ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

Σχολή Εφαρμοσμένων Μαθηματικών & Φυσικών Επιστημών

Διατμηματικό Μεταπτυχιακό Πρόγραμμα

«Μαθηματική Προτυποποίηση και Εφαρμογές στις Σύγχρονες Τεχνολογίες και την Οικονομία»

Διπλωματική εργασία:

Βέλτιστοι Σχεδιασμοί σε Βάσεις Μεγάλων Δεδομένων

Φοιτήτρια: Μπλετσογιάννη Ελευθερία

Επιβλέπων:Χρήστος Κουκουβίνος, Καθηγητής Τομέα Μαθηματικών Ε.Μ.Π

Αθήνα, Σεπτέμβριος 2016

ΠΡΟΛΟΓΟΣ

Η παρούσα διπλωματική εργασία με τίτλο «Βέλτιστοι Σχεδιασμοί σε Βάσεις Μεγάλων Δεδομένων» , εκπονήθηκε την περίοδο 2014-2016, με τη βοήθεια και τη συστηματική καθοδήγηση του επιβλέποντα Καθηγητή Χρήστου Κουκουβίνου. Σκοπός της πραγματοποίησής της ήταν η ολοκλήρωση των σπουδών μου στο Διατμηματικό Πρόγραμμα Μεταπτυχιακών Σπουδών με τίτλο «Μαθηματική Προτυποποίηση και Εφαρμογές στις Σύγχρονες Τεχνολογίες και την Οικονομία», με συντονίζουσα σχολή τη Σχολή Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών (ΣΕΜΦΕ) του Εθνικού Μετσόβιου Πολυτεχνείου. Αποτελεί μέρος μιας ευρύτερης έρευνας των βέλτιστων σχεδιασμών πάνω σε βάσεις υψηλών και πολύ υψηλών διαστάσεων. Η ολοκλήρωση της συγκεκριμένης εργασίας δεν θα ήταν εφικτή χωρίς τη συμβολή κάποιων προσώπων. Έτσι, από τη θέση αυτή, θα ήθελα να ευχαριστήσω τον κ. Χρήστο Κουκουβίνο, Καθηγητή του Τομέα Μαθηματικών του ΕΜΠ, χωρίς τον οποίο δεν θα βρισκόμουν σε αυτή τη θέση σήμερα. Εκτιμώντας τον ιδιαίτερα σαν καθηγητή και άνθρωπο, τον συμβουλεύομουν κάθε φορά που χρειαζόμουν καθοδήγηση στη διάρκεια των σπουδών μου. Η αφοσίωσή του στην επιστήμη του μου παρακίνησε το ενδιαφέρον κι έτσι αποφάσισα να ασχοληθώ με τον τομέα της Στατιστικής. Του εύχομαι να είναι πάντα καλά ώστε να εμπνέυσει κι άλλους φοιτητές με το ήθος και τον χαρακτήρα του. Θα ήθελα επίσης να ευχαριστήσω την υποψήφια διδάκτορα Κρυσταλλένια Δρόσου για τις συμβουλές της και τη βοήθειά της κατά τη διάρκεια εκπόνησης της εργασίας καθώς και όσους βοήθησαν με τον τρόπο τους χωρίς να το γνωρίζουν...

Τέλος, ίσως ένα «ευχαριστώ» δεν είναι αρκετό για να ανταποδώσω στην οικογένειά μου την υποστήριξη που μου δίνουν πάντα για να προχωράω μπροστά..

ΠΕΡΙΛΗΨΗ

Στην παρούσα εργασία γίνεται μελέτη των πειραμάτων που εκτείνονται σε μεγάλες βάσεις δεδομένων καθώς και αναζήτηση των βέλτιστων σχεδιασμών για την αντιμετώπιση τους.

Αρχικά, στο πρώτο κεφάλαιο της παρούσας εργασίας, γίνεται μια εισαγωγή στις μεγάλες βάσεις δεδομένων (Big Data). Αναλύεται η παρουσία και η επίδρασή τους στη σύγχρονη κοινωνία και κατ' επέκταση στην σύγχρονη επιστήμη της Στατιστικής. Μελετάται, επίσης, ο τρόπος επεξεργασίας μεγάλων βάσεων δεδομένων καθώς και η ερμηνεία των αποτελεσμάτων μέσω της μηχανικής μάθησης.

