότητα (specificity) του ταξινομητή, αλλά και σε πιο εύρωστα μέτρα όπως είναι η περιοχή κάτω από τη ROC καμπύλη και ο γεωμετρικός μέσος. Η γενικευμένη απόδοση εκτιμάται με τη μέθοδο επικύρωσης 10-fold cross validation.

5.1 Εφαρμογή σε μικρά σύνολα δεδομένων

Σε αυτή την ενότητα εφαρμόζουμε τις μεθόδους SVM και PSVM όπως επίσης και τις τροποποιημένες αυτών μεθόδους για μη ισορροπημένα δεδομένα TCSVM και MPSVM, σε τέσσερα σύνολα δεδομένων όλα εκ των οποίων σχετίζονται με την ιατρική επιστήμη. Τα σύνολα δεδομένων περιγράφονται στη συνέχεια:

Δεδομένα για τη μετάγγιση αίματος (Blood Transfusion)

Αυτή η βάση δεδομένων είναι διαθέσιμη στη βιβλιοθήκη Μηχανικής Μάθησης, UCI (UCI Machine Learning Repository (Bache Lichman, M. (2013)). Στο δείγμα υπάρχουν 748 δότες που επιλέγονται τυχαία από τη συγκεκριμένη βάση δεδομένων.

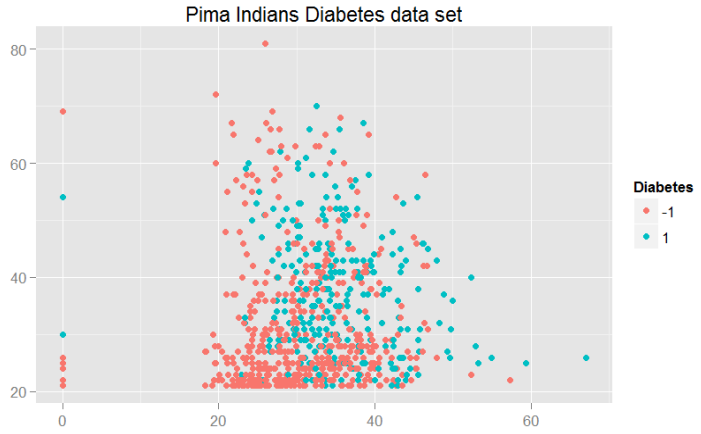


Σχήμα 5.1: Κατανομή των κλάσεων στο Blood Transfusion σύνολο δεδομένων (οι κόκκινες κουκίδες αναφέρονται στην κλάση -1(πλειοψηφία), οι μπλε κουκίδες στην κλάση 1(μειοψηφία))

Κάθε ένας από τους 748 δότες, περιλαμβάνουν στοιχεία σχετικά με την συχνότητα(recency) -μήνες από την τελευταία δωρεά-, συχνότητα(frequency) -συνολικός αριθμός των δωρεών-, την αξία (monetary) -συνολικό αίμα που δώρισε σε c.c.-, χρόνο(time) -μήνες από την πρώτη δωρεά- και μια δυαδική μεταβλητή η οποία αναπαριστά εάν αυτός/αυτή έδωσε αίμα το Μάρτιο του 2007 (1 σημαίνει δωρεά αίματος, -1 σημαίνει όχι δωρεά αίματος).

Δεδομένα για το Διαβήτη στους Ινδιάνους Pima[11] (Pima Indians Diabetes data set)

Αυτή η βάση δεδομένων είναι επίσης διαθέσιμη στη βιβλιοθήκη Μηχανικής Μάθησης, UCI, (UCI Machine Learning Repository (Bache Lichman, M. (2013)). Σύμφωνα με τους δημιουργούς αυτού του συνόλου δεδομένων, τοποθετήθηκαν αρκετοί περιορισμοί για την επιλογή αυτών των περιπτώσεων από μια μεγαλύτερη βάση δεδομένων. Συγκεκριμένα, όλοι οι ασθενείς είναι γυναίκες τουλάχιστον 21 ετών από το γένος των Ινδιάνων Pima.



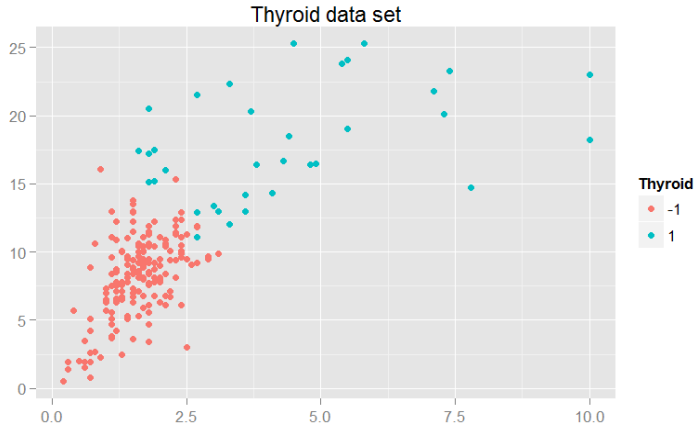
Σχήμα 5.2: Κατανομή των κλάσεων στο Pima Indians Diabetes σύνολο δεδομένων (οι κόκκινες κουκίδες ανφέρονται στην κλάση -1(πλειοψηφία), οι μπλε κουκίδες στην κλάση 1(μειοψηφία))

Στον καθένα από τους 768 ασθενείς, περιλαμβάνεται ο αριθμός των φορών που έμειναν έγκυος, η συγκέντρωση γλυκόζης πλάσματος 2 ωρών σε ένα τεστ ανοχής γλυκόζης από το στόμα, η διαστολική πίεση του αίματος (mm Hg), το πάχος του δέρματος στους τρικέφαλους (mm), η 2-ωρών ορός ινσουλίνης (MU U / ml), ο δείκτης μάζας σώματος (βάρος σε kg / (ύψος σε m) ^ 2), η γενεαλογία της συνάρτησης του διαβήτης, η ηλικία (έτη) και μία δυαδική μεταβλητή που αντιπροσωπεύει αν αυτοί ήταν θετικοί/αρνητικοί στο διαβήτη (τιμή κατηγορίας 1 ερμηνεύεται ως «θετικοί στο διαβήτη").

[11] Η Pima / pi:mə / [3] (ή Akimel O'odham, ή αλλιώς, "River Άνθρωποι", απλά γνωστοί ως Pima) είναι μια ομάδα των Αυτοχθόνων Αμερικανών που ζουν σε μια περιοχή που αποτελεί σήμερα την κεντρική και νότια Αριζόνα.

Δεδομένα για νόσο του θυρεοειδούς (Thyroid Disease data set)

Μία μη ισορροπημένη έκδοση του συνόλου δεδομένων της νόσου του θυρεοειδούς, όπου τα θετικά παραδείγματα ανήκουν στην δεύτερη κλάση (υπερθυρεοειδισμός) και τα αρνητικά παραδείγματα ανήκουν στο υπόλοιπο.

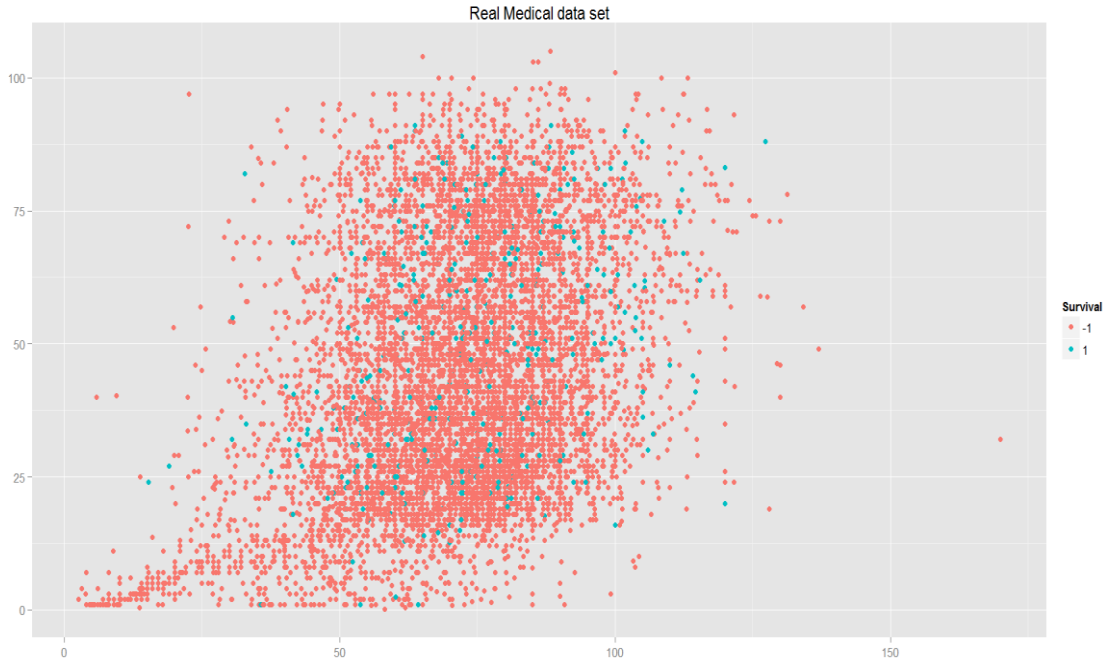


Σχήμα 5.3: Κατανομή των κλάσεων στο Thyroid σύνολο δεδομένων (οι κόκκινες κουκίδες ανφέρονται στην κλάση -1(πλειοψηφία), οι μπλε κουκίδες στην κλάση 1(μειοψηφία))

Υπάρχουν πέντε επεξηγηματικές μεταβλητές οι T3resin, Θυροξίνη, τριιωδοθυρονίνη, Thyroidstimulating και TSH_value, και 215 ασθενείς, εκ των οποίων 35 είναι θετικές στο θυρεοειδή και 180 είναι αρνητικές. Το σύνολο δεδομένων της νόσου του θυρεοειδούς είναι διαθέσιμο στη βιβλιοθήκη Μηχανικής Μάθησης, UCI, (UCI Machine Learning Repository (Bache Lichman, M. (2013)). Ωστόσο, τα σημερινά δεδομένα αποτελούν τροποποίηση των αρχικών δεδομένων που παρέχονται από τη βιβλιοθήκη KEEL. Για περισσότερες λεπτομέρειες σχετικά με αυτό το σύνολο δεδομένων βλέπε Παράρτημα Β.

Πραγματικά Ιατρικά Δεδομένα (Real Medical Data)

Πρόκειται για ένα σύνολο δεδομένων που αποτελεί υποβάση ενός μεγαλύτερου συνόλου που παρουσιάζεται στη συνέχεια και το οποίο αναλύθηκε χρησιμοποιώντας επιπρόσθετες μεθόδους των οποίων η χρήση ήταν επιτακτική λόγω της πολυπλοκότητας και του μεγέθους του. Για κάθε ασθενή η δυαδική μεταβλητή απόκρισης Y υποδηλώνει την πιθανότητα του θανάτου. Λήφθηκαν υπόψη 14 παράγοντες οι οποίοι περιλαμβάνουν δημογραφικά χαρακτηριστικά, στοιχεία μεταφορών καθώς και ενδονοσοκομειακά στοιχεία. Συνολικά 8862 ασθενείς ήταν διαθέσιμοι. Συγκεκριμένα, η δυαδική μεταβλητή y, εκφράζεται με τη μορφή δύο κατηγοριών -1 και 1, όπου -1 αντιπροσωπεύουν την επιβίωση, ενώ η τιμή 1 το θάνατο.



Σχήμα 5.4: Κατανομή των κλάσεων στο Real Medical σύνολο δεδομένων (οι κόκκινες κουκίδες ανφέρονται στην κλάση -1(πλειοψηφία), οι μπλε κουκίδες στην κλάση 1(μειοψηφία))

Σύμφωνα με ιατρικές συμβουλές, όλοι οι προγνωστικοί παράγοντες θα πρέπει να αντιμετωπίζονται ισότιμα κατά την στατιστική ανάλυση και δεν υπάρχει κανένα στοιχείο που θα πρέπει να διατηρείται πάντα στο μοντέλο. Τα ονόματα αυτών των παραγόντων περιλαμβάνονται στο Παράρτημα και είναι όλες συνεχείς μεταβλητές.

Πριν κάνουμε τις κατάλληλες συγκρίσεις των αλγορίθμων χωρίσαμε το σύνολο δεδομένων τυχαία σε σύνολο εκπαίδευσης και σύνολο ελέγχου. Στη συνέχεια, χρησιμοποιήσαμε ένα κλάσμα του συνόλου εκπαίδευσης για την εκπαίδευση αυτών των αλγορίθμων και ένα σύνολο δοκιμών για την αξιολόγηση της απόδοσης των ταξινομητών σε νέα δεδομένα. Για το διαχωρισμό αυτό, οι συνολικές παρατηρήσεις επιλέχθησαν τυχαία χωρίς επανάθεση για να δημιουργήσουμε τα σύνολα εκπαίδευσης και δοκιμής, σύμφωνα με το προκαθορισμένο μέγεθος τους, που περιέχει 75% και 25% των περιπτώσεων, αντίστοιχα. Προκειμένου να αξιολογήσουμε τις καλύτερες τιμές στο στάδιο της εκπαίδευσης των δεδομένων και για την επικύρωση των πειραματικών αποτελεσμάτων επιλέξαμε ως μέθοδο επικύρωσης, την διασταυρωμένη επικύρωση (10-fold cross validation). Η ανάλυση πραγματοποιήθηκε χρησιμοποιώντας κωδικούς της R.

Στο πρώτο στάδιο της ανάλυσης, πραγματοποιήσαμε μια επιλογή μοντέλου για τις εγγενείς παραμέτρους των μεθόδων των μηχανών διανυσματικής υποστήριξης, δεδομένου ότι αυτές οι παράμετροι επηρεάζουν πάρα πολύ τη συνολική απόδοση των ταξινομητών.

Πίνακας 5.1: Περιγραφή των τεσσάρων συνόλων δεδομένων

Σύνολο	**Περιγραφή**	**Αριθμός δειγμάτων:**	**Αριθμός Μεταβλητών:**	**Κατανομή των κλάσεων:**	
Blood Transfusion ***(UCI)***	Ταξινομεί εάν κάποιος έδωσε ή όχι αίμα	748	**4** + class	*Class 1* **178** / *Class -1* **570** (IR=0.31228)	1: έδωσαν αίμα -1: δεν έδωσαν αίμα
Pima Indians Diabetes ***(UCI)***	Προβλέπει εάν ο ασθενής έχει σημάδια διαβήτη	768	**8** + class	*Class 1* **268** / *Class -1* **500** (IR=0.536)	1: θετικός στο διαβήτη -1: αρνητικός στο διαβήτη
Thyroid1 ***(KEEL)***	Προβλέπει εάν ο ασθενής έχει σημάδια θυρεοειδή	215	**5** + class	*Class 1* **35** / *Class -1* **180** (IR=0.1944)	1: θετικός στο θυρεοειδή -1: αρνητικός στο θυρεοειδή
Real medical	Προβλέπει εάν ο ασθενής θα επιβιώσει	8862	**14** + class	*Class 1* **446**/ *Class -1* **8416** (IR=0.0529)	1 θάνατος -1 επιβίωση

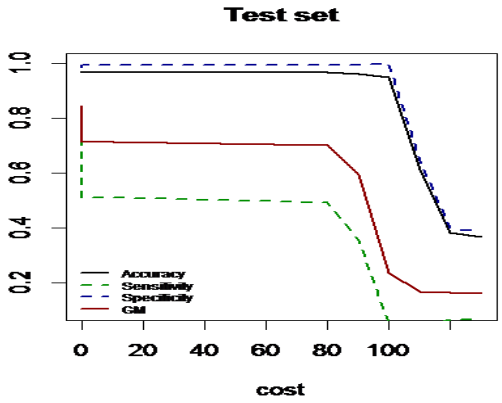
*Για περισσότερες λεπτομέρειες σχετικά με τις μεταβλητές των 4 συνόλων δεδομένων που χρησιμοποιούνται στη μελέτη μας παραπέμπουμε στο Παράρτημα Β

Ακολουθήσαμε την ίδια διαδικασία για τις παραμέτρους των μεθόδων μάθησης για τα μη ισορροπημένα δεδομένα που εφαρμόζονται στη μελέτη μας. Στη συνέχεια παρουσιάζουμε τις συγκρίσεις μεταξύ της προαναφερθείσας μεθοδολογίας για τις μηχανές διανυσματικής υποστήριξης όσον αφορά την ακρίβεια και το γεωμετρικό μέσο λαμβάνοντας υπόψη τόσο την ευαισθησία και την ειδικότητα, όσο και το χρόνο εκτέλεσης των μεθόδων αυτών.

5.1.1. Γραμμική περίπτωση

Υπάρχει μόνο μία παράμετρος, εκείνη του κόστους C που θα πρέπει να προσδιοριστεί στην περίπτωση του γραμμικού SVM και PSVM ταξινομητή. Πραγματοποιήσαμε μια έρευνα μεταξύ πολλών υποψηφίων προκειμένου να προσδιοριστεί η βέλτιστη τιμή της C που βελτιώνει την ακρίβεια της ταξινόμησης ή ισοδύναμα ελαχιστοποιεί το σφάλμα εσφαλμένης ταξινόμησης. Ωστόσο, παρατηρήσαμε ότι οι διαφορές στις επιδόσεις ήταν αμελητέες με την αλλαγή αυτής της παραμέτρου.

Ως εκ τούτου επιλέξαμε την παράμετρο αυτή ίση με το 1, έτσι ώστε να αποφευχθούν προβλήματα υπερπροσαρμογής.



Σχήμα 5.5: Μέτρα για διαφορετικές τιμές της παραμέτρου του κόστους (το κόκκινο είναι για τον GM, το μαύρο για την ακρίβεια, το μπλε για την ειδικότητα και το πράσινο για την ευαισθησία) χρησιμοποιώντας το γραμμικό SVM. Στον παραπάνω πίνακα παρουσιάζονται τα αποτελέσματα στο σύνολο δοκιμών.

Το Σχήμα 5.1 απεικονίζει τις αλλαγές στην ακρίβεια, την ευαισθησία, την εξειδίκευση και τον GM καθώς αυξάνεται η τιμή του C. Οι τιμές για το C (x-άξονας) κυμαίνονται από 0,01 έως 10^{13}. Όπως μπορούμε να συμπεράνουμε από το Σχήμα 1, τα πιο ακριβή αποτελέσματα, τα οποία είναι σχεδόν όμοια, δίνονται με την παράμετρο του κόστους να κυμαίνεται από 0,01 έως 10.

Πίνακας 5.2: PSVM and SVM training and testing correctness and running times using a linear classifier. Execution times include ten-fold training. Best results are in bold**.**

Σύνολο Δεδομένων	Γραμμική Μοντελοποίηση	Ακρίβεια	Ευαισθησία	Ειδικότητα	GM	Χρόνος
Blood Transfusion	SVM (Train,Test)	(0.768,0.763)	(0.0373,0.023)	(0.9976,0.992)	(0.192, 0.150)	0.29
	PSVM (Train,Test)	(**0.774, 0.768**)	(0.089, 0.045)	(0.9883, 0.993)	(**0.297,0.212**)	**0.14**
Pima Indians Diabetes	SVM (Train,Test)	(**0.786, 0.781**)	(0.5622, 0.537)	(0.907,0.912)	(0.713, **0.700**)	0.36
	PSVM (Train,Test)	(**0.786**,0.7604)	(0.5821,0.537)	(0.896, 0.880)	(**0.722**, 0.687)	**0.11**
Thyroid	SVM (Train,Test)	(**1.00 ,0.9434**)	(1.00 ,0.750)	(1.00 ,0.978)	(**1.00, 0.856**)	0.10
	PSVM (Train,Test)	(0.944, 0.925)	(0.667, 0.500)	(1.00, 1.00)	(0.816 ,0.707)	**0.08**
Real Medical	SVM (Train,Test)	(**0.975 , 0.9756**)	(0.719, 0.712)	(0.988, 0.989)	(**0.843, 0.839**)	10.21
	PSVM (Train,Test)	(0.969, 0.970)	(0.531, 0.559)	(0.993, 0.992)	(0.726, 0.744)	**5.60**

Με την προϋπόθεση ότι είχε αρκετά καλή επίδοση, τελικά ορίσαμε C=1, έτσι ώστε να μειωθεί ταυτόχρονα και η πολυπλοκότητα και το πρόβλημα της υπερπροσαρμογής των δεδομένων.

Στον Πίνακας 5.2 παρουσιάζεται μία σύγκριση των δύο μεθόδων, SVM και PSVM, χρησιμοποιώντας ένα γραμμικό ταξινομητή για τα τέσσερα διαφορετικά σύνολα δεδομένων. Οι συγκρίσεις που λαμβάνουν χώρα, γίνονται λαμβάνοντας υπόψη τόσο την ακρίβεια και το γεωμετρικό μέσο όσο και το χρόνο εκτέλεσης των αντίστοιχων μεθόδων. Με έντονο χρώμα παρουσιάζουμε τα καλύτερα αποτελέσματα. Οι συγκρίσεις με βάση το χρόνο δείχνουν ότι ο ταξινομητής PSVM έχει μια ξεκάθαρη υπεροχή στην ταχύτητα. Αυτό γίνεται ιδιαίτερα εμφανές στην περίπτωση του ιατρικού συνόλου δεδομένων όπου το PSVM έλυσε το πρόβλημα ταξινόμησης σε 5,60 δευτερόλεπτα, δηλαδή σχεδόν στο μισό χρόνο από το SVM που το έλυσε σε 10,21 δευτερόλεπτα.

Είναι σαφές, από τα παραπάνω αποτελέσματα ότι το γραμμικό SVM έχει καλύτερη ακρίβεια ταξινόμησης, καθώς και καλύτερο γεωμετρικό μέσο για το σύνολο δεδομένων του Διαβήτη στους Ινδιάνους Pima, στην ασθένεια του θυρεοειδούς και στην πραγματική ιατρική βάση δεδομένων. Ωστόσο, στην περίπτωση του συνόλου δεδομένων για τη μετάγγιση αίματος, το γραμμικό PSVM είχε καλύτερη επίδοση τόσο στο γεωμετρικό μέσο όσο και στην ακρίβεια ταξινόμησης. Ωστόσο, όπως έχουμε ήδη αναφέρει, ο χρόνος που χρησιμοποιήθηκε για να εκπαιδευτεί το μοντέλο είναι πολύ μικρότερος στην περίπτωση του PSVM.

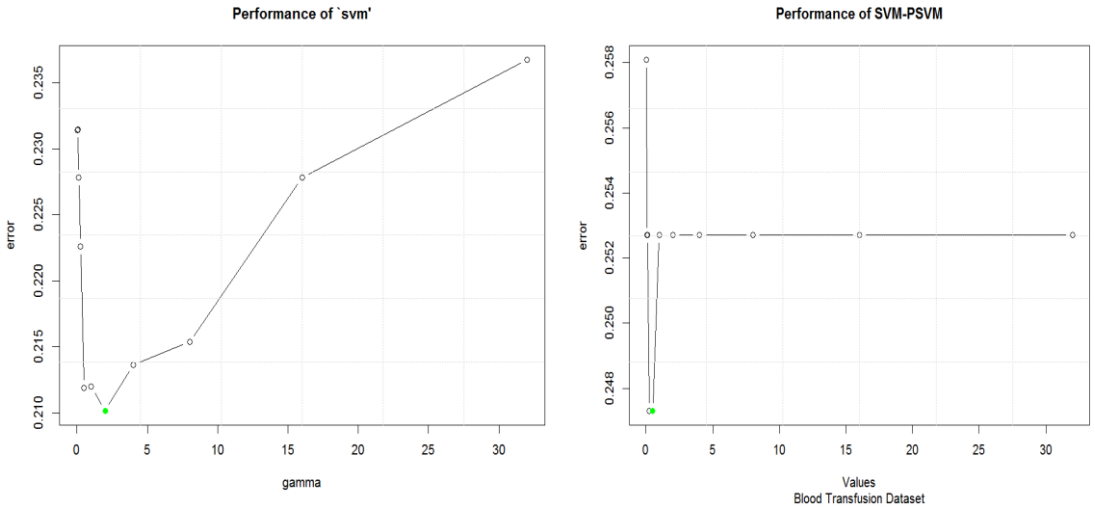
5.1.2. Μη γραμμική περίπτωση (Non-linear Case)

Εκτός από τις γραμμικό θα χρησιμοποιήσουμε τον Γκαουσιανό (Gaussian) πυρήνα έτσι ώστε να αξιολογηθεί η απόδοση των μεθόδων για το μη γραμμικό πρόβλημα ταξινόμησης. Στην εκπαίδευση του SVM εκτός από τη λειτουργία του πυρήνα, θα πρέπει να καθορίσουμε παράμετρο κανονικοποίησης C, η οποία ελέγχει το trade-off μεταξύ της μεγιστοποίησης του περιθωρίου και της ελαχιστοποίηση του σφάλματος εκπαίδευσης, καθώς και τις εγγενείς παραμέτρους της συνάρτησης του πυρήνα. Η τιμή της παραμέτρου του πυρήνα (g) σχετίζεται με την εξάπλωση των παρατηρήσεων και χρησιμοποιείται στον Γκαουσιανό πυρήνα. Έτσι, για να αναπτύξουμε το βέλτιστο ταξινομητή θα πρέπει να ορίσουμε το βέλτιστο σχέδιο όσον αφορά στις τιμές των παραμέτρων C και των παραμέτρων του πυρήνα. Παράλληλα, υπάρχει ένα πρόσθετο εμπόδιο το οποίο αναφέρεται στη διάσταση του πίνακα του πυρήνα. Για να το θέσουμε διαφορετικά, κατά το χειρισμό προβλημάτων ταξινόμησης μεγάλης κλίμακας θα πρέπει να υπολογίσουμε τον αντίστροφο του πίνακα nxn του πυρήνα. Ωστόσο, όπως έχουμε ήδη αναφέρει, υπάρχει μία δυσκολία στην εύρεση αυτού του αντίστροφου, δεδομένου ότι δεν είναι εφικτή η εύρεση του είτε σε πολλές περιπτώσεις η αποθήκευσή του. Για το Γκαουσιανό πυρήνα θα πρέπει να εκτιμηθεί τόσο το C όσο και το g. Ως εκ τούτου, κάναμε μια εκτίμηση των παραμέτρων εκτελώντας μία εκτενή αναζήτηση ανάμεσα στις πιθανές παραμέτρους (grid search). Ως μέθοδο αξιολόγησης χρησιμοποιήσαμε την 10-

fold cross validation, λαμβάνοντας υπόψη την ακρίβεια ταξινόμησης και το GM. Όπως ήταν αναμενόμενο, μεγαλύτερες τιμές της C δίνουν καλύτερη ακρίβεια. Ωστόσο, όπως ακριβώς στην γραμμική περίπτωση, χρησιμοποιώντας υψηλές τιμές της C μπορεί να προκαλέσει προβήματα υπερπροσαρμογής που οδηγούν σε υψηλότερες τιμές σφάλματος στο σύνολο ελέγχου σε σύγκριση με το σύνολο εκπαίδευσης. Θέτουμε τελικά την τιμή της παραμέτρου κόστους ίση με 1. Η τιμή g του πυρήνα, θα πρέπει κανονικά να είναι μεταξύ $1/k$ και $6/k$, όπου το k παριστά τη διάσταση των δεδομένων. Ως εκ τούτου ερευνήσαμε διάφορες τιμές στην προαναφερθείσα περιοχή και τελικά επιλέχθηκε η καλύτερη. Παρακάτω παρουσιάζουμε αυτή τη διαδικασία για κάθε ένα από τα τέσσερα σύνολα δεδομένων.

Blood Transfusion

Όπως μπορούμε να παρατηρήσουμε στο Σχήμα 5.6, θέτοντας g ίσο με 2 για το SVM και ίσο με 0,5 για το PSVM λαμβάνουμε υψηλές τιμές της ακρίβειας και χαμηλές τιμές του σφάλματος. Στη συνέχεια χρησιμοποιήσαμε αυτές τις τιμές, έτσι ώστε να κάνουμε τις απαραίτητες συγκρίσεις.

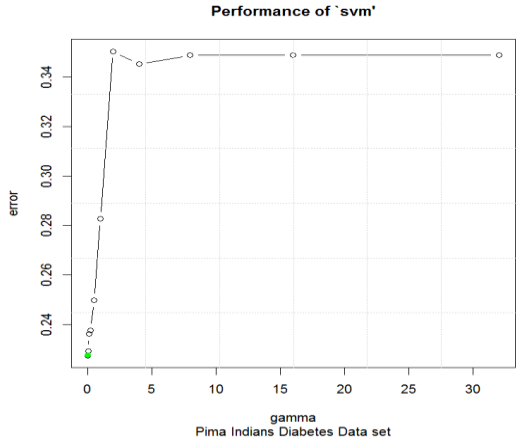


Σχήμα 5.6: Αναζήτηση πλέγματος **(**Grid search) για την παράμετρο γάμμα. Το αριστερό πάνελ αναφέρεται στην ορθότητα του συνόλου δοκιμής του μη γραμμικού SVM με Gaussian πυρήνα και ο δεξιός πίνακας αναφέρεται στο PSVM με ένα γκαουσιανό πυρήνα στο Blood transfusion σύνολο δεδομένων. Ο πράσινος συμπαγής κύκλος αντιπροσωπεύει την καλύτερη τιμή της g για το SVM (gamma=2) και για το PSVM (gamma=0.5).

Το αριστερό τμήμα του Σχήματος 5.6 παρουσιάζει την ορθότητα των αποτελεσμάτων στο σύνολο δοκιμής για το μη γραμμικό ταξινομητή SVM και η καλύτερη απόδοση σε σχέση με το ποσοστό σφάλματος εμφανίζεται με ένα πράσινο συμπαγή κύκλο. Το δεξιό πάνελ αναφέρεται στο PSVM και εδώ, όπως και προηγουμένως ένας πράσινος συμπαγής κύκλος αντιπροσωπεύει την καλύτερη τιμή του g η οποία είναι ίση με 0,5.

Pima Indians Diabetes

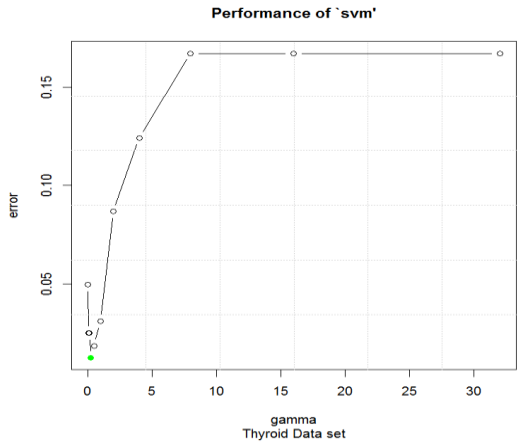
Το Σχήμα 5.7 παρουσιάζεται η ίδια διαδικασία αναζήτησης μόνο στην περίπτωση της μη γραμμικής SVM και η παράμετρο g = 0,03125 έδωσε την υψηλότερη απόδοση λαμβάνοντας υπόψη την ακρίβεια ταξινόμησης. Όσο για τη μη γραμμική PSVM επιλέξαμε ακριβώς την ίδια τιμή επειδή δεν υπήρχε σημαντική διαφορά μεταξύ των διαφόρων τιμών του g. Οι αναγκαίες συγκρίσεις έγιναν με τη χρήση της επιλεγμένης παραμέτρου g.



Σχήμα 5.7: Αναζήτηση πλέγματος **(**Grid search) για την παράμετρο γάμμα για τον SVM ταξινομητή. Παρουσιάζεται η ορθότητα του συνόλου δοκιμής του μη γραμμικού SVM με Gaussian πυρήνα στο Pima Indians Diabetes σύνολο δεδομένων (gamma=**0.03125**). Ο πράσινος συμπαγής κύκλος αντιπροσωπεύει την καλύτερη τιμή της g για το SVM.

Thyroid1

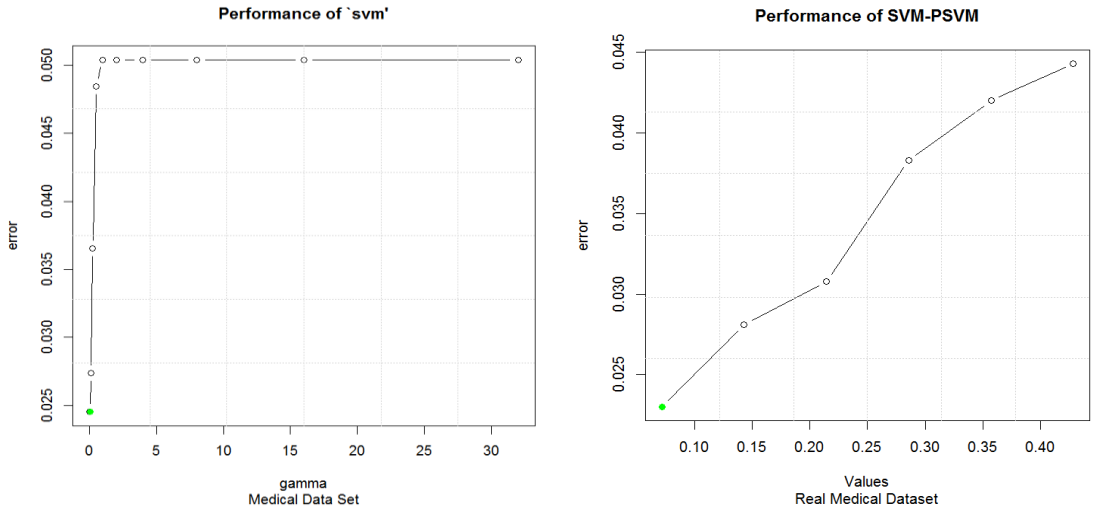
Το Σχήμα 5.8 δείχνει το σφάλμα εσφαλμένης ταξινόμησης στο σύνολο δεδομένων της ασθένειας του θυρεοειδή χρησιμοποιώντας τον μη γραμμικό SVM ταξινομητή. Η καλύτερη τιμή της g τέθηκε ίση με 0,25.



Σχήμα 5.8: Αναζήτηση πλέγματος **(**Grid search) για την παράμετρο γάμμα για τον SVM ταξινομητή. Παρουσιάζεται η ορθότητα του συνόλου δοκιμής του μη γραμμικού SVM με Gaussian πυρήνα (gamma=**0.25000)** στο Thyroid σύνολο δεδομένων. Ο πράσινος συμπαγής κύκλος αντιπροσωπεύει την καλύτερη τιμή της g για το SVM.

Real Medical Dataset

Τέλος η απόδοση των αλγορίθμων φαίνεται στην Σχήμα 5.9 και για την ιατρική βάση δεδομένων. Το αριστερό πάνελ αναφέρεται στην σφάλμα εσφαλμένης ταξινόμησης του μη γραμμικού SVM και η καλύτερη επίδοση εμφανίζεται όταν η παράμετρος g ορίστηκε ίση με 0,0625.



Σχήμα 5.9: Αναζήτηση πλέγματος **(**Grid search) για την παράμετρο γάμμα. Το αριστερό πάνελ αναφέρεται στην ορθότητα του συνόλου δοκιμής του μη γραμμικού SVM με Gaussian πυρήνα και ο δεξιός πίνακας αναφέρεται στο PSVM με ένα γκαουσιανό πυρήνα στο ιατρικό σύνολο δεδομένων. Ο πράσινος συμπαγής κύκλος αντιπροσωπεύει την καλύτερη τιμή της g για το SVM (gamma= 0.0625) και για το PSVM (gamma=0.0714).

Παράλληλα, όπως φαίνεται στον δεξί πίνακα του Σχήματος 5.9, το PSVM με τον Γκαουσιανό πυρήνα έδωσε τα καλύτερα αποτελέσματα όταν η παράμετρος g ορίστηκε ίση με 0,0714. Ο πράσινος συμπαγής κύκλος αντιπροσωπεύει την καλύτερη τιμή της g τόσο για το SVM όσο και για το PSVM. Ο Πίνακας 5.3 παρουσιάζει τη συγκριτική επίδοση των SVM και PSVM χρησιμοποιώντας ένα μη γραμμικό ταξινομητή για τα τέσσερα επιλεγμένα σύνολα δεδομένων. Παρατηρούμε ότι η μη γραμμική PSVM παρουσιάζει καλύτερες επιδόσεις στο σύνολο εκπαίδευσης και αρκετά χαμηλότερες στο σύνολο δοκιμής, κάτι που αποκαλύπτει μια υπερπροσαρμογή των δεδομένων. Ωστόσο, η μη γραμμική SVM έχει καλύτερες επιδόσεις στο σύνολο ελέγχου καθώς επίσης και στην ισορροπία μεταξύ της ορθότητας ταξινόμησης στο σύνολο εκπαίδευσης και στο σύνολο δοκιμής. Επιπλέον, για τα τρία από τα τέσσερα σύνολα δεδομένων οι επιδόσεις όσον αφορά τον χρόνο εκτέλεσης είναι πολύ καλύτερες με το μη γραμμικό SVM. Μόνο στο σύνολο δεδομένων της ασθένειας του θυρεοειδούς ο χρόνος εκτέλεσης με τη μη γραμμική PSVM είναι ελαφρώς μικρότερος από αυτόν του SVM.

Πίνακας 5.3: Ορθότητα ταξινόμησης του PSVM and του SVM στο σύνολο εκπαίδευσης και στο σύνολο δοκιμής χρησιμοποιώντας 10-fold cross validation καθώς επίσης και οι χρόνοι εκτέλεσης χρησιμοποιώντας μη γραμμικούς ταξινομητές. Οι χρόνοι περιλαμβάνουν το 10-fold training. Η τιμή του g που έδωσε την καλύτερη επίδοση χρησιμοποιήθηκε σε κάθε περίπτωση. Τα καλύτερα αποτελέσματα είναι με έντονα γράμματα.

Σύνολο Δεδομένων	Μη-Γραμμική Μοντελοποίηση	Ακρίβεια	Ευαισθησία	Ειδικότητα	GM	Χρόνος
Blood Transfusion	SVM (Train,Test)	(0.819 ,**0.79**)	(0.373,0.295)	(0.958, 0.944)	(0.598,**0.528**)	**0.55**
	PSVM (Train,Test)	(**0.927**,0.752)	(0.716,0.114)	(0.993, 0.951)	(**0.843**, 0.329)	1.94
Pima Indians Diabetes	SVM (Train,Test)	(0.795,**0.776**)	(0.571,0.552)	(0.915, 0.896)	(0.723,**0.703**)	**0.74**
	PSVM (Train,Test)	(**1.00**,0.6562)	(1.00, 0.015)	(1.00, 1.00)	(**1.00**, 0.122)	2.84
Thyroid1	SVM (Train,Test)	(**1.00,0.943**)	(1.00, 0.625)	(1.00, 1.00)	(**1.00, 0.791**)	0.13
	PSVM (Train,Test)	(1.00, 0.868)	(1.00, 0.125)	(1.00, 1.00)	(1.00, 0.354)	**0.09**
Medical Data	SVM (Train,Test)	(0.984, **0.978**)	(0.788, 0.630)	(0.995, 0.996)	(0.885,0.793)	**7.99**
	PSVM (Train,Test)	(**0.985,0. 977**)	(0.793, 0.644)	(0.995,0.994)	(**0.888, 0.800**)	623.02

Πολυωνυμικός πυρήνας

Εφαρμόσαμε τη μη γραμμική PSVM με πολυώνυμικό πυρήνα μόνο στο πραγματικό ιατρικό σύνολο δεδομένων. Για τον πολυώνυμικό πυρήνα θα πρέπει να επιλέξουμε τους βαθμούς ελευθερίας που ελέγχουν την πολυπλοκότητα (διάσταση) του χώρου χαρτογράφησης. Ο Πίνακας 5 δείχνει τη διαφορά ως προς την ακρίβεια, την ευαισθησία, την ειδικότητα και η GM, προκειμένου να επιλεγεί η πιο κατάλληλη τιμή των βαθμών ελευθερίας για τον πολυωνυμικό πυρήνα με την προϋπόθεση ότι έχουμε σταθερό το κόστος C ίσο με το 1 και το g με 0,0714 (Πίνακας 5.4).

Θα πρέπει να σημειωθεί ότι υπήρξε ένα υπολογιστικό εμπόδιο δεδομένου ότι είχαμε να λύσουμε ένα σύστημα εξισώσεων χρησιμοποιώντας έναν πίνακα για τον πυρήνα διάστασης nxn. Ως εκ τούτου παρέχουμε μόνο τα αποτελέσματα για τα οποία μπορούσε να υπάρξει λύση χωρίς να χρειάζεται να χρησιμοποιηθούν άλλες μεθόδους, προκειμένου να μειωθεί η διάσταση του πίνακα του πυρήνα. Με αυτό τον τρόπο επιλέχθησαν, στις περισσότερες περιπτώσεις, οι προεπιλεγμένες παράμετροι οι οποίες και δίνουν ικανοποιητικά αποτελέσματα. Για παράδειγμα, θέτουμε την offset παράμετρο, η οποία είναι απαραίτητη για πολυώνυμο πυρήνα, ίση με το μηδέν. Αλλάζοντας τους βαθμούς ελευθερίας και κρατώντας σταθερές τις παραμέτρους g και C παρατηρούμε ότι οι 2 βαθμοί ελευθερίας έδωσαν τα καλύτερα αποτελέσματα στο σύνολο δοκιμής σε σύγκριση με τους 1, 3 και 4. Επιπλέον, οι τιμές για τα g και το κόστος που αποκαλύπτουν την υψηλότερη απόδοση για όλα τα εξεταζόμενα μέτρα είναι 0.0714 και 1 αντίστοιχα. Σε

σχέση με την ακρίβεια και τον GM, οι 3 και 4 βαθμοί ελευθερίας έδωσαν την υψηλότερη τιμή, αλλά αυτό ήταν στο σύνολο εκπαίδευσης. Τέλος επιλέξαμε τις παραμέτρους που βελτιστοποιούν την ικανότητα γενίκευσης του ταξινομητή.

Πίνακας 5.4: Απόδοση του PSVM με πολυωνυμικό πυρήνα

g=0.0714 cost=1		$degree = 1$	$degree = 2$	$degree = 3$	$degree = 4$
Accuracy	Training set	0.9694	0.9734	0.9825	0.9825
	Testing set	**0.9689**	0.9695	0.9675	0.9675
Sensitivity	Training set	0.5123	0.5790	0.7151	0.7151
	Testing set	**0.5342**	0.5652	0.5229	0.5229
Specificity	Training set	0.9925	0.9934	0.9968	0.9968
	Testing set	**0.9939**	0.9928	0.9905	0.9905
GM	Training set	0.7131	0.7584	0.8369	0.8369
	Testing set	**0.7287**	0.7491	0.7197	0.7197

Χρησιμοποιώντας 10-fold cross validation, επικυρώσαμε τα παραπάνω αποτελέσματα μεταξύ 4 διαφορετικών τιμών των βαθμών ελευθερίας. Τα αποτελέσματα δεν ήταν τόσο καλά όσο εκείνα του Gaussian πυρήνα. Κατά συνέπεια, χρησιμοποιήσαμε Gaussian πυρήνα για τη μη γραμμική περίπτωση, έτσι ώστε να κάνουμε τις συγκρίσεις μεταξύ των διαφορετικών συνόλων δεδομένων και των μεθόδων που χρησιμοποιούνται στην ανάλυση μας.

5.1.3 Imbalanced Methods (Two cost/weight SVM and Modified Proximal SVM)

Στην περίπτωση των μη ισορροπημένων δεδομένων η απόδοση των τυποποιημένων αλγορίθμων μπορεί να βελτιωθεί με τη χρήση κατάλληλων "σταθμισμένων" μεθόδων. Για το τροποποιημένο PSVM, το MPSVM, υπάρχουν δύο παράμετροι που πρέπει να καθοριστούν. Το κόστος C και η παράμετρος δ. Η κατάλληλη επιλογή για την παράμετρο του κόστους, όπως έχουμε ήδη δει στο PSVM, είναι ίση με 1. Κατά συνέπεια, θα καθορίσουμε την παράμετρο δ. Επιπλέον, εφαρμόζοντας το TCSVM θα πρέπει να καθορίσουμε δύο κόστη, όπως μπορούμε να συμπεράνουμε από την προαναφερθείσα θεωρία που παρουσιάστηκε στε προηγούμενη ενότητα.

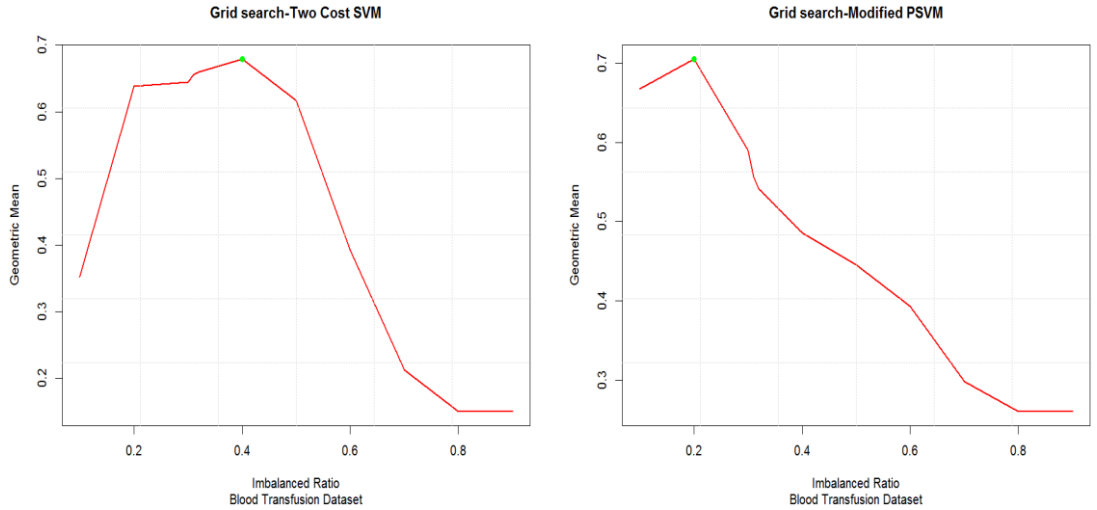
Για την επίτευξη των προσδοκώμενων αποτελεσμάτων της ταξινόμησης, τα δύο κόστη εσφαλμένης ταξινόμησης διαδραματίζουν σημαντικό ρόλο στην κατασκευή του

μοντέλου ευαίσθητης μάθησης. Καθορίσαμε λοιπόν, τις βέλτιστες παραμέτρους με βάση τη συνάρτηση αξιολόγησης του γεωμετρικού μέσου. Τα δύο κόστη είναι το κόστος της κλάσης μειοψηφίας (C^+) που αναφέρεται στις θετικές εμφανίσεις της μεταβλητής απόκρισης και το κόστος της κλάσης πλειοψηφία (C^-) που αναφέρεται στις αρνητικές περιπτώσεις. Μπορούμε να μειώσουμε τις επιπτώσεις της ανισορροπίας μεταξύ των κλάσεων θέτοντας ένα υψηλότερο κόστος ταξινόμησης για τα παραδείγματα της κλάσης μειοψηφίας από αυτές της κλάσης πλειοψηφίας. Οι Veropoulos et al. (1999) και Akbani et al. (2004) πρότειναν ως μια καλή επιλογή για αυτές τις παραμέτρους την αντίστροφη αναλογία μεταξύ των δύο μεγεθών των κλάσεων της μεταβλητής απόκρισης που βελτιώνει την απόδοση της μεθόδου TCSVM. Μετά την εκτέλεση αναζήτησης μεταξύ των διαφόρων τιμών για τα δύο κόστοι, επιβεβαιώνεται το προαναφερθέν αποτέλεσμα. Εμείς επιλέξαμε τις βέλτιστες παραμέτρους με βάση τη συνάρτηση αξιολόγησης του γεωμετρικού μέσου, δεδομένου ότι με την εφαρμογή αυτών των μεθόδων θέλουμε να ισορροπήσουμε την αναλογία ανάμεσα στην ευαισθησία και ειδικότητα.

Blood Transfusion

Το Blood Transfusion σύνολο δεδομένων αποτελείται από 748 παραδείγματα εκ των οποίων τα 178 ανήκουν στην κλάση 1 και δηλώνουν τους ανθρώπους που έδωσαν αίμα και τα 570 τους ανθρώπους που δεν έδωσαν αίμα.

Blood Transfusion	748	*Class 1* **178** / *Class -1* **570** (IR=0.31228)	1: donating blood -1: not donating blood

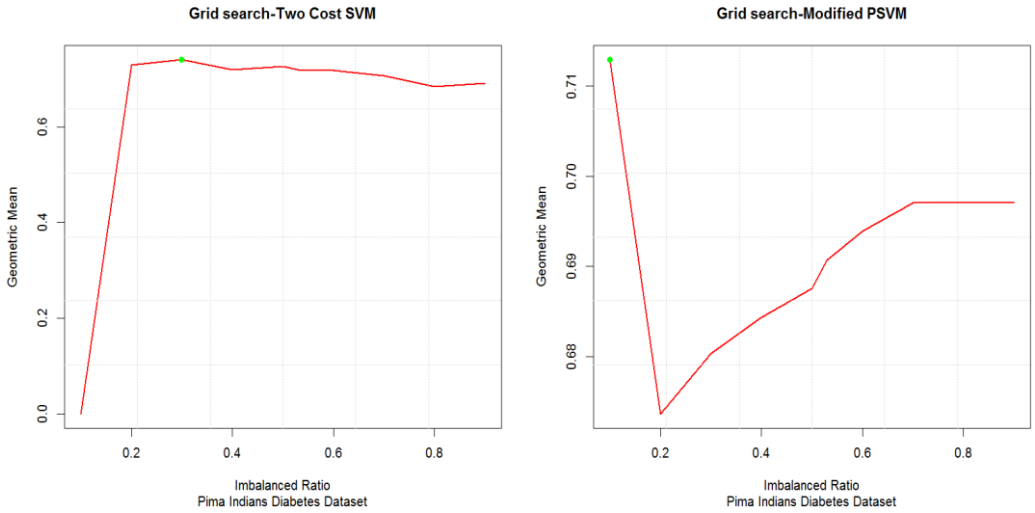


Σχήμα 5.10: Αναζήτηση πλέγματος **(**Grid search) για τις παραμέτρους των μεθόδων για μη ισορροπημένα δεδομένα χρησιμοποιώντας 10-fold cross validation. Το αριστερό πάνελ αναφέρεται στην ορθότητα του συνόλου δοκιμής θεωρώντας το μέτρο του Γεωμετρικού Μέσου (Geometric mean) δείχνοντας την απόδοση του TCSVM με γραμμικό πυρήνα. Το δεξί πάνελ αναφέρεται στο MPSVM με ένα γραμμικό πυρήνα θεωρώντας ως μέτρο τον Γεωμετρικό Μέσο. Οι μέθοδοι εφαρμόστηκαν στο Blood Transfusion σύνολο δεδομένων. Ο πράσινος συμπαγής κύκλος αντιπροσωπεύει το καλύτερο ποσοστό για τις παραμέτρους και των δυο μεθόδων.