Στη συνέχεια, στο δεύτερο κεφάλαιο, γίνεται εκτενής μελέτη του Σχεδιασμού Πειραμάτων. Αναλύονται οι βασικές αρχές πειραματικού σχεδιασμού , οι βασικοί πειραματικοί στόχοι καθώς και οι Παραγοντικοί Σχεδιασμοί. Δίνεται ιδιαίτερη έμφαση στους D – Βέλτιστους σχεδιασμούς με μελέτη της προσέγγισής τους και ανάλυση των κριτηρίων για την επιλογή του D – βέλτιστου σχεδιασμού.

Το τρίτο κεφάλαιο αφιερώνεται στη μελέτη της σύνδεσης μεταξύ Επιλογής Μεταβλητών και της Ταξινόμησής τους. Μέσα σε αυτό το κεφάλαιο περιλαμβάνονται μέθοδοι επιλογής μεταβλητών και κριτήρια που οδηγούν στην επιλογή του καλύτερου υποσυνόλου μεταβλητών. Σημαντικό μέρος του κεφαλαίου αυτού περιλαμβάνει την μελέτη της Ταξινόμησης Μεταβλητών καθώς και την ανάλυση του Αλγόριθμου CATCH.

Στο τελευταίο κεφάλαιο, αναλύονται λεπτομερώς οι D – Βέλτιστοι Σχεδιασμοί με μελέτη προσομοίωσης και σύνοψη των αποτελεσμάτων αυτής.

ABSTRACT

The focus of the present study is the research of the experiments which are expanded on big databases and the quest of the optimal designs to handle them.

Firstly, in the first chapter of the present study, there is an introduction in big databases. There is an analysis of the presence and the effect of big data in modern society and in Science of Statistics. Furthermore, there is paid attention to the processing of big databases and even more to the results' interpretation via Machine Learning.

Subsequently, in the second chapter, there is an extensive study of experiments' design. This section of the study presents the basic principles, the main goals while designing an experiment and analyses the factorial designs. There is given special emphasis on D – optimal designs by analyzing their approach and the criteria for the selection of the D – optimal plan.

The third chapter deals with the correlation between Variable Selection and Classification. This section of the study includes methods for variable selection and criteria which lead to the selection of optimal variables' subset. A big part of the chapter deals with variable classification and the analysis of CATCH Algorithm.

Finally, the last pages include a research in D – optimal plans by analyzing a simulation study and summarizing the results.