Τα παραδείγματα της κλάσης μειοψηφίας αφορούν την κατηγορία των θετικών παραδειγμάτων κάτι που σημαίνει ότι το ποσοστό των ατόμων που έδωσαν αίμα ήταν χαμηλότερο από το ποσοστό εκείνων που δεν το έκανε. Η αναλογία μεταξύ της κλάσης πλειοψηφίας και της κλάσης μειοψηφίας είναι $n^-/n^+ = 0.31228$. Για να εκτελέσουμε την ανάλυση για το TCSVM ορίσαμε το κόστος της κλάσης μειοψηφίας ίσο με 1, δηλαδή είναι $C^- = 1$, και αλλάξαμε τις τιμές της κλάσης πλειοψηφίας. Οι Veropoulos et al (1999) και οι Akbani et al (2004) πρότειναν ότι το ποσοστό του κόστους των δύο κατηγοριών θα πρέπει να είναι κοντά στο αντίστροφο ποσοστό μεταξύ των δύο κλάσεων του δείγματος. Κατά συνέπεια, πραγματοποιήσαμε την ανάλυση με τιμές για την κλάση πλειοψηφίας κοντά στις προτάσεις των παραπάνω ερευνητών. Όπως αναμενόταν, το αντίστροφο ποσοστό μεταξύ των δύο κατηγοριών έδωσε αρκετά καλά αποτελέσματα όχι μόνο βάσει της ακρίβεια ταξινόμησης, αλλά και βάσει του μέτρου GM. Στο Σχήμα 5.10 απεικονίζεται η απόδοση τόσο της TCSVM (αριστερό πάνελ) όσο και του MPSVM (δεξιά) με τη χρήση του γραμμικού πυρήνα λαμβάνοντας υπόψη το μέτρο GM. Αυτή η μη ισορροπημένη αναλογία μεταξύ των δύο κατηγοριών μπορεί να αλλάξει μέσα από τις εγγενείς παραμέτρους της κάθε μεθόδου. Η επιλογή των τιμών κυμαινεται από 0,1 έως 0,9 και χρησιμοποιώντας ως μέθοδο επικύρωσης την 10-fold cross validation παρουσιάζουμε τα αποτελέσματα της GM στην Εικόνα 7. Η καλύτερη επιλογή της τιμής παρουσιάζεται με ένα πράσινο συμπαγή κύκλο.

Pima Indians Diabetes

Pima Indians Diabetes Dataset	768	*Class 1* **268** / *Class -1* **500** (Imbalanced Ratio=0.536)	1: positive for diabetes -1: negative for diabetes



Σχήμα 5.11: Αναζήτηση πλέγματος **(**Grid search) για τις παραμέτρους των μεθόδων για μη ισορροπημένα δεδομένα χρησιμοποιώντας 10-fold cross validation. Το αριστερό πάνελ αναφέρεται στην ορθότητα του συνόλου δοκιμής θεωρώντας το μέτρο του Γεωμετρικού Μέσου (Geometric mean) δείχνοντας την απόδοση του TCSVM με γραμμικό πυρήνα. Το δεξί πάνελ αναφέρεται στο MPSVM με ένα γραμμικό πυρήνα θεωρώντας ως μέτρο τον Γεωμετρικό Μέσο. Οι μέθοδοι εφαρμόστηκαν στο Pima Indians Diabetes σύνολο δεδομένων. Ο πράσινος συμπαγής κύκλος αντιπροσωπεύει το καλύτερο ποσοστό για τις παραμέτρους και των δυο μεθόδων.

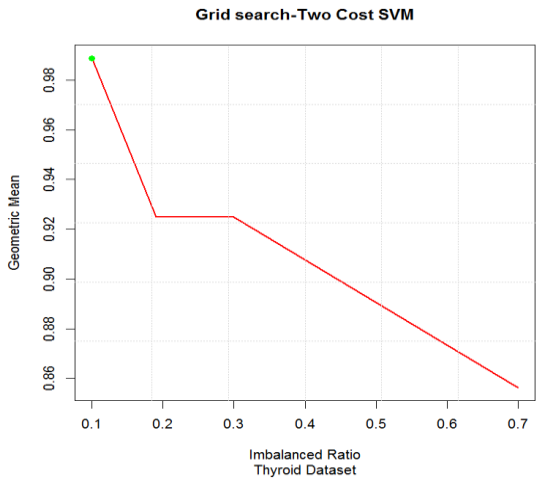
Το Pima Indians Diabetes σύνολο δεδομένων αποτελείται από 768 γυναίκες, 268 εκ των οποίων ανήκουν στην κατηγορία 1 η οποία δηλώνει γυναίκες που είναι θετικές στο διαβήτη και 500 γυναίκες στην κατηγορία -1 που είναι αρνητικές στο διαβήτη.

Για το Pima Indians Diabetes σύνολο δεδομένων η κλάση μειοψηφίας αποτελείται από άτομα θετικά στο διαβήτη και η κλάση πλειοψηφίας αποτελείται από άτομα αρνητικά στο διαβήτη. Η αναλογία μεταξύ της κλάσης πλειοψηφίας και της κλάσης μειοψηφίας είναι ίση με 0,536. Για να εκτελέσουμε την ανάλυση με τον αλγόριθμο TCSVM, καθορίσαμε το κόστος της κλάσης μειοψηφίας ίσο με 1, δηλαδή $C^- = 1$ και αλλάξαμε το ποσοστό αναλογίας για την πλειοψηφική κλάση. Πραγματοποιήσαμε την ανάλυση για τις τιμές του κόστους της κλάσης πλειοψηφίας από 0,1 έως 0,9. Όπως ήταν αναμενόμενο μια τιμή κοντά στο αντίστροφο ποσοστό μεταξύ των δύο κατηγοριών έδωσε τα καλύτερα αποτελέσματα για το μέτρο GM. Το Σχήμα 5.11 απεικονίζει την απόδοση και των δύο μεθόδων αλλάζοντας το κόστος της κλάσης πλειοψηφίας. Τα πιο ακριβή αποτελέσματα από την άποψη του GM δόθηκαν εξισώνοντας το κόστος για την κλάση πλειοψηφίας με την τιμή 0,3 για το TCSVM και την τιμή 0,1 για το MPSVM (Σχήμα 5.11).

Thyroid1

Το σύνολο δεδομένων Thyroid αποτελείται από 215 ασθενείς, 35 ανήκουν στην κατηγορία 1 και δηλώνουν τους ανθρώπους που είναι θετικοί στο θυρεοειδή (συγκεκριμένα στον υπερθυρεοειδισμό) και 180 ανήκουν στην κατηγορία -1 και δηλώνουν τους ανθρώπους οι οποίοι είναι αρνητικοί στο θυρεοειδή.

Thyroid Dataset	215	*Class 1* **35** */ Class -1* **180** (Imbalanced Ratio=0.1944)	1 positive for thyroid -1 negative for thyroid



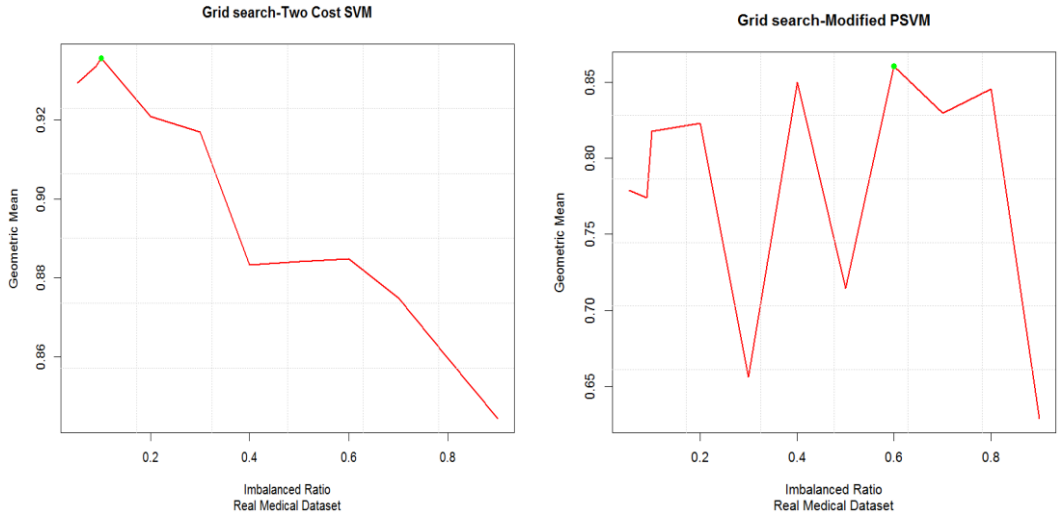
Σχήμα 5. 12: Αναζήτηση πλέγματος (Grid search) για τις παραμέτρους χρησιμοποιώντας 10-fold cross validation. Το γράφημα αναφέρεται στην ορθότητα στο σύνολο δοκιμής θεωρώντας το μέτρο του Γεωμετρικού Μέσου (Geometric mean) δείχνοντας την απόδοση του TCSVM με γραμμικό πυρήνα στο σύνολο δεδομένων Thyroid. Ο πράσινος συμπαγής κύκλος αντιπροσωπεύει το καλύτερο ποσοστό για την παράμετρο της μεθόδου TCSVM.

Για το σύνολο δεδομένων του θυρεοειδούς η κατηγορία μειοψηφίας αποτελείται από άτομα θετικά στον θυρεοειδή και η κατηγορία πλειοψηφίας αποτελείται από άτομα αρνητικά στο θυρεοειδή. Η αναλογία μεταξύ της κατηγορίας πλειοψηφίας και μειοψηφίας είναι ίση με 0.1944. Πραγματοποιήσαμε την ανάλυση για το TCSVM, όπως προηγουμένως, αλλάζοντας τις τιμές του κόστους της κλάσης πλειοψηφίας από 0,1 έως 0,9. Όπως ήταν αναμενόμενο μια τιμή κοντά στο αντίστροφο ποσοστό μεταξύ των δύο κατηγοριών έδωσε τα καλύτερα αποτελέσματα για το μέτρο GM. Έχουμε συγκεντρώσει επίσης την καλύτερη απόδοση λαμβάνοντας υπόψη και την ακρίβεια ταξινόμησης. Το Σχήμα 9 δείχνει την απόδοση του TCSVM και τα καλύτερα αποτελέσματα δόθηκαν χρησιμοποιώντας τιμή για το κόστος ίση με 0,1. Η καλύτερη επίδοση του MPSVM δόθηκε για το ίδιο ποσοστό, ίσο με 0.1.

Real Medical Data

Το Ιατρικό σύνολο δεδομένων αποτελείται από 8862 ασθενείς, 446 εκ των οποίων ανήκει στην κατηγορία 1 δηλώνοντας ανθρώπους που τελικά πέθαναν και 8416 ανθρώπους που επέζησαν.

Real medical Dataset	8862	*Class 1* **446**/ *Class -1* **8416** (Imbalanced Ratio=0.0529)	1 the death/ -1 survival



Σχήμα 5. 13: Αναζήτηση πλέγματος **(**Grid search) για τις παραμέτρους των μεθόδων για μη ισορροπημένα δεδομένα χρησιμοποιώντας 10-fold cross validation. Το αριστερό πάνελ αναφέρεται στην ορθότητα του συνόλου δοκιμής θεωρώντας το μέτρο του Γεωμετρικού Μέσου (Geometric mean) δείχνοντας την απόδοση του TCSVM με γραμμικό πυρήνα. Το δεξί πάνελ αναφέρεται στο MPSVM με ένα γραμμικό πυρήνα θεωρώντας ως μέτρο τον Γεωμετρικό Μέσο. Οι μέθοδοι εφαρμόστηκαν στο Real Medical σύνολο δεδομένων. Ο πράσινος συμπαγής κύκλος αντιπροσωπεύει το καλύτερο ποσοστό για τις παραμέτρους και των δυο μεθόδων.

Για το Ιατρικό σύνολο δεδομένων η κατηγορία μειοψηφίας αποτελείται από τα θετικά παραδείγματα και η κατηγορία πλειοψηφίας αποτελείται προφανώς από τα αρνητικά. Για τη μέθοδο TCSVM, δύο είναι παράμετροι του κόστους θα πρέπει να καθοριστούν, το κόστος της μειοψηφίας (C^+) που αναφέρεται στις θετικές εμφανίσεις και το κόστος πλειοψηφία (C^-) που αναφέρεται στις αρνητικές. Η αναλογία μεταξύ της κατηγορίας πλειοψηφίας και της κατηγορίας μειοψηφίας είναι ίση με 0,05299. Πιο συγκεκριμένα, με τον καθορισμό του κόστους της κλάσης των μειονοτήτων ίση με 1 και αλλάζοντας το κόστος της κλάσης πλειοψηφίας πραγματοποιήσαμε μια έρευνα μεταξύ πολλών διαφορετικών τιμών. Έχουμε εκτελέσει την ανάλυση για τις τιμές μεταξύ 0,0529 έως 0,9. Τα πιο ακριβή αποτελέσματα για μέτρο GM και την ακρίβεια ταξινόμησης δόθηκαν για την τιμή 0.1. Ωστόσο, τα αποτελέσματα που δόθηκαν για την τιμή 0,0529 (= λόγος) ήταν πραγματικά κοντά στην απόδοση που πραγματοποιείται με την τιμή 0,1. Όσο αφορά στο MPSVM, μεταξύ πολλών διαφορετικών τιμών για το δ, τα καλύτερα αποτελέσματα για το μέτρο GM λαμβάνονται για τιμές κοντά στο αντίστροφο ποσοστό μεταξύ των μεγεθών του δείγματος. Τελικά τα καλύτερα αποτελέσματα για το GM δόθηκαν για δ ίσο με 0,6. Στο Σχήμα 5.13 παρατηρούμε τις επιπτώσεις της αλλαγής του δ στο GM.

Πίνακας 5.5: Ορθότητα ταξινόμησης στο σύνολο εκπαίδευσης και δοκιμής και χρόνοι εκτέλεσης χρησιμοποιώντας μεθόδους για μη ισορροπημένα δεδομένα, γραμμικό TCSVM, μη γραμμικό TCSVM και MPSVM. Τα καλύτερα αποτελέσματα φαίνονται με έντονα γράμματα.

Σύνολο Δεδομένων	Μοντελοποίηση εξισορρόπησης	Ακρίβεια	Ευαισθησία	Ειδικότητα	GM	Χρόνος
Blood Transfusion	Two cost SVM (Train,Test)	(0.685,0.667)	(0.761 ,0.705)	(0.661,0.655)	(**0.7094**, 0.6792)	**0.33**
	Nonlinear TCSVM (Train,Test)	(**0.772**, **0.769**)	(0.559,0.546)	(0.839, 0.838)	(0.6852 ,0.676)	0.72
	MPSVM (Train,Test)	(0.714,0.747)	(0.634 , 0.636)	(0.738,0.782)	(0.684, **0.705**)	0.68
Pima Indians Diabetes	TCSVM (Train,Test)	(0.719,0.719)	(0.876,0.836)	(0.635,0.656)	(0.746, **0.740**)	**0.39**
	Nonlinear TCSVM (Train,Test)	(0.712,0.708)	(0.950,0.881)	(0.584,0.616)	(0.745, 0.737)	0.79
	MPSVM (Train,Test)	(**0.776 , 0.739**)	(0.706, 0.641)	(0.813,0.792)	(**0.758**,0.713)	0.81
Thyroid1	TCSVM (Train,Test)	(**0.994,0.981**)	(1.00,1.00)	(0.993,0.978)	(**0.996**,**0.989**)	0.14
	Non linear TCSVM (Train,Test)	(0.969, 0.962)	(1.00, 1.00)	(0.963, 0.956)	(0.981, 0.978)	0.14
	MPSVM (Train,Test)	(0.944, 0.925)	(0.667, 0.500)	(1.00, 1.00)	(0.8165,0.707)	**0.11**
Medical Data	TCSVM (Train,Test)	(0.962, 0.969)	(0.878,0.892)	(0.967 ,0.973)	(0.921, 0.9317)	**12.17**
	Non linear TCSVM (Train,Test)	(0.968, 0.968)	(0.919, 0.901)	(0.971, 0.972)	(**0.945, 0.936**)	15.94
	MPSVM (Train,Test)	(**0.971, 0.979**)	(0.707, 0.748)	(0.985, 0.991)	(0.835, 0.861)	939.36

Σε αυτό το πείραμα συγκρίνουμε τρεις μεθόδους ευαίσθητου κόστους, δύο από αυτές αποτελούν παραλλαγές του SVM και μία τροποποίηση του PSVM. Από την άποψη της ακρίβειας ταξινόμησης, όλες οι μέθοδοι παρέχουν σχεδόν παρόμοια αποτελέσματα με το MPSVM να ξεπερνά τις υπόλοιπες στις περισσότερες περιπτώσεις. Ωστόσο, λαμβάνοντας υπόψη το μέτρο GM, το γραμμικό και το μη γραμμικό TCSVM, είχε καλύτερα αποτελέσματα από το MPSVM. Λαμβάνοντας υπόψη το χρόνο εκτέλεσης των μεθόδων θα πρέπει να σημειώσουμε ότι το TCSVM με γραμμικό πυρήνα εμφανίζεται ως ο ταχύτερος αλγόριθμος όλων. Από τα τρία μικρά σύνολα δεδομένων συμπεραίνουμε ότι οι μη ισορροπημένες μέθοδοι που χρησιμοποιήθηκαν στο πείραμα είχαν σχεδόν ίδιο χρόνο εκτέλεσης. Ωστόσο, όσον αφορά το μεγαλύτερο σύνολο δεδομένων που είναι το ιατρικό, η μέθοδος MPSVM φαίνεται να μην αποτελεί και τόσο καλή επιλογή, σε σύγκριση με τις άλλες μεθόδους. Βλέπε Πίνακα 5.5 για περισσότερες λεπτομέρειες.

5.1.4 Comparisons

Έχοντας επιλέξει τις καλύτερες τιμές μετά την επιλογή των παραμέτρων στις προαναφερθείσες μεθόδους, παρουσιάζουμε μια σύγκριση μεταξύ του SVM και PSVM στην περίπτωση του γραμμικού πυρήνα. Ο Πίνακας 5.6 παρουσιάζει την απόδοση ως προς την ακρίβεια ταξινόμησης.

Πίνακας 5.6: Ορθότητα ταξινόμησης του PSVM και του SVM στο σύνολο εκπαίδευσης και στο σύνολο δοκιμής χρησιμοποιώντας 10-fold cross validation καθώς επίσης και οι χρόνοι εκτέλεσης χρησιμοποιώντας γραμμικό πυρήνα. Τα καλύτερα αποτελέσματα είναι με έντονα γράμματα.

linear Formulation **ACCURACY** / Data Set	SVM (Train,Test) Sec	PSVM (Train,Test) Sec
Pima Indians Data Set	**(0.7865,0.7812)** 0.36	**(0.7865,** 0.7604) **0.11**
Thyroid 1	**(1.00, 0.9434)** 0.10	(0.9444, 0.9245) **0.08**
Blood Transfusion	(0.7687, 0.7634) 0.29	**(0.774, 0.7688)** **0.14**
Real Medical Data Set	**(0.975, 0.9756)** 10.21	(0.9695, 0.9702) **5.60**

Από τα αποτελέσματα που δίνονται στον Πίνακα 5.6, η γραμμική SVM έχει καλύτερη ακρίβεια ταξινόμησης στο σύνολο των δοκιμών εκτός από το σύνολο Blood Transfusion. Το PSVM έχει περίπου την ίδια απόδοση με το SVM, αλλά ο χρόνος για την εκπαίδευση του μοντέλου είναι πολύ μικρότερος σε αυτή την περίπτωση.

Πιο συγκεκριμένα, για το Pima Indians σύνολο δεδομένων φαίνεται ότι στην εκπαίδευση και οι δύο μέθοδοι έχουν την ίδια ακρίβεια με το SVM να υπερτερεί ελαφρώς του PSVM στο σύνολο ελέγχου. Όσον αφορά το σύνολο δεδομένων του Thyroid, καταλήγουμε στο συμπέρασμα ότι το χαμηλότερο σφάλμα ταξινόμησης προκύπτει στην περίπτωση του SVM. Ωστόσο το SVM ήταν βραδύτερο από το PSVM. Από την ανάλυση του συνόλου Blood Transfusion, το PSVM είχε την καλύτερη ακρίβεια και στο σύνολο εκπαίδευσης και στο σύνολο δοκιμής αλλά και στο χρόνο. Στο ιατρικό σύνολο δεδομένων, οι δύο μέθοδοι έδωσαν καλά αποτελέσματα επιτυγχάνοντας τα ποσοστό των 0,9756 και 0,9702 για το SVM και το PSVM αντίστοιχα. Ωστόσο, το PSVM απαιτεί το μισό χρόνο για να εκπαιδευτεί σε σύγκριση με το SVM.

Ο Πίνακας 5.7 δείχνει την απόδοση του SVM και του PSVM για τη μη γραμμική περίπτωση για τα τέσσερα διαφορετικά σύνολα δεδομένων. Σε αντίθεση με γραμμική περίπτωση, το SVM φαίνεται να εκτελείται ταχύτερα από ότι το PSVM εκτός από το σύνολο δεδομένων Thyroid. Ωστόσο, το PSVM δίνει καλύτερα ποσοστά ταξινόμησης στο σύνολο των δεδομένων εκπαίδευσης επιτυγχάνοντας 100% ακρίβεια στα σύνολα δεδομένων Pima Indians Diabetes και Thyroid. Ωστόσο στα δεδομένα δοκιμών, το SVM έχει χαμηλότερα ποσοστά ταξινόμησης από το PSVM. Σχετικά με το πραγματικό Ιατρικό σύνολο δεδομένων το οποίο αποτελεί ένα μεγάλης κλίμακας στατιστικό πρόβλημα, το PSVM εμφανίζει μια εξαιρετική απόδοση συγκεντρώνοντας το ποσοστό 98.48% για το σύνολο εκπαίδευση και 97,7% για το σύνολο ελέγχου. Ωστόσο, χρειάζεται περισσότερο χρόνο για την εκπαίδευση σε σύγκριση με το πρότυπο SVM.

Πίνακας 5.7: Ορθότητα ταξινόμησης του PSVM και του SVM στο σύνολο εκπαίδευσης και στο σύνολο δοκιμής χρησιμοποιώντας 10-fold cross validation καθώς επίσης και οι χρόνοι εκτέλεσης χρησιμοποιώντας μη γραμμικό πυρήνα. Τα καλύτερα αποτελέσματα είναι με έντονα γράμματα.

Formulation **Radial kernel Accuracy** / Data Set	SVM (Train,Test) Sec	PSVM (Train, Test) Sec
Pima Indians Diabetes	0.795, **0.776** **0.74**	**1.00,** 0.656 2.84
Thyroid 1	1.00, **0.943** 0.13	**1.00,** 0.868 **0.09**
Blood Transfusion	0.819, **0.790** **0.55**	**0.927,** 0.753 1.94
Real Medical Data Set	0.984, **0.977** **7.99**	**0.989, 0. 977** **623.02**

Συγκρίνοντας τη γραμμική με τη μη γραμμική περίπτωση θα πρέπει να σημειωθεί ότι:

- Στο *Pima Indians Diabetes* σύνολο δεδομένων το PSVM με μη γραμμικό πυρήνα έδωσε ποσοστό ακρίβειας 100% επί του συνόλου εκπαίδευσης

επιτυγχάνοντας την υψηλότερη τιμή μεταξύ όλων των άλλων μεθόδων. Ενώ στο σύνολο ελέγχου το SVM με το γραμμικό (**linear SVM**) πυρήνα έδωσε τα πιο ακριβή αποτελέσματα. Κατά συνέπεια, δεδομένου ότι υπάρχει μια ισορροπία ανάμεσα στην ακρίβεια στο σύνολο εκπαίδευσης και στο σύνολο ελέγχου στην περίπτωση του γραμμικού SVM φαίνεται να είναι η καλύτερη επιλογή για το παρών σύνολο δεδομένων. Σαφώς, ο χρόνος για να εκπαιδεύσει το μοντέλο είναι πολύ μικρότερος στην περίπτωση του γραμμικού PSVM.

- Ομοίως, όσον αφορά το σύνολο δεδομένων *Thyroid*, το γραμμικό SVM (**linear SVM**) είναι η καλύτερη επιλογή θεωρώντας ως μέτρο αξιολόγησης την ακρίβεια ταξινόμησης. Παρ'όλα αυτά, το γραμμικό PSVM υπερτερεί στο χρόνο εκπαίδευσης.
- Στην περίπτωση του περίπτωση *Blood Transfusion* συνόλου δεδομένων φαίνεται να έχει καλύτερες επιδόσεις τόσο το SVM όσο και το PSVM. Στο σύνολο εκπαίδευσης το PSVM επιτυγχάνει ακρίβεια 92,7%, ακρίβεια υψηλότερη από εκείνη του SVM που είναι ίση με 81,85%. Στο σύνολο ελέγχου, το SVM έχει ακρίβειας ίση με 79.03% ένα ποσοστό που είναι πραγματικά κοντά στο 75,27% του PSVM. Ως εκ τούτου, το μη γραμμικό PSVM (**nonlinear PSVM**) φαίνεται να εφαρμόζεται καλύτερα στο παρόν σύνολο δεδομένων.
- Τέλος, το μη γραμμικό PSVM (**nonlinear PSVM**) έχει σαφώς την καλύτερη απόδοση στο πραγματικό ιατρικό σύνολο (*real medical*) επιτυγχάνοντας ακρίβεια 98.48% και 97,7% για την εκπαίδευση και την εξέταση αντίστοιχα.

Για πολλές εφαρμογές, ειδικά για την ιατρική διάγνωση, είναι πολύ σημαντικό να γίνει διάκριση ανάμεσα στην ακρίβεια των ψευδώς αρνητικών και των ψευδώς θετικών αποτελεσμάτων. Ο σκοπός της παρούσας μελέτης είναι αφενός να συγκριθούν οι διάφορες τροποποιήσεις του αλγορίθμου SVM σε ιατρικά δεδομένα και αφετέρου να αξιολογηθεί επιτυχώς η απόδοση των ταξινομητών, διατηρώντας την σωστή ισορροπία μεταξύ ευαισθησίας και εξειδίκευσης, προκειμένου να καταστεί δυνατή η επιτυχία της πρόβλεψης. Έχουν προταθεί πολλές στρατηγικές που ασχολούνται με μη ισορροπημένα δεδομένα, μερικές από τις οποίες έχουν εφαρμοστεί στην παρούσα ανάλυση και παρουσιάζονται στο θεωρητικό τμήμα. Για την επίτευξη των προσδοκώμενων αποτελεσμάτων ταξινόμησης, οι δαπάνες εσφαλμένης ταξινόμησης διαδραματίζουν καίριο ρόλο στην κατασκευή του μοντέλου ευαίσθητου κόστους. Κατά συνέπεια, μετά την επιλογή των παραμέτρων συγκρίνουμε αυτές τις μεθοδολογίες, το TCSVM και το MPSVM, χρησιμοποιώντας τις τιμές που έδωσαν την υψηλότερη απόδοση.

Θεωρώντας και πάλι ως μέτρο αξιολόγησης την ακρίβεια ταξινόμησης, το MPSVM παρέχει ικανοποιητικά αποτελέσματα σε όλα τα σύνολα δεδομένων που χρησιμοποιήθηκαν στη μελέτη μας δίνοντας την καλύτερη απόδοση σε σχέση με τις άλλες μεθόδους για τα σύνολα δεδομένων Pima *Indians Diabetes* και για το *Real Medical*. Θα πρέπει να σημειωθεί ότι αυτά τα δύο σύνολα είναι και τα μεγαλύτερα από τα τέσσαρα που λαμβάνουν χώρα στη συγκεκριμένη μελέτη. Στο σύνολο *Thyroid* and *Blood Transfusion* το **γραμμικό TCSVM** και το **μη γραμμικό TCSVM** είχαν τις υψηλότερες τιμές

ακρίβειας. Ωστόσο το MPSVM είναι περισσότερο χρονοβόρο από τη μέθοδο TCSVM (Πίνακας 5.8).

Πίνακας 5.8: Ορθότητα ταξινόμησης του TCSVM, του μη γραμμικού TCSVM και του MPSVM στο σύνολο εκπαίδευσης και στο σύνολο δοκιμής χρησιμοποιώντας 10-fold cross validation καθώς επίσης και οι χρόνοι εκτέλεσης. Τα καλύτερα αποτελέσματα είναι με έντονα γράμματα.

Formulation **Accuracy** / Data Set	TCSVM (Train, Test) Sec	Nonlinear TCSVM (Train,Test) Sec	Modified PSVM (Train,Test) Sec
Blood Transfusion	(0.685, 0.667) **0.33**	**(0.772,0.769)** 0.72	(0.714,0.747) 0.68
Pima Indians Diabetes	(0.719, 0.719) **0.39**	(0.712,0.708) 0.79	**(0.776,0.739)** 0.81
Thyroid 1	**(0.994,0.981)** 0.14	(0.969,0.962) 0.14	(0.944,0.925) **0.11**
Real Medical Data Set	(0.962,0.969) **12.17**	0.968,0.968 15.94	**0.971,0.979** 939.36

Οι τιμές για το γεωμετρικό μέσο τόσο στο σύνολο εκπαίδευσης όσο και στο σύνολο δοκιμής χρησιμοποιώντας τις τρεις διαφορετικές μεθόδους παρουσιάζεται στον Πίνακα 5.9. Η σύγκριση των διαφορών για όλα τα ζεύγη των μεθόδων που δείχνει ότι στο σύνολο *Thyroid* και στο *Real Medical*, το γραμμικό αλλά και το μη γραμμικό TCSVM έχει τις υψηλότερες τιμές του GM. Από την άλλη πλευρά, το MPSVM φαίνεται να έχει καλύτερες επιδόσεις στο σύνολο ελέγχου για τα δεδομένα *Blood Transfusion* και στο σύνολο εκπαίδευσης του *Pima Indians Diabetes*. Συμπερασματικά, το TCSVM πιστεύεται ότι είναι προτιμότερο από MPSVM όταν ο στόχος είναι να εξισορροπηθεί το ποσοστό μεταξύ ευαισθησίας και εξειδίκευσης.

Πίνακας 5.9: Ορθότητα ταξινόμησης του TCSVM, του μη γραμμικού TCSVM και του MPSVM στο σύνολο εκπαίδευσης και στο σύνολο δοκιμής θεωρώντας το μέτρο του Γεωμετρικού Μέσου χρησιμοποιώντας 10-fold cross validation. Περιλαμβάνονται και οι χρόνοι εκτέλεσης. Τα καλύτερα αποτελέσματα είναι με έντονα γράμματα.

Formulation **Geometric Mean** / Data Set	TCSVM Train,Test	Nonlinear TCSVM Train,Test	Modified PSVM Train,Test
Blood Transfusion	**(0.709,**0.679)	(0.685,0.676	0.684,**0.705)**
Pima Indians Diabetes	(0.745,**0.740)**	(0.745,0.737)	**(0.758,**0.713)
Thyroid 1	**(0.996,0.988)**	(0.981,0.978)	(0.817,0.707)
Real Medical Data Set	(0.921,0.931)	**(0.945,0.936)**	(0.835,0.861)

Επιπλέον, όπως προκύπτει από τους πίνακες που παρουσιάστηκαν παραπάνω (Πίνακας 5.6, 5.7, 5.8, 5.10) το SVM και το PSVM συγκέντρωσαν υψηλότερες τιμές ακρίβειας στις περισσότερες περιπτώσεις, σε σύγκριση με τις μεθόδους μάθησης ευαίσθητου κόστους. Ωστόσο, από την άποψη του γεωμετρικού μέσου (GM), το TCSVM και το MPSVM έχουν σαφώς καλύτερες επιδόσεις σε σχέση με τις συμβατικές μεθόδους οι οποίες δεν λαμβάνουν υπόψη τη μη ισορροπημένη αναλογία μεταξύ των δύο κλάσεων.

Συγκρίνοντας τις δύο *γραμμικές μεθόδους*, το γραμμικό PSVM και MPSVM, η δεύτερη υπερτερεί της πρώτης στο σύνολο εκπαίδευσης, καθώς και στο σύνολο ελέγχου. Όπως ήταν αναμενόμενο, η μέθοδος MPSVM είναι πιο κατάλληλη για να χρησιμοποιηθεί στην περίπτωση που θέλουμε μια ισορροπία ανάμεσα στην ευαισθησία και στην ειδικότητα και κατά συνέπεια μια καλή τιμή για το GM. Ειδικά, σε μεγαλύτερα σύνολα δεδομένων, όπως η Ιατρική βάση και το Pima Indians Diabetes σύνολο δεδομένων, ο αλγόριθμος MPSVM παρουσιάζει υψηλά ποσοστά ακρίβειας.

5.2 Εφαρμογή σε πραγματικά ιατρικά δεδομένα υψηλής διάστασης

Σε αυτή την ενότητα συγκρίνουμε την απόδοση των διαφορετικών μεθόδων που παρουσιάσαμε προηγουμένως σ' ένα πραγματικό σύνολο ιατρικών δεδομένων. Πιο συγκεκριμένα εφαρμόσαμε τις μεθόδους SVM και TCSVM, μεθόδους τυχαίας δειγματοληψίας (oversampling και undersampling) καθώς επίσης και ένα συνδυασμό του SMOTE και της τυχαίας undersampling δειγματοληψίας σε ένα σύνολων ιατρικών δεδομένων υψηλής διάστασης. Κύριος στόχος μας είναι να παρέχουμε μια αμερόληπτη εκτίμηση της απόδοσης του κάθε μοντέλου. Για το λόγο αυτό οι τιμές των κριτηρίων απόδοσης υπολογίζονται σ' ένα σύνολο δεδομένων το οποίο δεν χρησιμοποιείται στη διαδικασία οικοδόμησης του εκάστοτε μοντέλου. Το σύνολο αυτό, αποτελεί ένα τμήμα του αρχικού συνόλου δεδομένων και ονομάζεται σύνολο δοκιμής (test set). Παράλληλα, ένας ταξινομητής πρέπει να παρουσιάζει υψηλές τιμές της ακρίβειας, της ευαισθησίας, της εξειδίκευση, της AUROC και του γεωμετρικού μέσου. Η απόδοση γενίκευσης του μοντέλου συχνά εκτιμάται, όπως και στη συγκεκριμένη μελέτη από την holdout επικύρωση (holdout validation).

Στη παρούσα μελέτη για τη ασφαλή διεξαγωγή συμπερασμάτων, χωρίζουμε με τυχαίο τρόπο (απλή τυχαία δειγματοληψία) το σύνολο δεδομένων, που αποτελείται από σε ένα σύνολο εκπαίδευσης, που περιέχει 75% των περιπτώσεων και το σύνολο ελέγχου, που περιέχει 25% των περιπτώσεων, προκειμένου να αξιολογηθεί η απόδοση των ταξινομητών σε νέα δεδομένων.

Το ιατρικό σύνολο δεδομένων, του οποίου μία περιγραφή δίνεται στη συνέχεια, περιλαμβάνει μία μεταβλητή απόκρισης η οποία αποτελείται από δύο κλάσεις οι οποίες είναι μη ισορροπημένες (446 θετικές περιπτώσεις και 8416 αρνητικές περιπτώσεις). Αυτό καθιστά επιτακτική τόσο τη χρήση των μεθόδων «προ-επεξεργασίας» που συμβάλουν στην εξισορρόπηση των κλάσεων όσο και τις μεθόδους ευαίσθητης μάθησης (cost sensitive learning methods) που δίνουν διαφορετικά βάρη στις δύο διαφορετικές κλάσεις των δεδομένων. Επιπλέον, η χρήση πιο εύρωστων μέτρων (robust measures) από την ακρίβεια, όπως το εμβαδό κάτω από την ROC καμπύλη και το γεωμετρικό μέσο, παρέχει αδιαμφισβήτητα πιο αξιόπιστα συμπεράσματα. Το κίνητρό μας για την διεξαγωγή αυτής της μελέτης προέρχεται από την υποστήριξη ιατρικών αποφάσεων κάνοντας την επιλογή ενός συνόλου ιατρικών δεδομένων επιτακτική.

Η ανάλυση, η οποία περιλαμβάνει όλα τα στάδια προ-επεξεργασίας των δεδομένων και τα μοντέλα που αναπτύχθηκαν, πραγματοποιήθηκε χρησιμοποιώντας κώδικες της R και οι αλγόριθμοι υλοποιήθηκαν χρησιμοποιώντας ταυτόχρονα τα πακέτα «e1071» και «DMwR ».

5.2.1 Περιγραφή των δεδομένων

Το υπό εξέταση σύνολο δεδομένων αποτελείται από τη μεταβλητή απόκρισης y η οποία αναφέρεται στην επιβίωση ή μη του ασθενούς και κωδικοποιείται με -1 / 1 (ή 0/1) αντίστοιχα για κάθε κατηγορία. Αναλυτικότερα, για κάθε ασθενή το χαρακτηριστικό προορισμού, y, είναι μία δυαδική μεταβλητή και υποδηλώνει την πιθανότητα του θανάτου. Εκφράζεται υπό τη μορφή δύο κατηγοριών -1 και 1, όπου το -1 αντιπροσωπεύει την επιβίωση, ενώ η τιμή 1 το θάνατο. Οι δύο κλάσεις της μεταβλητής απόκρισης, και ως εκ τούτου το σύνολο δεδομένων, είναι μη ισορροπημένες. Αυτό γίνεται φανερό από τα στοιχεία που παρουσιάζονται στον παρακάτω πίνακα όπου παρατηρούμε ότι το σύνολο αποτελείται από 446 θετικές περιπτώσεις (θετική - majority class) και 8416 αρνητικές περιπτώσεις (αρνητική - minority class).

Πίνακας 5.11: Περιγραφή του ιατρικού συνόλου δεδομένων

Σύνολο δεδομένων	Αριθμός πειραματικών εκτελέσεων	Αριθμός μεταβλητών
Medical data set	8862 (θετικά (death):446, αρνητικά (survive): 8416)	41 (14 συνεχείς, 27:διακριτές)

Το σύνολο δεδομένων αποτελείται από $N = 8862$ ασθενείς και 41 επεξηγηματικές μεταβλητές που περιλαμβάνουν δημογραφικά στοιχεία, δεδομένα μεταφορών προς και από το νοσοκομείο αλλά και ενδονοσοκομειακά δεδομένα. Σύμφωνα με ιατρικές συμβουλές, όλοι οι προγνωστικοί παράγοντες θα πρέπει να τυγχάνουν ίσης μεταχείρισης κατά την στατιστική ανάλυση και δεν υπάρχει κανένα στοιχείο που θα πρέπει να διατηρείται πάντα στο μοντέλο. Τα ονόματα αυτών των παραγόντων παρατίθενται στο Παράρτημα Α.

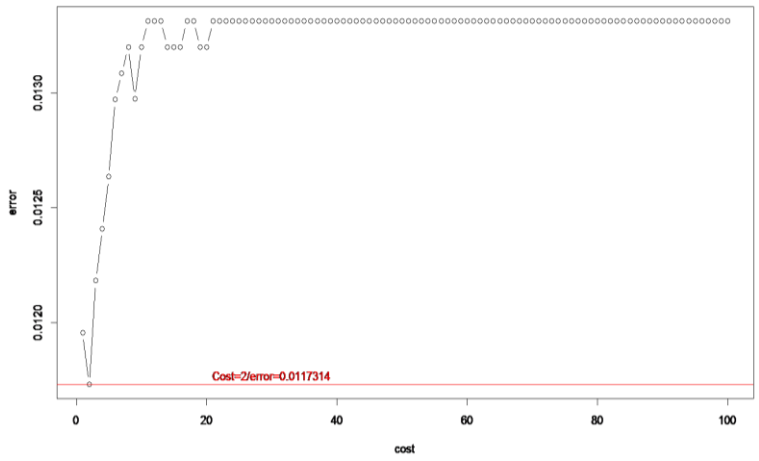
Για την αξιολόγηση της απόδοσης των ταξινομητών διασπάμε το σύνολο δεδομένων σ' ένα σύνολο εκπαίδευσης, που περιέχει 75% των περιπτώσεων (6647) και σε ένα σύνολο ελέγχου, που περιέχει 25% των περιπτώσεων (2215). Έτσι μπορούμε να αξιολογήσουμε την απόδοση των ταξινομητών σε νέα δεδομένα και να έχουμε μία πιο ασφαλή πρόβλεψη.

5.2.2 Κλασικό SVM (Standard SVM)

Για τον κλασικό SVM ταξινομητή αυτό που πρέπει να καθορίσουμε είναι τόσο η συνάρτηση πυρήνα όσο και η παράμετρος κανονικοποίησης C. Παράλληλα, ανάλογα με την επιλογή του πυρήνα (kernel) έχουμε να επιλέξουμε και τις αντίστοιχες παραμέτρους. Για παράδειγμα στην περίπτωση του Κανονικού πυρήνα (Gaussian / RBF) η παράμετρος

που χρειάζεται να καθοριστεί είναι η gamma και στην περίπτωση του πολυωνυμικού πυρήνα (Polynomial) είναι οι βαθμοί ελευθερίας. Το ζήτημα της επιλογής μεταβλητών στις μηχανές διανυσματικής υποστήριξης είναι ιδιαίτερα σημαντική και επηρεάζει σημαντικά την συνολική επίδοση των ταξινομητών κάνοντας το SVM ιδιαίτερα ευαίσθητο στην επιλογή αυτών των παραμέτρων. Εφαρμόζοντας 10-fold διασταυρωμένη επικύρωση (cross validation) εξασφαλίζουμε την τιμή του κόστους για την καλύτερη επίδοση με βάση το ποσοστό σφάλματος, η οποία ισούται με 2.

Στο Σχήμα 5.14 παρουσιάζεται η μεταβολή στο σφάλμα ταξινόμησης για τις διαφορετικές τιμές της παραμέτρου του κόστους C, στην περίπτωση του κλασικού γραμμικού SVM ταξινομητή. Πέραν του κόστους, οι εσωτερικές παράμετροι του SVM επιδρούν σημαντικά στην επίδοση, όπως έχουμε ήδη αναφέρει προγουμένως.



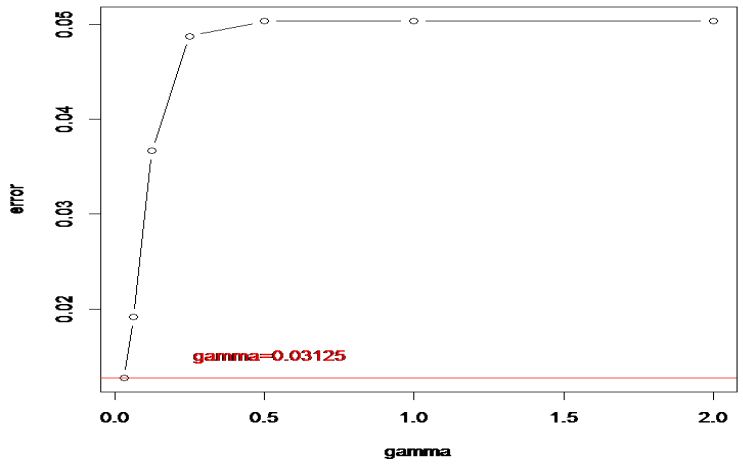
Σχήμα 5.14: Απόδοση του γραμμικού SVM για διαφορετικές τιμές του κόστους βάσει του ποσοστού σφάλματος. Η κόκκινη γραμμή δείχνει το κόστος που παρουσιάζει την καλύτερη επίδοση.

Για τον κανονικό πυρήνα, πέραν της παραμέτρου C, πρέπει να γίνει και επιλογή της παραμέτρου gamma ανάμεσα από πολλές υποψήφιες τιμές.

Πίνακας 5.12: επιλογή παραμέτρου στο SVM με Γκαουσιανό πυρήνα (RBF).

gamma	Error	dispersion
0.03125	0.01275349	0.003503342
0.06250	0.01918199	0.005812675
0.12500	0.03666481	0.008780307
0.25000	0.04874225	0.007506852
0.50000	0.05032355	0.007295674
1.00000	0.05032355	0.007295674
2.00000	0.05032355	0.007295674

Η τιμή αυτή θα πρέπει κανονικά να είναι μεταξύ 1/k (=0.0244) και 6/k (=0.14634), όπου το k αντιπροσωπεύει τη διάσταση των δεδομένων (41 στη μελέτη μας). Εκτελώντας μία εκτενή αναζήτηση (grid search) επιλέξαμε εκείνη την τιμή που καταλήγει στην καλύτερη επίδοση. Το Σχήμα 5.15 παρουσιάζει τη διαφορά που παρουσιάζεται στο σφάλμα, αλλάζοντας την παράμετρο gamma. Η βέλτιστη τιμή της gamma (=0.03125) που φαίνεται στο Σχήμα 5.15 (κόκκινη γραμμή), δίνει το μικρότερο ποσοστό σφάλματος. Στον Πίνακας 5.12 παραθέτουμε κάποιες ενδεικτικές τιμές.



Σχήμα 5.15: Επίδοση του SVM με κανονικό (RBF) πυρήνα για διαφορετικές τιμές της παραμέτρου gamma

Τα εκτιμώμενα μέτρα στον Πίνακας 5.1313 λαμβάνονται χρησιμοποιώντας την τιμή 2 για την παράμετρο C ($C = 2$) για τον γραμικό πυρήνα, $C = 1$ για τον σιγμοειδή, τον πολυωνυμικό και τον κανονικό πυρήνα. Η παράμετρος στην περίπτωση του κανονικού πυρήνα επιλέχθηκε, όπως προηγούμενως παρουσιάσαμε, ίση με $gamma = 0.03125$. Στην περίπτωση του πολυωνυμικού ή του σιγμοειδούς πυρήνα η offset παράμετρος τίθεται ίση με μηδέν το οποίο αποτελεί και default επιλογή στις περισσότερες περιπτώσεις. Τέλος μόνο στην περίπτωση του πολυωνυμικού πυρήνα υπάρχει η παράμετρος που σχετίζεται με τους βαθμούς ελευθερίας και στην παρούσα μελέτη την θέσαμε ίση με 3.

Ο Πίνακας 5.133 δείχνει την επίδοση του SVM χρησιμοποιώντας διαφορετικούς πυρήνες. Τόσο το γραμμικό SVM όσο και το SVM με κανονικό πυρήνα παρουσιάζουν την υψηλότερη απόδοση σε ακρίβεια, ευαισθησία, ειδικότητα, AUC και στο γεωμετρικό μέσο (GM).

Ο κανονικός πυρήνας φτάνει το ποσοστό του 0.9848, 0.77922, 0.99821, 0.8866 και 0.8796 για την ακρίβεια, την ευαισθησία, την ειδικότητα, την AUC και το γεωμετρικό μέσο (GM) αντίστοιχα. Σχεδόν παρόμοια αποτελέσματα δίνει και ο γραμμικός πυρήνας. Ο δεύτερος καλύτερος προκύπτει να είναι ο σιγμοειδής πυρήνας όσον αφορά στην ακρίβεια (accuracy). Ωστόσο ο Sigmoid έχει τη χειρότερη επίδοση αν θεωρήσουμε τα αποτελέσματα για τα πιο εύρωστα μέτρα, δηλαδή για την AUC και το γεωμετικό μέσο.

Πίνακας 5.13: Σύγκριση της επίδοσης του κλασικού SVM για διαφορετικούς πυρήνες, στο ιατρικό σύνολο δεδομένων

Kernel	Ακρίβεια		Ευαισθησία		Ειδικότητα		AUC		Γεωμετρικός μέσος	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Γραμμικός (Linear)	0.993	0.988	0.869	0.844	0.999	0.994	0.920	0.904	0.917	0.899
Κανονικός (Gaussian)	0.992	0.985	0.849	0.779	0.999	0.998	0.947	0.887	0.946	0.879
Πολυωνυμικός (polynomial)	0.975	0.970	0.523	0.500	0.998	0.996	0.970	0.904	0.97	0.899
Sigmoid	0.98	0.977	0.757	0.786	0.992	0.983	0.805	0.832	0.785	0.818

Στο σημείο αυτό πρέπει να σημειώσουμε ότι υπάρχει μία υπερπροσαρμογή των δεδομένων, ιδιαίτερα στην περίπτωση των μη γραμμικών πυρήνων λαμβάνοντας υπόψη τα προηγούμενα μέτρα απόδοσης. Παράλληλα, εκτελώντας ταξινόμηση SVM χωρίς την επιλογή της δειγματοληψίας, παρατηρούμε ότι οι τιμές του γεωμετρικού μέσου προκύπτουν συστηματικά χαμηλές.

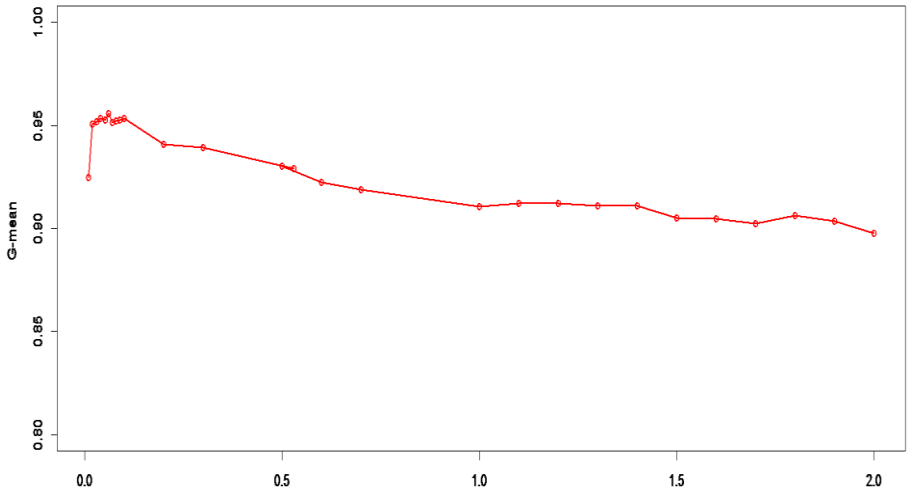
5.2.3 Προσεγγίσεις για μη ισορροπημένα δεδομένα

Two-cost SVM (TC SVM)

Για να εφαρμόσουμε το TCSVM πρέπει να καθορίσουμε δύο κόστη, όπως ήδη έχουμε αναφέρει από την περιγραφή της μεθόδου σε προηγούμενο κεφάλαιο. Για την επίτευξη των αναμενόμενων αποτελεσμάτων ταξινόμησης, το σημαντικότερο ρόλο στην κατασκευή του μοντέλου ευαίσθητης μάθησης (cost sensitive learning model) παίζουν τα προαναφερθέντα κόστη (misclassification costs).