Περιεχόμενα

Κεφάλαιο 1

Big Data

1.1 Εισαγωγή

Στη σημερινή εποχή δεχόμαστε καθημερινά καταιγισμό δεδομένων και πληροφοριών. Ως εκ τούτου, χρησιμοποιείται πλέον όλο και συχνότερα η φράση «Big Data» με την οποία γίνεται αναφορά σε μεγάλα δεδομένα. Η φράση «Big Data» σαν όρος, βρίσκεται ανάμεσα στις μεγαλύτερες τάσεις της τελευταίας δεκαετίας καθώς φαίνεται πως πάρα πολλοί τομείς της καθημερινότητας κι όχι μόνο χαρακτηρίζονται από υπεράριθμα δεδομένα.Ο όρος «Big Data» αναφέρεται σε βάσεις δεδομένων των οποίων το μέγεθος είναι πέραν των δυνατοτήτων των λογισμικών τυπικών βάσεων δεδομένων. Τα όργανα των συνήθων λογισμικών, αδυνατούν να συλλάβουν, να αποθηκεύσουν, να διαχειριστούν και να αναλύσουν βάσεις τέτοιων δεδομένων. Αυτός ο όρος είναι σκόπιμα υποκειμενικός και ενσωματώνει τον ορισμό που δηλώνει πόσο μεγάλο χρειάζεται να είναι ένα σύνολο δεδομένων προκειμένου να θεωρηθεί σαν σύνολο «Big Data». Με άλλα λόγια, ο όρος αυτός δεν προσδιορίζει τα «Big Data» χρησιμοποιώντας τη σύγκριση ενός συνόλου δεδομένων με ένα σύνολο συγκεκριμένου αριθμού terabytes (εκατοντάδες gigabytes). Υποθέτουμε ότι καθώς η τεχνολογία αναπτύσσεται, θα αυξάνεται επίσης και το μέγεθος των συνόλων δεδομένων που θα πληρούν τις προϋποθέσεις για να ανήκει στην κατηγορία των «Big Data». Θα πρέπει επίσης, να σημειώσουμε ότι ο ορισμός μπορεί να ποικίλλει κατά τομέα, εξαρτώμενος από τα διαθέσιμα είδη λογισμικών και τα μεγέθη των συνόλων δεδομένων που είναι συνήθη στην κάθε βιομηχανία. Λαμβάνοντας υπόψη όλα τα παραπάνω, σε πολλούς τομείς της σύγχρονης κοινωνίας, τα «Big Data» θα κυμαίνονται από μερικές δωδεκάδες terabytes μέχρι πολλαπλά petabytes. Καθώς, λοιπόν, περιστοιχιζόμαστε από βάσεις δεδομένων με εκατοντάδες πεδία, με εκατομμύρια δεδομένα και δισεκατομμύρια πληροφορίες, θεωρείται πλέον σίγουρο από τους ερευνητές πωςοι επιστήμες, οι επιχειρήσεις και η βιομηχανία θα υποβληθούν σε τεράστιες αλλαγές από την επιρροή των «Big Data». Τα δεδομένα θεωρούνται από τους ερευνητές, ένα πανίσχυρο ακατέργαστο υλικό το οποίο δύναται να επηρεάσει αποφάσεις που αφορούν κυβερνήσεις και επιχειρήσεις- αποφάσεις που παλαιότερα στηρίζονταν απλά σε εικασίες. Γίνεται, λοιπόν, φανερό πως η σωστή διαχείριση μεγάλου όγκου δεδομένων μπορεί να οδηγήσει σχεδόν κάθε όψη της μοντέρνας κοινωνίας, συμπεριλαμβανομένων των υπηρεσιών κινητής τηλεφωνίας, των πωλήσεων, των βιομηχανιών, των οικονομικών υπηρεσιών και των επιστημών.

1.2 Τα Big Data στη σύγχρονη κοινωνία

Τα «Big Data» παρουσιάζουν απαράμιλλες ευκαιρίες σε διάφορους τομείς της ζωής μας. Κάποιοι από τους δρόμους που μας ανοίγουν είναι: η επιτάχυνση των επιστημονικών ανακαλύψεων και καινοτομιών, η βελτίωση της υγείας και της ιατρικής περίθαλψης, η δημιουργία νέων τομέων μελέτης που μέχρι σήμερα δεν ήταν δυνατή, η βελτίωση λήψης αποφάσεων μέσω των προβλέψεων που προσφέρει η ανάλυση δεδομένων, η συνηδειτοποίηση της δυναμικής της ανθρώπινης συμπεριφοράς καθώς και η δυνατότητα επίδρασης στο εμπόριο μέσα στην παγκόσμια

οικονομία. Ας δούμε λίγο πιο αναλυτικά τους τομείς τους οποίους έχουν επηρεάσει σημαντικά τα «Big Data».

Η *αστρονομία* είναι μια φανταστική εφαρμογή των «Big Data» καθώς οδηγείται από την πρόοδο των αστρονομικών οργάνων. Κάθε pixel που λαμβάνεται από τα νέα αστρονομικά όργανα μπορεί να έχει μερικές εκατοντάδες χαρακτηριστικά τα οποία μπορούν να μεταφραστούν, πλέον πολύ γρήγορα, σε πρόβλημα κλίμακας petabyte (εκατοντάδες gigabyte). Αυτή η ραγδαία αύξηση δεδομένων δημιούργησε ένα νέο τομέα ο οποίος ονομάζεται «Astro-informatics». Ο νέος τομέας της «Αστρο–πληροφορικής » δημιουργεί συνεργασίες μεταξύ της επιστήμης των υπολογιστών, της στατιστικής και της αστρονομίας. Στο πεδίο της *αστρονομίας* έχει έλθει επανάσταση χάρη στο Sloan Digital Sky Survey [SDSS2008]. Πρόκειται για ένα τεράστιο πολύ-φίλτρο που απεικονίζει φασματοσκοπική έρευνα του σύμπαντος, χρησιμοποιώντας ένα οπτικό τηλεσκόπιο σε παρατηρητήριο στο Νέο Μεξικό. Η μεγάλη εξέλιξη στον τομέα της *αστρονομίας*γίνεται κατανοητή αν σκεφτούμε ότι πριν από το Sloan Digital Sky Survey μεγάλο κομμάτι της δουλειάς ενός αστρονόμου ήταν η συλλογή εικόνων από το σύμπαν. Από το 2000 που ξεκίνησε η συλλογή δεδομένων από το SDSS, οι απαραίτητες εικόνες βρίσκονται ήδη σε βάση δεδομένων. Ύστερα, το έργο που έχει να επιτελέσει ο αστρονόμος είναι να βρει ενδιαφέροντα αντικείμενα και φαινόμενα μέσα σ'αυτή τη βάση δεδομένων.