Ανακαλύπτουμε τις βέλτιστες παραμέτρους βασιζόμενοι σε διαφορετικές συναρτήσεις αξιολόγησης, όπως το γεωμετρικό μέσο και το AUC. Για το ιατρικό σύνολο δεδομένων η μειονοτική κλάση (minority class) αποτελείται από τις θετικές πειραματικές εκτελέσεις (δηλαδή το θάνατο του ασθενούς) και η πλειοψηφική κλάση (majority class) αποτελείται από τις αρνητικές πειραματικές εκτελέσεις (δηλαδή την επιβίωση του ασθενούς). Οι δύο παράμετροι του κόστους είναι το κόστος της μειοψηφίας (C^+) που όπως προείπαμε αναφέρεται στις θετικές περιπτώσεις και το κόστος πλειοψηφίας (C^-) αναφέρεται στις αρνητικές περιπτώσεις. Μπορούμε να μειώσουμε τις επιπτώσεις που έχει η ανισορροπία των δύο κλάσεων της μεταβλητής απόκρισης, αναθέτοντας υψηλότερο κόστος (ή ένα μεγαλύτερο βάρος-weight) ταξινόμησης στις πειραματικές εκτελέσεις που αποτελούν τη μειονοτική κλάση (minority class) απ' ότι στις πειραματικές εκτελέσεις που αποτελούν την πλειοψηφική κλάση (την κλάση δηλαδή με τα περισσότερες πειραματικές εκτελέσεις).

Οι Veropoulos et al. (1999) και Akbani et al. (2004) πρότειναν, ως μία καλή επιλογή για το κόστος, το αντίστροφο ποσοστό μεταξύ των δύο κλάσεων (minority/majority) τονίζοντας ότι βελτιώνει την απόδοση της μεθόδου TCSVM. Εκτελώντας μια εκτενή αναζήτηση ανάμεσα σε διαφορετικές τιμές για τα δύο κόστοι, επιβεβαιώσαμε το προαναφερθέν αποτέλεσμα.

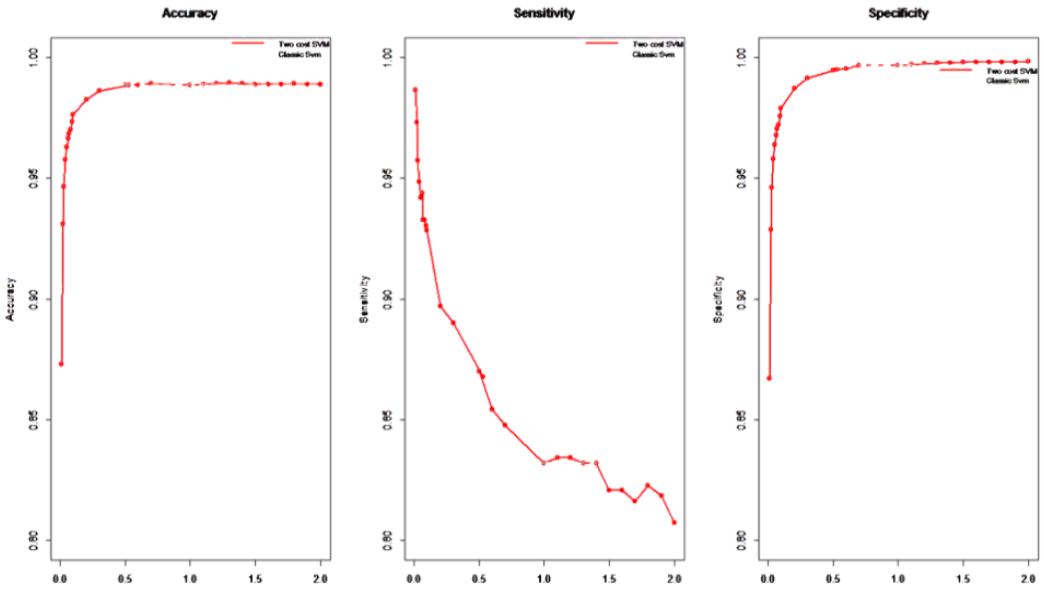


Σχήμα 5.16: Γεωμετρικός μέσος (y-άξονας) αλλάζοντας το κόστος της κλάσης πλειοψηφίας (x-άξονας)

Το ποσοστό μεταξύ της minority και majority κλάσης για το ιατρικό σύνολο δεδομένων ισούται με 0.05299. Πιο συγκεκριμένα, με τον καθορισμό του κόστους της κλάσης με τις λιγότερες εκτελέσεις (minority) να ισούται με 1 και την αλλαγή του κόστους της κλάσης με τις περισσότερες εκτελέσεις (majority) πραγματοποιήσαμε μια ενδελεχή έρευνα μεταξύ πολλών τιμών.

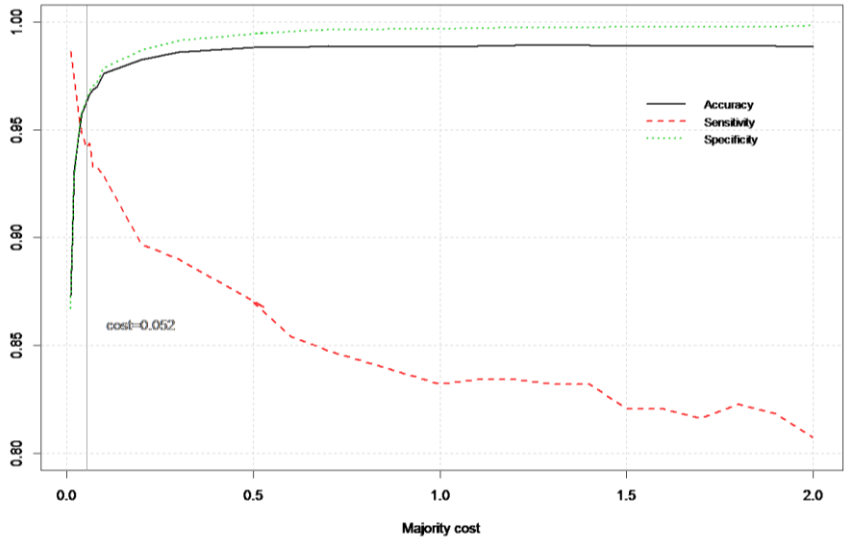
Εκτελέσαμε την ανάλυση για τιμές που κυμαίνονταν μεταξύ 0.01 και 2.0. Τα πιο ακριβή αποτελέσματα όσον αφορά την τιμή του γεωμετρικού μέσου προέκυψαν για τις τιμές 0.04, 0.0529 (= minority/majority) και 0.06 του κόστους της κλάσης πλειοψηφίας, όπως συμπεραίνουμε και από το Σχήμα 5.16. Την καλύτερη απόδοση δίνει η τιμή 0.06. Ωστόσο, οι δύο άλλες τιμές έδωσαν σχεδόν παρόμοια αποτελέσματα. Τελικώς, επιλέξαμε το αντίστροφο μεταξύ των δύο κλάσεων, θέτοντας το ποσοστό ίσο με την αναλογία της κατηγορίας μειοψηφίας προς την κατηγορία πλειοψηφίας ($C^- = C^+ *$ 0.05299).

Το Σχήμα 5.17 απεικονίζει σε ξεχωριστά γραφήματα την απόδοση στην ακρίβεια, την ευαισθησία και την ειδικότητα, αλλάζοντας το κόστος της κατηγορίας πλειοψηφίας (majority class). Η διακεκομμένη γκρι γραμμή δείχνει την τιμή του κάθε μέτρου στην περίπτωση του κλασικού SVM ταξινομητή.



Σχήμα 5.17: Επίδοση βάσει των τριών μέτρων **(Accuracy, Sensitivity, Specificity)** αλλάζοντας το κόστος της κλάσης πλειοψηφίας (x-axis) (συμπαγής κόκκινη γραμμή: TCSVM; διακεκκομένη γραμμή: Classic SVM)

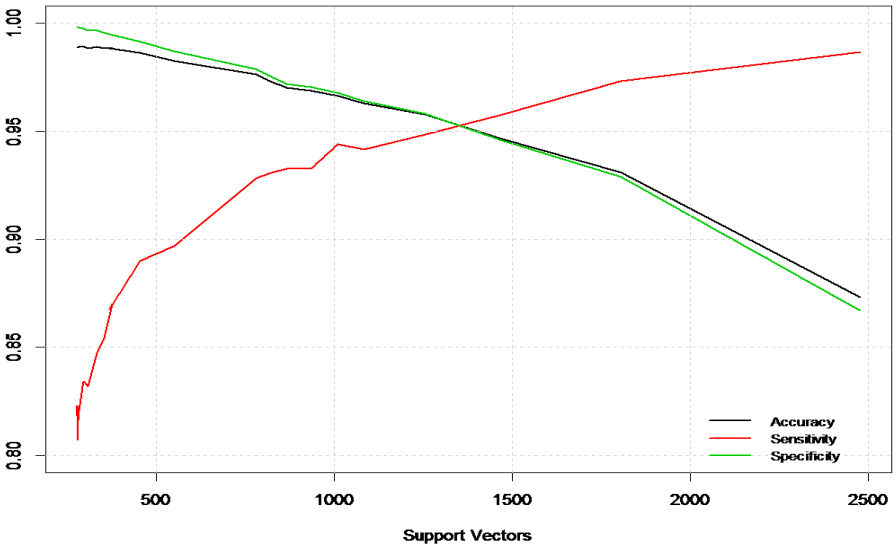
Στο Σχήμα 5.18 θεωρούμε τις συγκρίσεις που αναφέρονται παραπάνω, αλλά στο ίδιο γράφημα. Η κάθετη γκρι γραμμή υποδεικνύει το κόστος της κλάσης πλειοψηφίας που θέσαμε να είναι ίσο με την αναλογία μεταξύ των δύο κλάσεων.



Σχήμα 5.18: Επίδοση ως προς την ακρίβεια, την ευαισθησία και την ειδικότητα της majority κλάσης (x-άξονας) (Μαύρη συνεχής γραμμή: Ακρίβεια, Διακεκομμένη κόκκινη γραμμή: Ευαισθησία, Διακεκομμένη πράσινη γραμμή: Εξειδίκευση). Η κάθετη γραμμή υποδεικνύει το κόστος της κλάσης πλειοψηφίας όταν ορίστηκε να είναι ίσο με το ποσοστό μεταξύ των δύο κλάσεων.

Αντιθέτως, η ακρίβεια (accuracy) και η ειδικότητα (specificity) συγκέντρωσαν υψηλότερες τιμές για λιγότερα διανύσματα υποστήριξης.

Σημειώνουμε εδώ ότι με την αύξηση του majority κόστους έχουμε και πάλι λιγότερα διανύσματα υποστήριξης. Όπως μπορούμε να συμπεράνουμε από το Σχήμα 5.19, η ευαισθησία (sensitivity) αυξάνεται συνεχώς καθώς αυξάνονται τα διανύσματα υποστήριξης.



Σχήμα 5.19: Τρια διαφορετικά μέτρα (y-άξονας) σε σχέση με τα διανύσματα υποστήριξης (x-άξονας) για το TCSVM.

Σύγκριση C και TC SVM

Παρουσιάζουμε ορισμένες συγκρίσεις μεταξύ του C και του TC SVM προκειμένου να τονιστεί η σημασία της εφαρμοζόμενης μεθοδολογίας σε δεδομένα με ανισορροπία μεταξύ των κλάσεων. Πρώτα απ' όλα, παρουσιάζουμε την απόδοση για τη γραμμική περίπτωση και στη συνέχεια ακολουθεί η μη γραμμική με την παράθεση των τριών διαφορετικών πυρήνων. Τα παρακάτω αποτελέσματα υπολογίστηκαν μετά την επιλογή των βέλτιστων παραμέτρων που παραθέσαμε στην προηγούμενη ενότητα. Στον Πίνακας 5.144 παρουσιάζονται τα αποτελέσματα όπου φαίνεται ότι το SVM συγκεντρώνει μεγαλύτερη ακρίβεια τόσο στο σύνολο εκαπίδευσης όσο και στο σύνολο ελέγχου-δοκιμής.

Πίνακας 5.14: Συγκρίσεις επίδοσης για το C και το TC SVM (γραμμική περίπτωση)

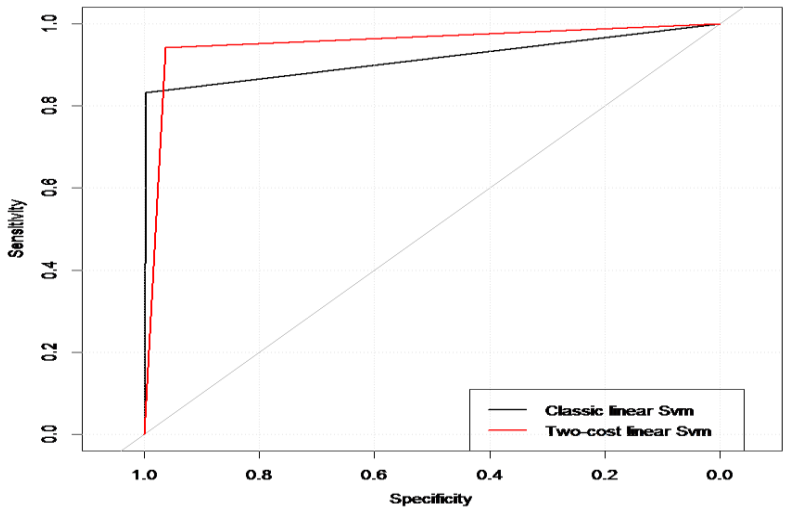
	Ακρίβεια		Ευαισθησία		Ειδικότητα	
SVM	Train	Test	Train	Test	Train	Test
C	0.9929	0.9875	0.86986	0.84416	0.99929	0.99536
TC	0.9717	0.9648	0.94863	0.93506	0.97293	0.96643

Συγκρίνοντας τον κλασικό και τον TC SVM ταξινομητή, ο πρώτος παρουσιάζει μεγαλύτερη ειδικότητα που σημαίνει ότι ο ταξινομητής αναγνωρίζει περισσότερα αληθώς αρνητικά (actual negatives). Με άλλα λόγια, αυτό σημαίνει ότι με τη χρήση του TCSVM παίρνουμε μικρότερο ποσοστό σφάλματος τύπου I. Αυτό το μέτρο από μόνο του δεν μας λέει πόσο καλά ο ταξινομητής αναγνωρίζει τις θετικές περιπτώσεις και γι' αυτό είναι απαραίτητο να ληφθούν υπόψη τόσο η ευαισθησία όσο και η ειδικότητα των ταξινομητών. Όταν οι δύο αλγόριθμοι αξιολογούνται με βάση την ευαισθησία, ο TCSVM έχει σαφές πλεονέκτημα συγκεντρώνοντας υψηλότερο ποσοστό, πράγμα που σημαίνει ότι το ποσοστό του σφάλματος τύπου II είναι χαμηλότερο από εκείνο του CSVM (κλασικό SVM).

Πίνακας 5.15: Συγκρίσεις επίδοσης σε πιο εύρωστα μέτρα για το C και το TC SVM (γραμμική περίπτωση)

	AUC		Geometric mean	
SVM	Train	Test	Train	Test
C	0.9346	0.9198	0.9323	0.9166
TC	0.9608	0.9507	0.9607	0.9506

Λαμβάνοντας υπόψη τα πιο εύρωστα και στην περίπτωση μας πιο αξιόπιστα μέτρα, όπως είναι αυτά του AUC και του Γεωμετρικού μέσου λαμβάνουμε τις τιμές που παρουσιάζονται στον Πίνακας 5.15. Η AUROC επιτυγχάνει την τιμή του 0,9198 για το γραμμικό SVM ενώ το η αντίστοιχη τιμή για το TC SVM ισούται με 0,9507, σαφώς υψηλότερη. Όχι μόνο από την άποψη της AUC, αλλά και λαμβάνοντας υπόψη τον γεωμετρικό μέσο, η μέθοδος ευαίσθητης μάθησης που χρησιμοποιεί τα κόστη αντισταθμίζει τα ποσοστά που δίνει το κλασικό SVM.



Σχήμα 5.20: Σύγκριση των Roc καμπύλων για τη γραμμική περίπτωση

Στο Σχήμα 5.20 εμφανίζονται οι ROC καμπύλες που προέρχονται από τις δύο μεθόδους (C και TC SVM). Όσο πιο πάνω από τη γραμμή αναφοράς βρίσκεται η καμπύλη, τόσο πιο ακριβές είναι το τεστ. Η AUROC φτάνει την τιμή 0.9198 για τη γραμμική περίπτωση του κλασικού SVM ταξινομητή και υψηλότερη τιμή (= 0.9507) φαίνεται να δίνει το TC SVM. Οχι μόνο βάσει του AUC αλλά και βάσει του Γεωμετρικού μέσου η TC μέθοδος υπερτερεί της κλασικής περίπτωσης. Ο Πίνακας 5.16 περιγράφει την απόδοση του κλασικού SVM και του TC SVM για τη μη γραμμική περίπτωση θεωρώντας τα τρία βασικά μέτρα, δηλαδή την ακρίβεια, την ευαισθησία και την ειδικότητα.

Πίνακας 5.16: Κριτήρια επίδοσης για τις δύο διαφορετικές SVM τεχνικές (μη γραμμική περίπτωση)

Kernel	SVM	Accuracy		Sensitivity		Specificity	
		Train	Test	Train	Test	Train	Test
Gaussian	C	0.9924	0.9848	0.84932	0.77922	0.99982	0.99607
	TC	0.9655	0.9563	0.93493	0.94805	0.96706	0.95679
Polynomial	C	0.9748	0.9702	0.52397	0.50000	0.99822	0.99607
	TC	0.9749	0.9692	0.70890	0.71429	0.98878	0.98321
Sigmoid	C	0.98	0.977	0.75685	0.78571	0.99163	0.98750
	TC	0.969	0.9631	0.91096	0.94156	0.97204	0.96429

Στον παραπάνω πίνακα, C είναι η συντομογραφία για το κλασικό SVM και TC για το TC SVM

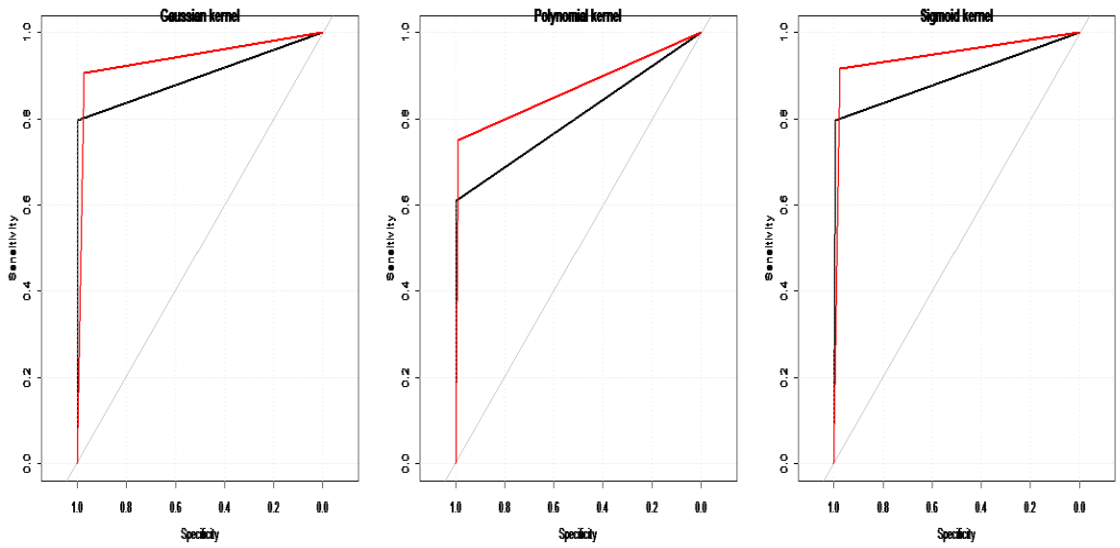
Ο Πίνακας 5.17 παρουσιάζει την αποδόση λαμβάνοντας υπόψη τα πιο εύρωστα της AUC και του Γεωμετρικού μέσου. Την κάλυτερη απόδοση βάσει του Γεωμετρικού μέσου για το TC SVM εμφανίζει ο κανονικός (Gaussian) πυρήνας. Συγκριτικά αποτελέσματα παίρνουμε χρησιμοποιώντας τον σιγμοειδή πυρήνα για όλα τα μέτρα φτάνοντας το ποσοστό του 95.20% για το GM στην TC μέθοδο.

Πίνακας 5.17: Κριτήρια επίδοσης για τις δύο διαφορετικές SVM τεχνικές (μη γραμμική περίπτωση)

Kernel	SVM	AUC		GM	
		Train	Test	Train	Test
Gaussian	C	0.9246	0.8876	0.9215	0.8809
	TC	0.951	0.9524	0.9509	0.9524
Polynomial	C	0.7611	0.748	0.7232	0.7057
	TC	0.8488	0.8488	0.8372	0.8380
Sigmoid	C	0.8742	0.8866	0.8663	0.8808
	TC	0.9415	0.9520	0.9410	0.9520

Στον παραπάνω πίνακα, C είναι η συντομογραφία για το κλασικό SVM και TC για το TC SVM

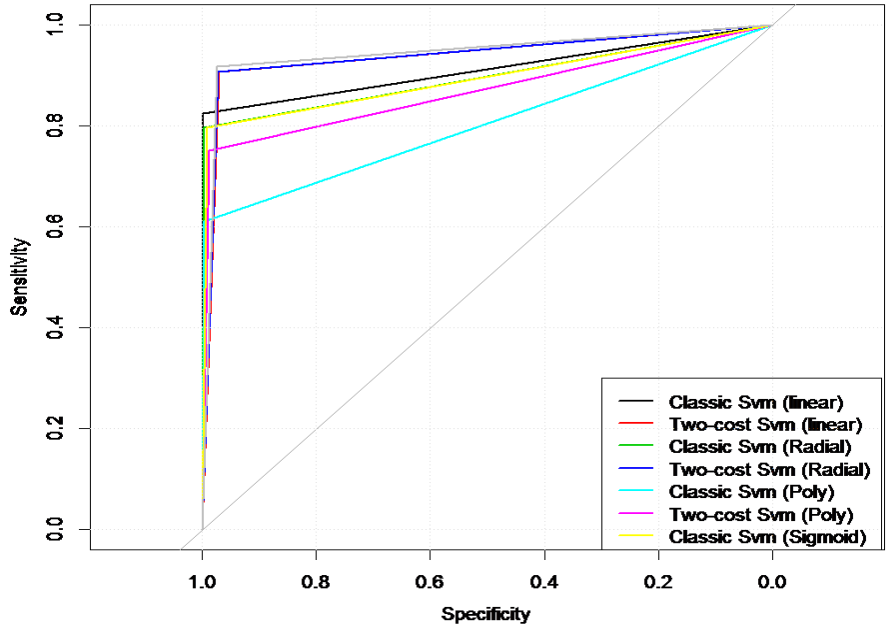
Θα πρέπει να σημειωθεί ότι η χρήση της μεθόδου ευαίσθητης μάθησης μειώνει το πρόβλημα της υπερπροσαρμογής. Σχεδόν παρόμοια αποτελέσματα δόθηκαν λαμβάνοντας υπόψη την AUC αντί του Γεωμετρικού Μέσου ως μέτρο απόδοσης. Ο κανονικός πυρήνας έχει σαφώς τα υψηλότερα GM και AUC σε σύγκριση με όλους τους μη-γραμμικούς πυρήνες θεωρώντας το TC SVM ενώ ο πολυωνυμικός πυρήνας εμφανίζει τα χαμηλότερα ποσοστά. Η διαφορά μεταξύ των δύο πυρήνων, κανονικού και σιγμοειδούς, είναι τόσο μικρή που θα λέγαμε ότι και οι δύο να επιτυγχάνουν καλά αποτελέσματα για όλα τα μέτρα. Επιπλέον, το TC SVM αποδίδει καλά στη γραμμική περίπτωση. Στο Σχήμα 5.21 παρουσιάζεται μια σύγκριση σε σχέση με την απόδοση, θεωρώντας τον γεωμετρικό μέσο, που επιβεβαιώνει τα παραπάνω συμπεράσματα. Συγκρίνοντας το TC με το C SVM για τον κανονικό πυρήνα, μπορούμε να συμπεράνουμε ότι η πρώτη μέθοδος υπερτερεί της δεύτερης από την άποψη του γεωμετρικού μέσου και της AUC. Αντίθετα, όσον αφορά στον πολυώνυμικό πυρήνα, η διαφορά μεταξύ των δύο μεθόδων είναι σημαντικά υψηλότερη από ότι στους άλλους δύο πυρήνες.



Σχήμα 5.21: Σύγκριση των καμπύλων ROC για τη μη γραμμική περίπτωση στο σύνολο ελέγχου.
Οι κόκκινες καμπύλες αναπαριστούν το TC SVM και οι μαύρες το κλασικό SVM.

Το Σχήμα 5.21 εμφανίζει τις καμπύλες ROC που προέρχονται από όλες τις SVMs με τους τρεις μη γραμμικούς πυρήνες. Για τις καμπύλες ROC στο Σχήμα 5.21, όσον αφορά στον κανονικό πυρήνα, η μέθοδος TC αποδίδει καλύτερα κατά μέσο όρο σε σύγκριση με τους άλλους πυρήνες αν και η διαφορά με τον σιγμοειδή πυρήνα δεν είναι στατιστικά σημαντική. Όπως μπορούμε να συμπεράνουμε από το Σχήμα 5.21, ο πολυωνυμικός πυρήνας παρουσιάζει τη χειρότερη απόδοση.

Το Σχήμα 5.22 απεικονίζει την απόδοση αυτών των δύο μεθόδων για το ιατρικό σύνολο δεδομένων για όλους τους υπό εξέταση πυρήνες. Επιπλέον, κατατάσσει τα καλύτερα υποψήφια μοντέλα σύμφωνα με το κριτήριο της AUC και μας βοηθά να επιλέξουμε την καλύτερη προσέγγιση για τη συγκεκριμένη ανάλυση.



Σχήμα 5.22: ROC καμπύλες για όλους τους πυρήνες για τις μεθόδους C και TC.

Η υψηλότερη AUC ελήφθη για τη μέθοδο TC με Gaussian πυρήνα (AUC = 0,9524) και η δεύτερη υψηλότερη σημειώθηκε για τη μέθοδο TC, τόσο χρησιμοποιώντας το γραμμικό πυρήνα (AUC = 0,9507) όσο και το Sigmoid πυρήνα (AUC = 0,9520), με το δεύτερο να υπερτερεί ελαφρώς του πρώτου. Σχεδόν παρόμοια αποτελέσματα προέκυψαν για το κλασικό γραμμικό SVM (AUC = 0,9198) αλλά και για το κλασικό SVM με Gaussian πυρήνα (AUC = 0,8876). Η AUROC για τον πολυωνυμικό πυρήνα έδωσε τη χαμηλότερη απόδοση που ισούται με 0,748 και 0,8488 για το κλασικό και το TC SVM αντίστοιχα.

Μέθοδοι επαναδειγματοληψίας (Resampling Methods)

Random Sampling SVM

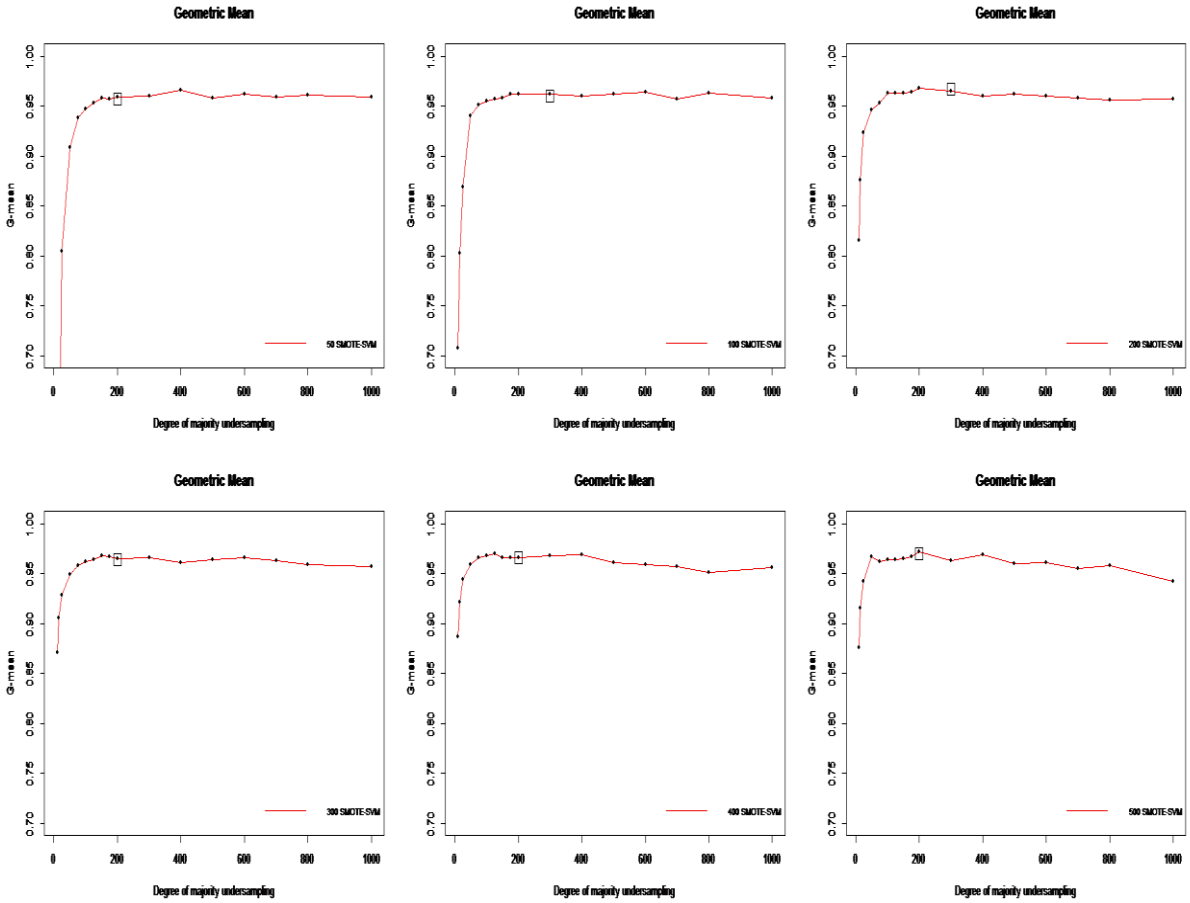
- Τυχαία Υπερ-δειγματοληψία (Random Over-sampling) (SVM-RO)
  Η μάθηση με δείγματα εκπαίδευσης όπου εφαρμόσαμε υπερ-δειγματοληψία επαναλήφθηκε 20 φορές για κάθε μέγεθος των αυξημένων συνόλων εκπαίδευσης. Στη συνέχεια, επιλέξαμε το σύνολο εκπαίδευσης που παρήγαγε τη μέγιστη τιμή του Γεωμετρικού Μέσου για το αρχικό σύνολο εκπαίδευσης.

- Τυχαία Υπό-δειγματοληψία (Random Under-sampling) (SVM-RU)
  Επίσης εφαρμόσαμε τυχαία υπο-δειγματοληψία των πειραματικών εκτελέσεων της κλάσης πλειοψηφίας. Όπως στην περίπτωση της υπερδειγματοληψίας, η μάθηση με

δείγματα εκπαίδευσης όπου εφαρμόσαμε υπο-δειγματοληψία επαναλήφθηκε 20 φορές για κάθε μέγεθος του μειωμένου σύνολο εκπαίδευσης. Στη συνέχεια, εμείς επιλέξαμε το μειωμένο σύνολο εκπαίδευσης που παρήγαγε τη μέγιστη τιμή του Γεωμετρικού Μέσου για το αρχικό σύνολο εκπαίδευσης.

SMOTE-SVM and undersampling combination

Όσον αφορά τον αλγόριθμο SMOTE, για τον υπολογισμό των k-πλησιέστερων γειτόνων, θέσαμε για k ίσο με 5. Η διαδικασία της μάθησης πραγματοποιήθηκε χρησιμοποιώντας 20 ανεξάρτητα "συνθετικά" επαυξανόμενα (synthetically enhanced) σύνολα δεδομένων και στη συνέχεια, προκειμένου να προσδιορίσουμε το καλύτερο "συνθετικό" μέγεθος του δείγματος υπολογίζουμε τη μέγιστη τιμή του γεωμετρικού μέσου.



Σχήμα 5.23: Τιμές του γεωμετρικού μέσου στο σύνολο εκπαίδευσης σε σχέση με την αύξηση των συνθετικών παραδειγμάτων στην κλάση μειωψηφίας από την μέθοδο SMOTE

Στο παράδειγμα, επιλέγεται μια αύξηση της τάξεως του 300%, αν η μέγιστη τιμή του γεωμετρικού μέσου του αρχικού συνόλου εκπαίδευσης εμφανίζεται όταν προστίθενται στο σύνολο δεδομένων εκπαίδευσης 300% των νέων συνθετικών περιπτώσεων.

Ενώ αυξάνουμε σταδιακά τις πειραματικές εκτελέσεις της κλάσης με τις λιγότερες πειραματικές εκτελέσεις (minority class) και ταυτόχρονα μειώνουμε τις πειραματικές εκτελέσεις της κλάσης πλειοψηφίας (majority class), παρατηρήσαμε για κάθε συνδυασμό τις τιμές του γεωμετρικού μέσου των αρχικών συνόλων εκπαίδευσης για κάθε πειραματικό σύνολο δεδομένων. Χρησιμοποιώντας το SVM-SMOTE, ο αριθμός των συνθετικών περιπτώσεων για να επιτευχθεί η επιθυμητή ισορροπία μεταξύ των κλάσεων είναι άγνωστος και για το λόγο αυτό θα πρέπει να εκτελεστούν εμπειρικές μελέτες.

Πίνακας 5.18: **Grid search για διαφορετικούς συνδυασμούς του SMOTE SVM και της τυχαίας υποδειγματοληψίας.**

Gaussian Kernel	**Geometric mean**					
Under-sampling %	50 SMOTE	100 SMOTE	200 SMOTE	300 SMOTE	400 SMOTE	500 SMOTE
10%	0.230069	0.707719	0.815265	0.871231	0.886499	0.875904
15%	0.521103	0.802191	0.876140	0.905183	0.921272	0.915015
25%	0.804562	0.868991	0.923027	0.928704	0.944614	0.941737
50%	0.908342	0.940197	0.946149	0.948743	0.959388	0.966566
75%	0.938477	0.950819	0.953434	0.958444	0.965468	0.961461
100%	0.946934	0.955301	0.962648	0.962282	0.967766	0.964178
125%	0.953191	0.956700	0.963171	0.964419	0.969509	0.963771
150%	0.958011	0.958382	0.962707	**0.967735**	0.966161	0.964967
175%	0.956627	0.961705	0.964181	0.966504	0.965536	0.966931
200%	0.958892	0.962331	**0.967519**	0.964554	0.965711	**0.971865**
300%	0.960314	0.961802	0.964521	0.966229	0.967882	0.963321
400%	**0.966179**	0.960241	0.959626	0.960879	**0.968944**	0.968494
500%	0.957841	0.961965	0.962146	0.963540	0.961394	0.960416
600%	0.961914	**0.963472**	0.959781	0.966203	0.959285	0.960570
700%	0.958990	0.956811	0.957587	0.963160	0.957262	0.955441
800%	0.961374	0.963265	0.956116	0.959472	0.951091	0.957553
1000%	0.959371	0.957840	0.957446	0.956986	0.956018	0.942184
2000%	0.945764	0.943434	0.935558	0.930169	0.940361	0.928435

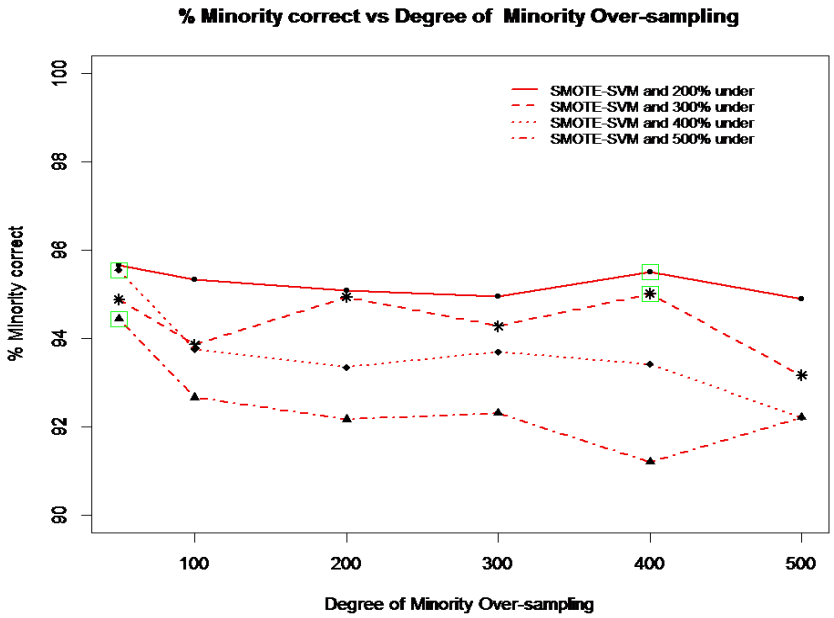
Στην κλάση μειοψηφίας (minority class) εκτελέσαμε υπερ-δειγματοληψία σε 50%, 100%, 200%, 300%, 400%, 500% και στην κλάση πλειοψηφίας εκτελέσαμε υποδειγματοληψία 10%, 15%, 25%, 50%, 75%, 100%, 125%, 150%, 175%, 200%,

300%, 400%, 500%, 600%, 700%, 800%, 1000%, 2000%, όπως παρουσιάζονται στον Πίνακας 5.18.

Πίνακας 5.19: Σύγκριση του ποσοστού επιτυχίας στην κλάση μειωψηφίας για διαφορετικά ποσοστά υποδειγματοληψίας αλλάζοντας το ποσοστό υπερδειγματοληψίας

Gaussian Kernel	% Minority correct			
Smote	200% under-sampling	300% under-sampling	400% under-sampling	500% under-sampling
50%	**0.9566563**	0.9487952	0.9554896	0.9444444
100%	0.9534161	0.9386503	0.9375000	0.9266055
200%	0.9509202	0.9494048	0.9335260	0.9216301
300%	0.9495549	0.9427711	0.9369369	0.9230769
400%	0.9549550	0.9501466	0.9341317	0.9120235
500%	0.9489489	0.9316770	0.9221557	0.9221557

Επιλέξαμε τα προαναφερθέντα ποσοστά, σύμφωνα με τους Chawla et αl. (2002). Λόγω των καλών επιδόσεων του κανονικού (Gaussian) πυρήνα επιλέξαμε αυτόν για την SVM ταξινόμηση.



Σχήμα 5.24: Γραφική απεικόνηση του ποσοστού επιτυχίας στην κλάση μειωψηφίας για διαφορετικά ποσοστά υποδειγματοληψίας αλλάζοντας το ποσοστό υπερδειγματοληψίας.

Υπενθυμίζουμε ότι παρουσιάζει την καλύτερη απόδοση μεταξύ όλων των πυρήνων (γραμμικών και μη γραμμικών) στο σύνολο δεδομένων μας.

Για το SVM-SMOTE, η μέγιστη τιμή του γεωμετρικού μέσου βρέθηκε για τα ποσοστά 500% σε συνδυασμό με το πόσο του 200% υποδειγματοληψίας, επιτυγχάνοντας την τιμή του 97,18652% για την τιμή του ΓΜ. Ο Πίνακας 5.19 δείχνει μία αναζήτηση μεταξύ των διαφόρων ποσοστών υπό-δειγματοληψίας για το 50%, 100%, 200%, 300%, 400%, 500% του SMOTE-SVM αντιστοίχως. Το Σχήμα 5.24 δείχνει το ποσοστό των τιμών των minority correct των αρχικών συνόλων εκπαίδευσης, καθώς προστίθενται πειραματικές εκτελέσεις από τη μέθοδο SMOTE με 4 διαφορετικά ποσοστά υπο-δειγματοληψίας. Η υψηλότερη τιμή παρουσιάζεται στην περίπτωση με το συνδυασμό του 50 SMOTE-SVM και 200% υπό-δειγματοληψία.

5.2.4 Πειραματικά αποτελέσματα και συγκρίσεις

Οι τιμές του Γεωμετρικού μέσου για το πραγματικό-αρχικό σύνολο δεδομένων χρησιμοποιώντας τους 5 διαφορετικούς πυρήνες, φαίνονται στον παρακάτω πίνακα (Πίνακας 5.20). Οι συγκρίσεις μεταξύ όλων των μεθόδων, έδειξαν ότι το SMOTE-SVM και η μέθοδος της υπερδειγματοληψίας έχουν την καλύτερη επίδοση στο σύνολο ελέγχου.

Πίνακας 5.20: Γεωμετρικός Μέσος του συνόλου εκπαίδευσης που αποκτήθηκαν από τις 4 διαφορετικές μεθόδους

Gmean of training set in original data		
Linear kernel	Train	Test in original train data
C SVM	0.9192	0.9248457
TC SVM	0.98138	0.9699514
SVM-RU	0.9824	0.9579060
SVM-RO	0.9792	0.9789352
SMOTE	0.9681217	0.9687785

Ωστόσο χρησιμοποιώντας τυχαία-υπερδειγματοληψία υπάρχει ένα πρόβλημα, αυτό της υπερ-προσαρμογής των δεδομένων που παράλληλα είναι πιο πιθανό να συμβεί με τη χρήση μη γραμμικών πυρήνων. Για το λόγο αυτό, το SMOTE-SVM φαίνεται να έχει την καλύτερη δυνατή απόδοση. Το SMOTE SVM ξεπέρασε ελαφρώς το SVM-RU και την μεροληπτική μέθοδο με τα κόστη (TC SVM). Συγκρίνοντας τους τρεις μη γραμμικούς πυρήνες για τις μεθόδους που θεωρήσαμε παρατηρούμε ότι ο Γκαουσιανός πυρήνας παρουσιάζει την υψηλότερη απόδοση για το ΓΜ. Όσον αφορά στις μεθόδους, το

SMOTE-SVM έχει την υψηλότερη απόδοση σε σύγκριση με το CSVM, το TCSVM και το SVM-RU.

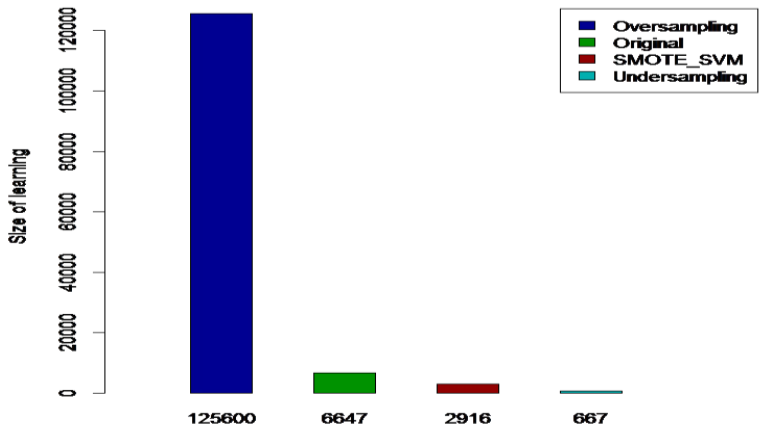
Στην περίπτωση του SVM-RO δεν φαίνεται σημαντική διαφορά για τον γκαουσιανό και τον σιγμοειδή πυρήνα, ενώ το SVM-RO φαίνεται να υπερτερεί του SMOTE-SVM στην περίπτωση του πολυωνυμικού πυρήνα.

Η υπερδειγματοληψία, παρά τις υψηλές τιμές του Γεωμετρικού μέσου, οδηγεί σε μεγάλη αύξηση του συνόλου εκπαίδευσης κάτι που όχι μόνο αυξάνει την υπολογιστική επιβάρυνση του αλγορίθμου μάθησης, αλλά οδηγεί επίσης σε overfitting προβλήματα.

Πίνακας 5.21: Γεωμετρικός Μέσος του συνόλου εκπαίδευσης που αποκτήθηκαν από τις 4 διαφορετικές μεθόδους (μη-γραμμική περίπτωση)

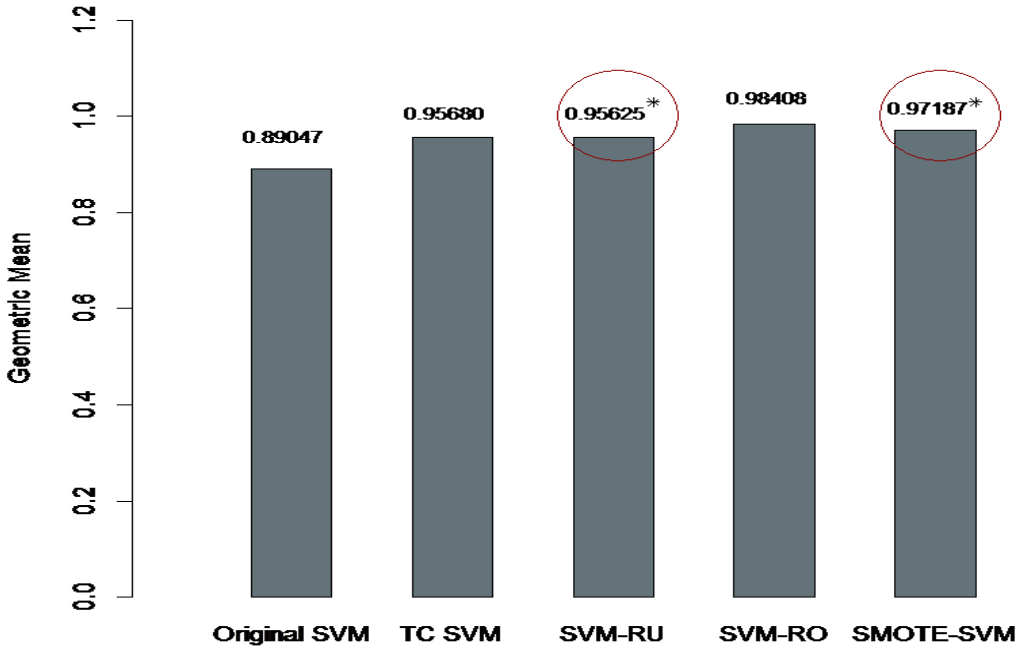
	Gmean of training set in original data		
	Gaussian kernel	Polynomial	Sigmoid
C SVM	0.8904698	0.7866470	0.8596290
TC SVM	0.9567920	0.8732783	0.9561292
SVM-RU	0.9562475	0.8782818	0.9479898
SVM-RO	0.9840769	0.9529313	0.9568208
SMOTE SVM	0.9718652	0.880000	0.9606813

Το SMOTE-SVM σε συνδυασμό με το undersampling δεν φαίνεται να προκαλεί προβλήματα υπερπροσαρμογής και ταυτόχρονα διατηρεί μικρότερο σύνολο δεδομένων σε σύγκριση με τη μέθοδο της υπερ-δειγματοληψίας. Στο Σχήμα 5.25 απεικονίζονται τα μεγέθη των δειγμάτων εκπαίδευσης για τις 4 διαφορετικές μεθόδους που εφαρμόσαμε.



Σχήμα 5.25: Μέγεθος των δειγμάτων εκπαίδευσης

Κατά συνέπεια, όπως αναμενόταν, η τυχαιότητα της υπο-δειγματοληψίας δεν παράγει συνεπή αποτελέσματα σε σχέση με τη μέθοδο SMOTE. Όπως μπορούμε να συμπεράνουμε, η υπερδειγματοληψία αυξάνει σε μεγάλο βαθμό το μέγεθος των δειγμάτων εκπαίδευσης, παρέχοντας μια υπολογιστική επιβάρυνση στον αλγόριθμο SVM που εφαρμόζεται στο δεύτερο στάδιο.



Σχήμα 5.26 Σύγκριση των τιμών του Γεωμετρικού Μέσου για τα σύνολα εκπαίδευσης, για τις 5 διαφορετικές μεθόδους. Έγινε χρήση του Γκαουσιανού πυρήνα στο ιατρικό σύνολο δεδομένων.

Σχήμα 5.26 παρουσιάζονται οι θεωρούμενες μέθοδοι λαμβάνοντας ως μέτρο απόδοσης την τιμή του Γεωμετρικού Μέσου. Κατά συνέπεια τα αποτελέσματα που αναφέρθηκαν παραπάνω επιβεβαιώνονται με αυτή την γραφική παρουσίαση.

ΚΕΦΑΛΑΙΟ 6

ΣΥΜΠΕΡΑΣΜΑΤΑ

Ένα ιδιαίτερα σημαντικό θέμα σήμερα είναι εκείνο της στατιστικής ανάλυσης μεγάλων συνόλων δεδομένων, καθώς και το θέμα της μάθησης από πολύπλοκες και μεγάλες δομές δεδομένων. Η φύση των ιατρικών συνόλων δεδομένων είναι πραγματικά περίπλοκη και η ανάλυσή τους έχει μεγάλη σημασία, έτσι ώστε να έχουμε ακριβή και έγκαιρη ιατρική διάγνωση. Στον τομέα της ιατρικής, όπου οι περιπτώσεις που δεν έχουν κάποιο νόσημα αποτελούν την πλειοψηφία, είναι πιο σημαντική η σωστή ισορροπία μεταξύ ευαισθησίας και της ειδικότητας κάτι που σημαίνει ότι πρέπει να είναι σαφής η διάκριση των ψευδώς αρνητικών αποτελεσμάτων από τα ψευδώς θετικών αποτελεσμάτων. Για το λόγο αυτό είναι επιτακτική η ανάγκη να εκτελεστεί μια ακριβής ανάλυση ευαισθησίας. Ανάμεσα στις μεθόδους και τα εργαλεία ταξινόμησης, οι SVM έχουν αποκτήσει μεγάλη δημοτικότητα λόγω της απόδοσης και της ικανότητάς τους για γενίκευση. Εκτός από τις συνήθεις τεχνικές, υπάρχουν και πολλές τροποποιήσεις που βελτιώνουν την απόδοσή τους. Το κίνητρό μας για την εκπόνηση της συγκεκριμένης μελέτης προέρχεται από την υποστήριξη λήψης ιατρικών αποφάσεων κάτι που δείχνει ότι η επιλογή των ιατρικών συνόλων δεδομένων ήταν επιτακτική. Ο κύριος στόχος της εργασίας μας είναι να αναπτύξουμε μια συγκριτική ανάλυση μεταξύ ορισμένων τροποποιήσεων του SVM και με αυτόν τον τρόπο, να παρέχουμε την πιο αποδοτικό αλγόριθμο που θα επιτρέψει την επιτυχή πρόβλεψη για την έκβαση του αποτελέσματος στα σύνολα δεδομένων που επελέγησαν. Ουσιαστικά διεξήχθησαν δύο μελέτες. Οι πρώτη εξ αυτών περιλαμβάνει 4 μικρότερα σύνολα δεδομένων και βάσει αυτού επιλέχθησαν και οι κατάλληλες μέθοδοι για την ανάλυση τους και η δεύτερη αφορά στην ανάλυση ενός μεγαλύτερου συνόλου για την ανάλυση του οποίου χρειάστηκε η εφαρμογή και άλλων τεχνικών.

6.1 Συμπεράσματα πρώτης μελέτης

Στην πρώτη μελέτη διερευνήσαμε την επίδραση της ενσωμάτωσης του PSVM στο μοντέλο μάθησης των Μηχανών Διανυσματικής υποστήριξης τόσο στη γραμμική όσο και στη μη γραμμική ταξινόμηση, ώστε να μειώσουμε όχι μόνο το χρόνο εκτέλεσης για την εκπαίδευση των δεδομένων, αλλά και τα σφάλματα ταξινόμησης. Πιο συγκεκριμένα αυτή η εργασία παρουσιάζει μια συγκριτική ανάλυση των διαφορετικών στρατηγικών SVM. Ειδικά εκτελείται μια παραλλαγή του προτύπου SVM, το PSVM, που οδηγεί σε ένα απλό πρόβλημα βελτιστοποίησης, μαζί με δύο μεθόδους μάθησης ευαίσθητου κόστους, το TCSVM και μια τροποποίηση του PSVM, το MPSVM, όλα για ιατρικά σύνολα δεδομένων. Η αξιολόγηση της αξιοπιστίας των αλγορίθμων ταξινόμησης είναι απαραίτητη για τη διασφάλιση της ποιότητας των δεδομένων. Χρησιμοποιήσαμε το μέτρο της ακρίβειας και το γεωμετρικό μέσο (GM) που προκύπτει ως συνδυασμός της ευαισθησίας και της ειδικότητας, για τη σύγκριση των αλγορίθμων προκειμένου να παρέχουμε χρήσιμα αποτελέσματα. Επιπλέον, οι προσεγγίσεις ευαίσθητης μάθησης εφαρμόζονται σε ιατρικά σύνολα δεδομένων σε μια προσπάθεια να προβλέψουμε την επιβίωση του ασθενούς, εάν κάποιος είναι αρνητικός ή θετικός σε μια ασθένεια (διαβήτης ή θυρεοειδούς). Η εφαρμογή αυτών των μεθόδων ήταν απαραίτητη προκειμένου να μειωθούν οι συνέπειες της μη ισορροπημένης φύσης των ιατρικών δεδομένων που χρησιμοποιήσαμε. Τα αποτελέσματα της μελέτης επιβεβαιώνουν τα πλεονεκτήματα των εξεταζόμενων προσεγγίσεων και δείχνουν την πολλά υποσχόμενη προοπτική και τη νέα αντίληψη των νέων διαδικασιών.

Σε γενικές γραμμές, οι τροποποιήσεις που παρουσιάζονται αποκαλύπτουν μια εξαιρετική απόδοση σε όλα τα σύνολα δεδομένων κάτι που επιβεβαιώνει τη σημαντικότητα της εφαρμογής εναλλακτικών μεθόδων, ιδίως στον τομέα της ιατρικής. Πιο συγκεκριμένα, για τα δεδομένα Μετάγγισης Αίματος η βέλτιστη τιμή της GM για το γραμμικό TCSVM αγγίζει το ποσοστού του 99,62% και 98,88% για το σύνολο εκπαίδευση και το σύνολο δοκιμής αντίστοιχα. Επιπλέον, το υψηλότερο ποσοστό για το μέτρο της ακρίβειας συγκέντρωσε η μη γραμμική PSVM (92,7%) στο σύνολο εκπαίδευσης και η μη γραμμική SVM (79.03%) στο σύνολο δοκιμών. Η πιο γρήγορη μέθοδος, όπως ήταν αναμενόμενο ήταν η Γραμμική PSVM (0.14). Συγκρίνοντας τις μεθόδους στο σύνολο δεδομένων για το διαβήτη στους Ινδιάνους Pima, παρατηρούμε ότι η MPSVM μαζί με μη γραμμική TCSVM απέδωσε την υψηλότερη απόδοση για το γεωμετρικό μέσο. Θεωρώντας το μέτρο της ακρίβειας ταξινόμησης, οι γραμμικές μέθοδοι, όπως η γραμμική SVM και PSVM, έδωσαν υψηλότερες τιμές της τάξεως του 78,65% για την εκπαίδευση και του 78,12%, 76,04% για τα δεδομένα δοκιμών, αντίστοιχα. Για άλλη μια φορά η γραμμική PSVM (0,11) ήταν ο ταχύτερος αλγόριθμος. Όσον αφορά τις μεθόδους για τα δεδομένα της ασθένειας του θυρεοειδούς, το SVM και το TCSVM είχαν μια εξαιρετική απόδοση στην ακρίβεια ταξινόμησης και ως εκ τούτου το χαμηλότερο ποσοστό εσφαλμένης ταξινόμησης.

Ωστόσο, όπως αναμενόταν, η μη γραμμική TCSVM ήταν η προτιμότερη μεθοδολογία όσον αφορά το μέτρο GM. Σημειώνουμε ότι το σύνολο δεδομένων για την ασθένεια του θυρεοειδούς είναι το μικρότερο ανάμεσα στα τέσσερα που παρουσιάζονται στη μελέτη μας. Κατά συνέπεια, οι συμβατικοί αλγόριθμοι φαίνεται να αποδίδουν πολύ καλά. Στην περίπτωση των πραγματικών ιατρικών δεδομένων, η μη γραμμική TCSVM (94.46% (στην εκπαίδευση) 93,57% (για τη δοκιμή) πέτυχε την υψηλότερη απόδοση σε σύγκριση με τις άλλες μεθόδους για το GM, κάτι που επιβεβαιώνει τη σημασία των σταθμισμένων μεθόδων των μη ισορροπημένων δεδομένων. Όπως μπορούμε να συμπεράνουμε, την υψηλότερη ακρίβεια ταξινόμησης είχε η μη γραμμική PSVM. Αξίζει να αναφέρουμε το γεγονός ότι από την άποψη του χρόνου η γραμμική PSVM (5.60) εμφανίστηκε ο ταχύτερος από όλους τους αλγορίθμους. Θα πρέπει επίσης να αναφέρουμε ότι η MPSVM συγκέντρωσε υψηλά ποσοστά και στα δύο. Και στην ακρίβεια ταξινόμησης αλλά και στην τιμή του Γεωμετρικού Μέσου και φάνηκε να ξεπερνάει το γραμμικό PSVM, κάτι που επιβεβαιώνει τη σημασία αυτής της επανασταθμισμένης μεθόδου για μη ισορροπημένα δεδομένα. Τα πειραματικά αποτελέσματα που παρέχονται στην παρούσα εργασία έχουν αποδείξει ότι οι μέθοδοι PSVM και οι μέθοδοι ευαίσθητου κόστους παρέχουν μια πολύ ανταγωνιστική λύση σε σχέση με τις άλλες υπάρχουσες τεχνικές στη βελτιστοποίηση του μέτρου GM και της ακρίβειας για την διαχείρηση μεγάλων προβλημάτων ταξινόμησης αλλά και μη ισορροπημένων προβλημάτων. Ειδικότερα, σε μεγάλης κλίμακας στατιστική μοντελοποίηση το MPSVM και το PSVM έδειξαν μια εξαιρετική απόδοση σε σχέση με αυτά τα δύο μέτρα.

Τέλος το γραμμικό PSVM είναι ένα εξαιρετικά γρήγορος αλγόριθμος που θα μπορούσε να αποδειχθεί ιδιαίτερα αποδοτικός σε προβλήματα πραγματικού χρόνου ή σε προβλήματα της πραγματικής ζωής. Οι Μηχανές Διανυσματικής Υποστήριξης και οι παραλλαγές τους, είναι ένα ισχυρό εργαλείο για την ταξινόμηση και τη χρήση των SVMs ως μια εναλλακτική μέθοδο με σκοπό την υποστήριξη της ιατρικής έρευνας ως μίας από τις πλέον υποσχόμενες μεθόδους. Ελπίζουμε ότι η μελέτη αυτή θα πείσει όχι μόνο τους ερευνητές στην ιατρική επιστήμη αλλά και τους πειραματιστές σε μεγάλης κλίμακας προβλήματα ταξινόμησης να χρησιμοποιούν όχι μόνο τυπικές τεχνικές SVM αλλά και τις τροποποιήσεις αυτού του αλγορίθμου, όπως PSVM, το TCSVM και το MPSVM με σκοπό όχι μόνο να εξαγάγουν χρήσιμα πρότυπα, αλλά και να μειώσουν την πολυπλοκότητα των υπολογισμών.

6.2 Συμπεράσματα δεύτερης μελέτης

Έχουν προταθεί πολλές στρατηγικές που ασχολούνται με τα μη ισορροπημένα δεδομένα, μερικές από τις οποίες έχουν εφαρμοστεί στην παρούσα ανάλυση. Σε επίπεδο δεδομένων, η δειγματοληψία είναι η συνηθέστερη προσέγγιση, με την τυχαία υπερδειγματοληψία να ξεπερνά σε απόδοση την τυχαία υποδειγματοληψία. Σε αλγοριθμικό επίπεδο, έχουν προταθεί ευρέως αποδεκτές λύσεις χρησιμοποιώντας προσαρμοσμένα κόστη. Χρησιμοποιήσαμε μια εναλλακτική μέθοδο ευαίσθητου κόστους, την TCSVM, που αναφέραμε και προηγουμένως, δεδομένου ότι οι κλασικές SVMs αποδείχθηκαν ακατάλληλες για την αντιμετώπιση του συγκεκριμένου προβλήματος με μη ισορροπημένα δεδομένα. Ερευνήσαμε το αποτέλεσμα της ενσωμάτωσης της TCSVM στο μοντέλο μάθησης SVM χρησιμοποιώντας μία μεγάλη βάση πραγματικών ιατρικών δεδομένων. Σε γενικές γραμμές, το TC SVM φαίνεται να ξεπερνάει το κλασικό SVM για όλους τους πυρήνες που χρησιμοποιούνται σε αυτή τη συγκριτική μελέτη με βάσει τα κριτήρια του AUC και του Γεωμετρικού μέσου, κάτι που επιβεβαιώνει τη σημασία της μεθόδου TC για μη ισορροπημένα δεδομένα. Τα πειραματικά αποτελέσματα που παρουσιάζονται στην παρούσα μελέτη έδειξαν ότι η μέθοδος TC παρέχει μια πολύ ανταγωνιστική λύση σε σχέση με άλλες υφιστάμενες τυποποιημένες μεθόδους στη βελτιστοποίηση κριτηριών που είναι ευαίσθητα στην αλλαγή μεταξύ της ευαισθησίας και της ειδικότητας. Τα αποτελέσματα αυτά επιβεβαιώνουν τα πλεονεκτήματα της εξεταζόμενης προσέγγισης, που δείχνει την πολλά υποσχόμενη προοπτική και νέα κατανόηση της μάθησης ευαίσθητου κόστους. Από την άλλη πλευρά, οι μέθοδοι δειγματοληψίας φαίνεται να είναι καλύτεροι από τον κλασικό SVM αλγόριθμο. Ειδικά ένας συνδυασμός της μεθόδου SMOTE-SVM με την τεχνική της υποδειγματοληψίας αποδεικνύει την καλύτερη απόδοση λαμβάνοντας υπόψη όχι μόνο το γεωμετρικό μέσο και τον υπολογιστικό χρόνο, αλλά και τα προβλήματα υπερπροσαρμογής που έχουν δημιουργηθεί χρησιμοποιώντας τις άλλες μεθόδους.

Αυτή η εργασία παρουσιάζει μια συγκριτική ανάλυση των διαφορετικών στρατηγικών SVM σε πραγματικά ιατρικά δεδομένα. Η αξιολόγηση της αξιοπιστίας των αλγορίθμων ταξινόμησης είναι απαραίτητη για τη διασφάλιση της ποιότητας των δεδομένων. Χρησιμοποιήσαμε λοιπόν το γεωμετρικό μέσο και την περιοχή κάτω από την καμπύλη ROC, που και τα δύο λαμβάνονται από την ευαισθησία και ειδικότητα, για τη σύγκριση των αλγορίθμων προκειμένου να παρέχουμε χρήσιμα αποτελέσματα σχετικά με την απόδοση των ταξινομητών. Σημειώνουμε ότι αυτές οι δύο μετρήσεις έδωσαν σχεδόν όμοια αποτελέσματα. Για το λόγο αυτό κάναμε μερικές συγκρίσεις μόνο για το γεωμετρικό μέσο. Είναι προφανές ότι η προσπάθεια της φροντίδας της υγείας για την πρόληψη του θανάτου των ασθενών είναι ένα τεράστιο πρόβλημα που ανακύπτει, αναγκάζοντας τους ερευνητές να είναι πιο προσεκτικοί στις έρευνές τους. Η ευαισθησία και η ειδικότητα μετράνε την ικανότητά του προγνωστικού μοντέλου για την αναγνώριση των ασθενών μιας συγκεκριμένης ομάδας (επιζώντες ή μη επιζώντες). Η αξία αυτής της συγκριτικής μελέτης είναι η ικανότητα να υπολογίζουμε τα σφάλματα Τύπου I και Τύπου II. Οι μέθοδοι μάθησης ευαίσθητου κόστους αλλά και οι μέθοδοι

προεπεξεργασίας των δεδομένων μας παρέχουν μικρότερο σφάλμα Τύπου ΙΙ και κατά συνέπεια μεγαλύτερη ευαισθησία σε σύγκριση με το CSVM. Το θέμα αυτό έχει μεγάλη σημασία για την ιατρική διάγνωση λόγω του γεγονότος ότι η παρουσιαζόμενη μεθοδολογία μας δίνει τη δυνατότητα να αναγνωρίσουμε τους ασθενείς που πρόκειται να πεθάνουν και αυτοί χρίζονται ιδιαίτερης μεταχείρησης. Με τον τρόπο αυτό, θα μπορούσαν να αποφευχθούν πολλοί θάνατοι. Αυτή η μέθοδος μπορεί να παρέχει κάποιες κατευθυντήριες γραμμές για τη βελτίωση της ποιότητας της θεραπείας και, ως εκ τούτου επιβίωσης ενός ασθενούς μέσω της βέλτιστης διαχείρισης της κατάστασης υγείας του. Μολονότι, οι Parpoula et αl. (2013) έχουν ήδη ασχοληθεί με την ανάλυση του συγκεκριμένου συνόλου δεδομένων η μελέτη τους εστιάζεται στη σύγκριση των διαφόρων τεχνικών εξόρυξης δεδομένων, συμπεριλαμβανομένου και του κλασικού SVM. Το κίνητρό μας για την εκπόνηση της μελέτης αυτής είναι διαφορετικό, διότι αυτό που θέλουμε να πετύχουμε είναι η ισορροπία ανάμεσα στην ευαισθησία και την ειδικότητα που θα μας επιτρέψουν την επιτυχία της πρόγνωσης ενός θανάτου. Η αποτελεσματικότητα της εξεταζόμενης προσέγγισης είναι προφανής από τα πειραματικά αποτελέσματα (Κεφάλαιο 5.2).

Ελπίζουμε ότι και αυτή η μελέτη θα πείσει τους πειραματιστές να χρησιμοποιήσουν όχι μόνο τις τυπικές τεχνικές SVM, αλλά και αναδιατυπώσεις τους για την εξαγωγή χρήσιμων σχεδίων, στην περίπτωση που ασχολούμαστε με μη ισορροπημένες βάσεις ιατρικών δεδομένων. Οι μηχανές διανυσματικής υποστήριξης είναι ένα ισχυρό εργαλείο πρόβλεψης και η χρήση των ταξινομητών SVM ως μια εναλλακτική μέθοδο για την υποστήριξη των ιατρικών αποφάσεων αποτελεί ένα από τα πιο προσοδοφόρα θέματα για περαιτέρω έρευνα.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Akbani, R., Kwek, S., Japkowicz, N., (2004). Applying support vector machines to imbalanced datasets, in *Proceedings of the 15th European Conference on Machine Learning*, pp. 39-50.

[2] Alaiz-Rodriguez, R, Japkowicz, N., (2008). Assessing the impact of changing environments on classifier performance. In *Proceedings of the 21st Canadian Conference on Advances in Artificial Intelligence (CCAI '08)*, Springer-Verlag, Berlin, Heidelberg, pp. 13‑24.

[3] Alberto Fernandez, Maria Jose del Jesus, Francisco Herrer, (2009). Hierarchical fuzzy rule based classification systems with genetic rule selection for imbalanced data-sets, International Journal of Approximate Reasoning 50, 561–577.

[4] Bache, K. & Lichman, M., (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[5] Barandela, R., Sanchez, J.S., Garcia, V., & Rangel, E., (2003). Strategies for learning in class imbalance problem. *Pattern Recognition* 36, 849-851.

[6] Batuwita, R., Palade, V., (2010). Efficient resampling methods for training support vector machines with imbalanced datasets. In *Proceedings of the International Joint Conference on Neural Networks*, pp. 1-8.

[7] Borrajo, L., Romero, R., Iglesias E. L., Redondo Marey, C. M., (2011). Improving imbalanced scientific text classification using sampling strategies and dictionaries. *Journal of Integrative Bioinformatics*, 8(3):176.

[8] Boser, B., Guyon, I., and Vapnik, V., (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*. pp. 144-152, ACM Press.

[9] Bradley, P. and Mangasarian, O.L., (1998). Feature selection via concave minimization and support vector machines. In Proceedings of the Fifteenth International Conference (ICML), pp. 82-90.

[10] Breiman, L., Friedman, J., Olshen, R., & Stone, C., (1984). *Classification and regression trees*, Wadsworth.

[11] Buckland, M. and Gey, F., (1994). The Relationship Between Recall and Precision. *Journal of the American Society for Information Science*, 45(1): 12-19.

[12] Burges, C., (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121-167.

[13] Chan, P. & Stolfo, S., (2001). Toward scalable learning with non-uniform class and cost distribution: a case study in credit card fraud detection. *In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*.

[14] Chang C.-C. and Lin, C.-J. (2011). LIBSVM: A library for support vector machines," ACM Transactions on Intelligent Systems and Technology, vol. 2, pp. 27:1-27:27.

[15] Chawla, N. V., (2010). Data mining for imbalanced datasets: An overview. *Data mining and knowledge discovery Handbook*, Springer.

[16] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, Vol 16, p.p. 321-357.

[17] Chawla, N.V., Japkowicz, N., Kolcz, A., (2004). Editorial: special issue on learning from imbalanced data sets. SIGKDD Explorations *Special Issue on Learning from Imbalanced Datasets 6* (1):1-6.

[18] Chawla, N., Lazarevic, A., Hall, L., & Bowyer, K. (2003). SMOTEBoost: Improving prediction of the minority class in boosting. *In Proceedings of the Seventh European Conference on Principles and Practice of Knowledge Discovery in Databases*.

[19] Chen, J., Casique, M. and Karakoy, M. (2004). Classification of lung data by sampling and support vector machine," in Proceedings of the 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, vol. 2, pp. 3194-3197.

[20] Chen, S., He, H., Garcia, E.A. (2010). Ramoboost: Ranked minority oversampling in boosting. *IEEE Transactions on Neural Networks* 21 (10) 1624‑1642.

[21] Choi, J. M., (2010). A Selective Sampling Method for Imbalanced Data Learning on Support Vector Machines. *Iowa State University*, cjm7331@gmail.com.

[22] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, vol.20, no. 3, pp. 273-297.

[23] Cristianini, N., Kandola, J., Elissee, A., Shawe-Taylor, and J., (2002). On kernel-target alignment. *In Advances in Neural Information Processing Systems* 14, pp. 367-373, MIT Press.

[24] Cristianinio N., and Shawe-Taylor, J. (2000). *An introduction to support Vector Machines: and other kernel-based learning methods*. Cambridge University Press.

[25] Diao, L., Yang, C., Wang, H. (2012). Training SVM email classifiers using very large imbalanced dataset. *Journal of Experimental & Theoretical Artificial Intelligence*, Vol. 24, No. 2, 193‑210.

[26] Domingos, P. (1999). Metacost: A General Method for Making Classifiers Cost-sensitive. In *Proceedings of the Fifrh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 155-164, San Diego, CA. ACM Press.

[27] Drummond, C., Holte, R. (2003). C4.5, Class Imbalance and Cost Sensitivity: Why Undersampling beats Over-sampling. *ICML 2003 Workshop on Learning from Imbalanced Data Sets II*, Washington, DC.

[28] Drummond, C., Holte, R.C., (2006). Cost curves: An improved method for visualizing classifier performance, *Mach Learn*, 65:95‑130, DOI 10.1007/s10994-006-8199-5.

[29] Drosou, K, (2013). Statistical Methods for the Analysis of High Dimensional Data. National Technical University of Athens. Bachelor Thesis.

[30] Drosou, K., Georgiou, S., Koukouvinos, C., S. Stylianou, S., (2014). Support Vector Machines classification on class imbalanced data: a case study with real medical data, *Journal of Data Science* (accepted for publication).

[31] Elkan, C. (2001). The Foundations of Cost-Sensitive Learning. In *International Joint Conference on Artificial Intelligence*, pp. 973-978.

[32] Ertekin, S., Huang, J., Bottou, L., Giles, C., (2007). Learning on the border: active learning in imbalanced data classification. In: *Proceedings of the sixteenth ACM conference on information and knowledge management*, pp.127-136.

[33] Estabrooks, A., Jo, T. & Japkowicz, N. (2004). A Multiple Resampling Method for Learning from Imbalanced Data Sets. *Computational Intelligence*. Vol. 20, pp 18-36.

[34] Fan, W. Stolfo, S., Zhang, J., & Chan, P.(1999). AdaCost: misclassification cost-sensitive boosting. In *Proceedings of the Sixteenth International Conference on Maching Learning*.

[35] Fernándeza, A., Garcíaa, S., Jesusb, M., Herreraa, F. (2008). A Study of The Behaviour of Linguistic Fuzzy Rule Based Classification Systems in the Framework of Imbalanced Data-Sets. *Fuzzy Sets and Systems*, vol: 159, 2378-2398.

[36] Freund, Y. & Schapire, E.(1997). A decision-theoretic generation of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1) pp. 119-139.

[37] Fu, Y., Ruixiang, S., Yang, Q., Simin, H., Wang, C., Wang, H., Shan, S., Liu, J. and Gao, W. (2004). A block-based support vector machine approach to the protein homology prediction task in kdd cup 2004. SIGKDD Exploration Newsletters, vol. 6, pp. 120-124.

[38] Fung, G., M., Mangasarian, O. L., (2001a). Proximal support vector machine classifiers. In F. Provost and R. Srikant, editors, *Proceedings KDD-2001: Knowledge Discovery and Data Mining, San Francisco, CA*, pages 77-86, New York, 2001. Association for Computing Machinery. ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/01-02.ps.

[39] Garcia, V., Sanchez, J. S., Mollineda, R. A., Alejo, R., Sotoca, J. M., (2007). The class imbalance problem in pattern classification and learning, Tamida 2007, Saragossa, Spain, pp. 283-291.

[40] Ghazikhani, A., Monsefi, R., Yazdi, H. S. (2013). Ensemble of online neural networks for nonstationary and imbalanced data streams, *Neurocomputing* 122, 535‑544.

[41] Cristianini, N., Kandola, J. Elisseeff, A., Shawe-Taylor, J. (2002). On kernel-target alignment. In *Advances in Neural Information Processing Systems 14*, pp. 367-373, MIT Press.

[42] Han, H., Wang, W., & Mao, B., (2005). Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. In *Proceedings of the International Conference on Intelligent Computing 2005*, Part I, LNCS.

[43] Hand, D.J., Till, R.J. (2001). A simple generalisation of the area under the ROC curve for multiple class classification problems- *Machine Learning*, - Springer.

[44] Hanley J.A., McNeil B.J. (1982). The meaning and use of the area under ROC. *Radiology* 143:29-36

[45] Hart, P., (1968). The condensed nearest neighbor rule. *IEEE Transactions on Information Theory*, IT-14, pp. 515-516.

[46] Hastie, T., Tibshirani, R., Friedman, J., (2001). *The elements of statistical learning*, Springer Series in Statistics, Springer-Verlag, New York. Data mining, Inference and Prediction.

[47] He, H., Bai, Y., Garcia, E.A., & Li, S., (2008). ADASYN: Adaptive Synthetic sampling approach for imbalanced learning. *Proceedings of International Joint Conference on Neural Networks*, pp. 1322-1328.

[48] He, H., Garcia, E. A., (2009). Learning from Imbalanced Data. In *IEEE Transactions on Knowledge and Data Engineering*, Vol. 21, No. 9, pp 1263 – 1284

[49] He H, Shen X (2007) A ranked subspace learning method for gene expression data classification. In: *Proceedings of international conference artificial intelligence*, pp 358–364.

[50] Hong, X. (2007). A Kernel-Based Two-Class Classifier for Imbalanced Data Set, *IEEE Transactions on Neural Networks*, vol. 18, no. 1.

[51] Hong, X., Chen, S., Harris, C. (2007). A kernel-based two-class classifier for imbalanced data sets. *IEEE Transactions on Neural Networks*, vol. 18, no. 1, pp. 28-41.

[52] Huang, S. & Lee, Y., (2004). Reduced support vector machines: a statistical theory. Technical report, Institute of Statistical Science, *Academia Sinica*, Taiwan.

[53] Japkowicz, N., (2000). The Class Imbalance Problem: Significance and Strategies. In *Proceedings of the 2000 International Conference on Artificial Intelligence*: Special Track on Inductive Learning, Las Vegas, Nevada.

[54] Japkowicz, N., (2000). Learning from Imbalanced Data Sets: a Comparison of Various Strategies. *AAAI Workshop on Learning from Imbalanced Data Sets*, *AAAI Press*, Menlo Park, CA.

[55] Jo, T., Japkowicz. N., (2004). Class imbalances versus small disjuncts. *ACM SIGKDD Explorations Newsletter*, Vol. 6, Issue 1, pages 40‑49.

[56] Joshi, M., Kumur, V., Agarwal, R. (2001). Evaluating boosting algorithms to classify rare cases: comparison and improvements. *In Proceedings of the First IEEE International Conference on Data Mining*.

[57] Kang, P. and Cho, S. (2006). Eus svms: ensemble of under-sampled svms for data imbalance problems. In *Proceedings of the 13th international conference on Neural Information Processing*, pp. 837-846, Springer-Verlag, 2006.

[58] Kandola, J., Shawe-Taylor, J., (2003). Refining kernels for regression and uneven classification problems. In *Proceedings of International Conference on Artificial Intelligence and Statistics.*

[59] Khoshgoftaar, T.M, Van Hulse, J., Napolitano, A., (2011). Comparing boosting and bagging techniques with noisy and imbalanced data" , IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans 41 (3), 552‑568.

[60] Kotsiantis, S., Kanellopoulos, D., Pintelas, P. (2006). Handling unbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering* , 30 (1), pp. 25-36

[61] Kowalczyk A. and Raskutti, B. (2002). One class svm for yeast regulation prediction. *SIGKDD Exploration Newsletters*, vol. 4, no. 2, pp. 99-100.

[62] Krawczyk, B., Schaefer, G., Wozniak, M., (2012). Breast thermogram analysis using a cost-sensitive multiple classifier system. In: *Proceedings of the IEEE-EMBS International Conference on Biomedical and Health Informatics* (BHI 2012), pp. 507‑510.

[63] Kubat, M. & Martin, S., (1997). Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. *Proceedings of the 14th International Conference on Machine Learning*.

[64] Kubat, M., Holte, R., Matwin, S., (1998). Machine learning for the detection of oil spills in satellite radar images. *Machine Learning*, 30, 195-215

[65] Laurikkala, J., (2001). Improving identification of difficult small classes by balancing class distribution. *Proceedings of AI in Medicine in Europe: Artificial Intelligence Medicine*, pp. 63-66.

[66] Lebrun, G., Lezoray, O., Charrier, C. and Cardot, H., (2006). A New Model Selection Method for SVM, E. Corchado et al. (Eds.): IDEAL 2006, LNCS 4224, pp. 99–107, Springer-Verlag Berlin Heidelberg 2006.

[67] Leng, M., Cheng, J., Wang, J., Zhang, Z., Zhou, H., Chen, X., (2013). Active Semi supervised Clustering Algorithm with Label Propagation for Imbalanced and Multidensity Datasets. Hindawi Publishing Corporation, Mathematical Problems in Engineering, Article ID 641927, 10 pages, http://dx.doi.org/10.1155/2013/641927.

[68] Lessmann S. (2004). Solving imbalanced classification problems with support vector machines. In Proceedings of the International Conference on Artificial Intelligence, pp. 214-220.

[69] Lin, Z., Hao, Z., Yang, X., Liu, X. (2009). Several svm ensemble methods integrated with under-sampling for imbalanced data learning. In *Proceedings of the 5th International Conference on Advanced Data Mining and Applications*, pp. 536-544, Springer-Verlag.

[70] Liu, Y., An, A., & Huang, X., (2006a). Boosting prediction accuracy on imbalanced datasets with svm ensembles. In the *Proceedings of the 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Singapore, 107-118.

[71] Liu, W., Chawla, S. (2011). Class Confidence Weighted kNN Algorithms for Imbalanced Data Sets. *Advances in Knowledge Discovery and Data Mining, Lecture Notes in Computer Science,* Volume 6635, 2011, pp 345-356.

[72] Liu, W., Chawla, S., Cieslak, D.A., Chawla, N.V. (2010). A Robust Decision Tree Algorithm for Imbalanced Data Sets. SDM, 2010 – SIAM, pp 766-777.

[73] Liu Q., He Q., Zhongzhi S., (2007). Incremental Nonlinear Proximal Support Vector Machine, *ISNN, Part III, LNCS 4493*, 336-341.

[74] Liu, X., Wu, J., Zhou, Z., (2006b). Exploratory under-sampling for class imbalance learning. In Proceedings of the 6th IEEE International Conference on Data Mining. pp. 965-969.

[75] Lopez, V., Fernandez, A., Garcia, S., Palade, V., Herrera, F., (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences* 250, p.p. 113–141.

[76] Maheshwari, S., Agrawal, J., Sharma, S., (2011). A New approach for Classification of Highly Imbalanced Datasets using Evolutionary Algorithms. In *International Journal of Scientific & Engineering Research*, Volume 2, Issue 7, ISSN 2229-5518.

[77] Margineantu, D. (2002). Class probability estimation and cost-sensitive classification decisions. *Machine Learning:* ECML 2002, 167-185.

[78] Mazurowski, M. A., Habas, P. A., Zurada, J. M., Lo, J. Y. Baker, J. A., Tourassi, G. D. (2008). Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance, *Neural Networks* 21, 427‑436.

[79] McCarthy, K., Zabar, B., & Weiss, G., (2005). Does Cost-sensitive learning Beat Sampling for Classifying rare classes?. In *Proceedings of International Workshop Utility-based Data Mining*, pp. 69-77.

[80] Mena, L., Gonzalez, J.A. (2006). Machine Learning for Imbalanced Datasets: Application in Medical Diagnostic, *Proceedings of the 19th International FLAIRS Conference (FLAIRS-2006)*, Melbourne Beach, Florida, May 11-13.

[81] Moreno-Torres, J.G, Herrera. F., (2010). A preliminary study on overlapping and data fracture in imbalanced domains by means of genetic programming-based feature extraction. In *Proceedings of the 10th International Conference on Intelligent Systems Design and Applications* (ISDA' 10), 2010, pp. 501-506.

[82] Nallapati, R., (2004). Discriminative Models for Information Retrieval. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, Pages 64-71.

[83] Nguyen, H., Franke, K., Petrovic, S. (2010). Towards a Generic Feature-Selection Measure for Intrusion Detection. In *Proc. International Conference on Pattern Recognition (ICPR)*, Istanbul, Turkey.

[84] Nguyen, H. T., Franke, K., Petrovic, S. (2011). On General Definition of L1-norm Support-Vector Machine for Feature Selection. In *Proceedings of the International Journal of Machine Learning and Computing*, ISSN: 2010-3700.

[85] Ou, Y.Y., Chen, C.Y., Hwang, S.C., Oyang, Y.J. (2003). Expediting model selection for support vector machines based on data reduction. In: *Proc. IEEE International Conference on Systems, Man and Cybernetics (SMC2003)*, pp. 786–791.

[86] Parpoula, C., K. Drosou K., Koukouvinos, C. (2013). Large-Scale Statistical Modelling via Machine Learning Classifiers, *Journal of Statistics Applications & Probability* 2, No. 3, 1-20.

[87] Pearson, R., Goney, G., Shwaber, J., (2003). Imbalanced Clustering for Microarray Time-Series. In: *Proceedings of international conference Machine Learning*, Workshop Learning from Imbalanced Data Sets II.

[88] Prez-Godoy, M. D., Fernndez, A., Rivera, A. J., Jesus, M. J., (2010). Analysis of an evolutionary RBFN design algorithm, CO2RBFN, for imbalanced data sets, *Pattern Recognition Letters*, vol. 31, no. 15, pp. 2375-2388, 2010.

[89] Qin, A., Suganthan, P. (2004). Kernel neural gas algorithms with application to cluster analysis. In *Proceedings of the 17th International Conference on Pattern Recognition*, pp. 617-620, IEEE Computer Society.

[90] Raskutti, B. & Kowalczyk, A., (2004). Extreme Re-balancing for SVMs: a case study. *SIGKDD Explorations*, 6(1), pp. 60-69.

[91] Raudys, S.J, Jain, A.K., (1991). Small sample size effects in statistical pattern recognition: recommendations for practitioners, IEEE Transactions on Pattern Analysis and Machine Intelligence 13 (3), 252‑264.

[92] Rosset, S., Zhu, J. and Hastie, T. (2004). Margin maximizing loss functions, in S. Thrun, L. Saul and B. Sch¨olkopf (eds). *Advances in Neural Information Processing Systems* 16, MIT Press, Cambridge, MA.

[93] Scholkopf, B. and Smola, A. (2001). Learning with Kernels: Support Vector Machines. *Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press, 2001.

[94] Sun, Y, Kamel, M., Wong, A., Wang, Y. (2007). Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition*, 40, 3358‑3378, 2007.

[95] Tashk, A. and Faez, K. (2007). ”Boosted bayesian kernel classifier method for face detection”. *In Proceedings of the Third International Conference on Natural Computation*, pp. 533-537, IEEE Computer Society.

[96] Thanathamathee, P., Lursinsap, C., (2013). Handling imbalanced data sets with synthetic boundary data generation using bootstrap re-sampling and AdaBoost techniques, Pattern Recognition Letters, vol. 34, pp. 1339-1347.

[97] Tikhonov, A. N., Arsenin, V. Y., (1977). *Solutions of III-Posed Problems*. John Wiley & Sons, New York.

[98] Tomek, I., (1976). Two Modifications of CNN. *IEEE Transactions on Systems Man and Communications SMC-6*, pp. 760-772.

[99] Turney, P. (2000). Types of Cost in Inductive Concept Learning. In *Workshop on Cost-Sensitive Learning at the Seventeenth International Conference on Machine Learning*, pages 15-2 1, Stanford, CA.

[100] Vapnik, V.N., (2000). The Nature of Statistical Learning Theory. 2nd end. Springer-Verlag, Berlin Heidelberg New York.

[101] Veropoulos, K., Campbell, C., Cristianini, N., (1999). Controlling the sensitivity of support vector machines. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp.55-60.

[102] Wang B. and Japkowicz N. (2010). Boosting support vector machines for imbalanced data sets. *Knowledge and Information Systems*, vol. 25, pp. 1-20.

[103] Wasikowski. M, Chen X.-W, (2010). Combating the small sample class imbalance problem using feature selection. IEEE Transactions on Knowledge and Data Engineering 22 (10), 1388‑1400.

[104] Weiss, G., (2003). The Effect of Small Disjuncts and Class Distribution on Decision Tree Learning, *Ph.D. Dissertation, Department of Computer Science, Rutgers University*, New Brunswick, New Jersey.

[105] Weiss, G.M., (2004). Mining with rarity: a unifying framework. *ACM SIGKDD Explorations Newsletter*, Vol. 6, Issue 1, pages 7-19.

[106] Wilson, D., (1972). Asymptotic Properties of Nearest Neighbor Rules using Edited Data. *IEEE Transactions on Systems, Man and Cybernetics, Vol. 2*, pp 408-420.

[107] Wu, G. & Chang, E., (2003a). "Class-Boundary Alignment for Imbalanced Dataset Learning". *In ICML 2003 Workshop on Learning from Imbalanced Data Sets II,* Washington, DC.

[108] Wu, G. and Chang, E. (2003b). "Adaptive feature-space conformal transformation for imbalanced-data learning". *In Proceedings of the 20th International Conference on Machine Learning*, pp. 816-823.

[109] Wu, G., Chang, E. (2005). Kba: Kernel boundary alignment considering imbalanced data distribution. *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 6, pp. 786-795

[110] Xiao-yan, T., Hong-bing J., (2007). A Modified PSVM and its Application to Unbalanced Data Classification: *Third International Conference on Natural Computation* (ICNC 2007), Pages:488 – 490.

[111] Xu, X., Ye, Q., Ye, N., Wu, B., (2008). Nonlinear Proximal Support Vector Classifiers Aiming At Large Scale Classification Problems. *Proceedings of the 11th Joint Information Sciences.*

[112] Yang, C.-Y., Yang, J.-S. and Wang, J.-J. (2009). Margin calibration in svm class imbalanced learning. *Neurocomputing*, vol. 73, no. 1-3, pp. 397-411.

[113] Yang, Y., Pedersen, J. (1997). A comparative study on feature selection in text categorization, *The Fourteenth International Conference on Machine Learning*, pages 412-420.

[114] Yang Yong (2012). The Research of Imbalanced Data Set of Sample Sampling Method Based on K-Means Cluster and Genetic Algorithm, *Energy Procedia* 17, pages 164 ‑ 170.

[115] Yao, D., Yang, J., Zhan, X., (2013). An Improved Random Forest Algorithm for Class-Imbalanced Data Classification and its Application in PAD Risk Factors Analysis, The Open Electrical & Electronic Engineering Journal, 7, (Supple 1: M7) 62-70.

[116] Yoon, K., Kwek, S. (2005). An Unsupervised Learning Approach to Resolving the Data Imbalanced Issue in Supervised Learning Problems in Functional Genomics. *Proceedings of the 5th International Conference on Hybrid Intelligent Systems*, pp. 303-308, Rio de Janeiro, Brazil.

[117] Yu, X.-P. and Yu, X.-G., (2007). "Novel text classification based on k-nearest neighbor". In Proceedings of the International Conference on Machine Learning and Cybernetics, pp. 3425-3430.

[118] Yuan, J., Li, J., Zhang, B. (2006). Learning concepts from large scale imbalanced data sets using support cluster machines. In *Proceedings of the 14th annual ACM international conference on Multimedia*, pp. 441-450, ACM.

[119] Zhang, Y., Zhang, D., Mi, G., Ma, D., Li, G., Guo, Y., Li, M., Zhu, M., (2012). Using ensemble methods to deal with imbalanced data in predicting protein–protein interactions. *Computational Biology and Chemistry* 36, 36–41.

[120] Zhang,Y., Zhang, Y., Wang, D., A split and boost active learning method for classimbalance data sets. School of Computer and Information Technology, Liaoning Normal University, Dalian 116081, China.

[121] Zhang, Y., Wang, D. (2013). A Cost-Sensitive Ensemble Method for Class-Imbalanced Datasets. Hindawi Publishing Corporation, *Abstract and Applied Analysis*, Volume 2013, Article ID 196256, 6 pages.

[122] Zheng, Z., Wu, X., Srihari, R., (2004). Feature selection for text categorization on imbalanced data. *SIGKDD Explorations*, 6(1).

[123] Zhu, Y., Li, C., and Zhang, Y., (2004). A Practical Parameters Selection Method for SVM, ISNN 2004, LNCS 3173, pp. 518–523.

ΠΑΡΑΡΤΗΜΑ Α

Πίνακας Α-1: Περιγραφή του ιατρικού (medical) συνόλου δεδομένων

*Μεταβλητή Απόκρισης-Δίτιμη-**output***	*Y*	**0: survival, 1: death**
*Συνεχείς Μεταβλητές-**inputs***	x_1	*βάρος (kg)*
	x_2	*Ηλικία (χρόνια)*
	x_3	*Σκορ κώματος Γλασκώβης (GCS)*[12]
	x_4	*Σφυγμός (N/min)*
	x_6	*συστολική αρτηριακή πίεση του αίματος (mmHg)*
	x_7	*διαστολική αρτηριακή πίεση του αίματος (mmHg)*
	x_8	*αιματοκρίτης(Ht), %*
	x_9	*αιμοσφαιρίνη (Hb), g/dl*
	x_{11}	*Αριθμός λευκών αιμοσφαιρίων (/ml)*
	x_{15}	*γλυκόζη, mg%*
	x_{16}	*Κρεατινίνη mg %*
	x_{18}	*amylase, score*
	x_{20}	*Injury Severity Score, score*
	x_{21}	*Revised Trauma Score, score*

[12] Μια κλίμακα που χρησιμοποιείται για τον προσδιορισμό του επιπέδου συνείδησης του ασθενή. Αποτελεί μια βαθμολόγηση από το 3 έως το 15 της ικανότητας του ασθενή να ανοίξει τους οφθαλμούς του, να αντιδράσει λεκτικά και να κινηθεί φυσιολογικά. Η κλίμακα GCS χρησιμοποιείται κυρίως κατά τη φυσική εξέταση ασθενών με τραύμα ή αγγειακό εγκεφαλικό επεισόδιο. Η επαναλαμβανόμενη αξιολόγηση της κλίμακας καθορίζει αν η εγκεφαλική λειτουργία του ασθενή βελτιώνεται ή επιδεινώνεται.

Πίνακας Α-2: Συνέχεια Πίνακα Α-1

Κατηγορικές μεταβλητές-**inputs**	x_{19}	*evaluation of disability* *(0 = expected permanent big, 1 = expected permanent small, 2 = expected impermanent big, 3 = expected impermanent small, 4 = recovery)*
	x_{23}	*cause of injury* *(0 = fall, 1 = trochee accident, 2 = athletic, 3 = industrial, 4 = crime, 5 = other)*
	x_{24}	*means of transportation* *(0 = airplane, 1 = ambulance, 2 = car, 4 = on foot)*
	x_{25}	*Ambulance (0 = no, 1 = yes)*
	x_{26}	*hospital of records*
	x_{27}	*substructure of hospital* *(0 = orthopaedic, 1 = CT, 2 = vascular surgeon, 3 = neurosurgeon, 4 = Intensive Care Unit)*
	x_{28}	*comorbidities (0 = no, 1 = yes)*
	x_{35}	*doctor's speciality* *(0 = angiochirurgeon, 1 = non specialist, 2 = general doctor 3 = general surgeon, 4 =jawbonesurgeon, 5 = gynaecologist, 6 = thoraxsurgeon, 7 = neurosurgeon, 8 = orthopaedic, 9 = urologist, 10 = paediatrician, 11 =children surgeon, 12 = plastic surgeon)*
	x_{36}	*major doctor (0 = no, 1 = yes)*
	x_{41}	*dysphoria (0 = no, 1 = yes)*
	x_{52}	*Collar (0 = no, 1 = yes)*
	x_{55}	*immobility of limbs (0 = no, 1 = yes)*
	x_{56}	*fluids (0 = no, 1 = yes)*
	x_{64}	*Radiograph E.R. (0 = no, 1 = yes)*
	x_{66}	*US (0 = no, 1 = yes)*
	x_{67}	*urea test (0 = no, 1 = yes)*
	x_{71}	*destination after the emergency room* *(0 = other hospital, 1 = clinic, 2 = unit of high care, 3 = intensive care unit I.C.U, 4 = mortuary, 5 = operating room)*
	x_{72}	*surgical intervention (0 = no, 1 = yes)*

	x_{86}	*arrival at emergency room* *(0 = 00:00-04:00, 1 = 04:01-08:00, 2 = 08:01-12:00, 3 = 12:01-16:00, 4 =16:01-18:00, 5 = 18:01-20:00, 6 = 20:01-24:00)*
	x_{87}	*exit from emergency room (0 = 00:00-04:00, 1 = 04:01-08:00, 2 = 08:01-12:00, 3 = 12:01-16:00, 4 =16:01-18:00, 5 = 18:01-20:00, 6 = 20:01-24:00)*
	x_{101}	*head injury (0 = none, 1 = AIS ≤ 2, 2= AIS > 2)*
	x_{102}	*face injury (0 = none, 1 = AIS ≤ 2, 2= AIS > 2)*
	x_{104}	*breast injury (0 = none, 1 = AIS ≤ 2, 2= AIS > 2)*
	x_{106}	*spinal column injury (0 = none, 1 = AIS ≤ 2, 2= AIS > 2)*
	x_{107}	*upper limbs injury (0 = none, 1 = AIS ≤ 2, 2= AIS > 2)*
	x_{108}	*lower limbs injury (0 = none, 1 = AIS ≤ 2, 2= AIS > 2)*

ΠΑΡΑΡΤΗΜΑ Β

Πίνακας Β-1: Περιγραφή του ιατρικού (medical) συνόλου δεδομένων μόνο για τις συνεχείς τυχαίες μεταβλητές

*Μεταβλητή Απόκρισης-Δίτιμη-**output***	Y	**0: survival, 1: death**
*Συνεχείς Μεταβλητές-**inputs***	x_1	*βάρος (kg)*
	x_2	*Ηλικία (χρόνια)*
	x_3	*Σκορ κώματος Γλασκώβης (GCS)*[13]
	x_4	*Σφυγμός (N/min)*
	x_6	*συστολική αρτηριακή πίεση του αίματος (mmHg)*
	x_7	*διαστολική αρτηριακή πίεση του αίματος (mmHg)*
	x_8	*αιματοκρίτης(Ht), %*
	x_9	*αιμοσφαιρίνη (Hb), g/dl*
	x_{11}	*Αριθμός λευκών αιμοσφαιρίων (/ml)*
	x_{15}	*γλυκόζη, mg%*
	x_{16}	*Κρεατινίνη mg %*
	x_{18}	*amylase, score*
	x_{20}	*Injury Severity Score, score*
	x_{21}	*Revised Trauma Score, score*

[13] Μια κλίμακα που χρησιμοποιείται για τον προσδιορισμό του επιπέδου συνείδησης του ασθενή. Αποτελεί μια βαθμολόγηση από το 3 έως το 15 της ικανότητας του ασθενή να ανοίξει τους οφθαλμούς του, να αντιδράσει λεκτικά και να κινηθεί φυσιολογικά. Η κλίμακα GCS χρησιμοποιείται κυρίως κατά τη φυσική εξέταση ασθενών με τραύμα ή αγγειακό εγκεφαλικό επεισόδιο. Η επαναλαμβανόμενη αξιολόγηση της κλίμακας καθορίζει αν η εγκεφαλική λειτουργία του ασθενή βελτιώνεται ή επιδεινώνεται.

Pima Indians Diabetes Data Set (UCI Repository)

Πίνακας Β-2: Περιγραφή των δεδομένων για το διαβήτη στους Ινδιάνους Pima

Number of Instances: **768**		Number of Attributes: **9**
*Μεταβλητή Απόκρισης-Δίτιμη-**output***	Y	**-1:** "tested negative for diabetes", **1:** "tested positive for diabetes"
*Συνεχείς Μεταβλητές-**inputs***	x_1	1. Number of times pregnant
	x_2	2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
	x_3	3. Diastolic blood pressure (mm Hg)
	x_4	4. Triceps skin fold thickness (mm)
	x_5	5. 2-Hour serum insulin (mu U/ml)
	x_6	6. Body mass index (weight in kg/(height in m)^2)
	x_7	7. Diabetes pedigree function
	x_8	8. Age (years)
	x_9	9. Class variable (0 or 1)

Πίνακας Β-3: Brief statistical analysis

Attribute number:	Mean:	Standard Deviation:
1.	3.8	3.4
2.	120.9	32.0
3.	69.1	19.4
4.	20.5	16.0
5.	79.8	115.2
6.	32.0	7.9
7.	0.5	0.3
8.	33.2	11.8

Blood Transfusion (UCI Repository)

Πίνακας Β-4: Περιγραφή των δεδομένων για τη μετάγγιση αίματος (Blood Transfusion)

Number of Instances: **748**		Number of Attributes: **5**
*Μεταβλητή Απόκρισης-Δίτιμη-**output***	Y	**-1:** "not donated blood ", **1:** " donated blood"
*Συνεχείς Μεταβλητές-**inputs***	x_1	Recency quantitative Months Input 0.03 74.4 9.74 8.07
	x_2	Frequency quantitative Times Input 1 50 5.51 5.84
	x_3	Monetary quantitative c.c. blood Input 250 12500 1378.68 1459.83
	x_4	Time quantitative Months Input 2.27 98.3 34.42 24.32
	x_5	Recency quantitative Months Input 0.03 74.4 9.74 8.07

Attribute Characteristics: Real
Date Donated: 2008-10-03
Associated Tasks: Classification

Thyroid Disease (New Thyroid) data set

Πίνακας Β-5: Περιγραφή των δεδομένων για τη νόσο του θυρεοειδούς (Thyroid Disease data set)

Number of Instances: **215**		Number of Attributes: **5**
*Μεταβλητή Απόκρισης-Δίτιμη-**output***	Y	**-1:** "otherwise", **1:** "tested positive for hyperthyroidism "
*Συνεχείς Μεταβλητές-**inputs***	x_1	1) attribute T3resin integer [65, 144]
	x_2	2) attribute Thyroxin real [0.5, 25.3]
	x_3	3) attribute Triiodothyronine real [0.2, 10.0]
	x_4	4) attribute Thyroidstimulating real [0.1, 56.4]
	x_5	5) attribute TSH_value real [-0.7, 56.3]