Στον τομέα της *βιολογίας*, υπάρχει πλέον, καθιερωμένη κατάθεση επιστημονικών δεδομένων σε δημόσια αποθήκη πληροφοριών. Επιπλέον, κερδίζει όλο και περισσότερο έδαφος στους χώρους της επιστημονικής κοινότητας, η δημιουργία δημόσιων βάσεων δεδομένων, έτοιμων προς χρήση από πολλούς επιστήμονες. Στην πραγματικότητα, υπάρχει μεγάλη πειθαρχία στον τομέα της *βιοπληροφορικής*. Η πειθαρχία αυτή φαίνεται να είναι εξ'ολοκλήρου αφιερωμένη στην επιμέλεια και την ανάλυση δεδομένων που αφορούν σε τέτοιες επιστήμες. Το μέγεθος καθώς και ο αριθμός των διαθέσιμων συνόλων δεδομένων αυξάνεται εκθετικά, ειδικότερα με την έλευση της Next Generation Sequencing. Πρόκειται για την διαδικασία προσδιορισμού της ακριβούς αλληλουχίας των νουκλεοτιδίων μέσα στο μόριο του DNA. Όπως είναι φανερό, η χρήση αυτής της τεχνολογίας παρέχει συνεχώς αναρίθμητα δεδομένα στους επιστήμονες σχετικά με τη γνώση των ανθρώπινων γονιδίων καθώς και πολυάριθμων τύπων ζωής.

Τα «Big Data» έχουν τη δυναμική να φέρουν επανάσταση ακόμα και στον τομέα της *εκπαίδευσης*. Πρόσφατα διεξήχθη λεπτομερής, ποσοτική σύγκριση ανάμεσα στις διαφορετικές προσεγγίσεις που ελήφθησαν από 35 πειραματικά σχολεία στη Νέα Υόρκη. Τα αποτελέσματα έδειξαν ότι μία από τις κορυφαίες πέντε εφαρμοσμένες τακτικές στα σχολεία, ήταν η χρήση δεδομένων που οδηγούσαν την κατάρτιση των μαθητών. Με άλλα λόγια, φανταστείτε, έναν κόσμο όπου έχουμε πρόσβαση σε μία τεράστια βάση δεδομένων, από την οποία συλλέγουμε οποιαδήποτε λεπτομερή μέτρηση για την ακαδημαϊκή επίδοση των μαθητών. Αυτά τα δεδομένα θα μπορούσαν να χρησιμοποιηθούν για τον σχεδιασμό κάποιων περισσότερο αποδοτικών προσεγγίσεων για την εκπαίδευση. Οι σχεδιασμοί αυτοί θα μπορούσαν

να αναφέρονται σε προχωρημένα, πανεπιστημιακού επιπέδου μαθήματα. Βρισκόμαστε ακόμη μακρυά από μια τέτοιου είδους πρόσβαση σε τέτοια δεδομένα, αλλά υπάρχουν ισχυρές τάσεις προς αυτή την κατεύθυνση. Συγκεκριμένα, υπάρχει μεγάλη τάση προς την ανάπτυξη ιστοσελίδων που αφορούν εκπαιδευτικές δραστηριότητες. Αυτό το γεγονός θα οδηγήσει στην παραγωγή ολοένα αυξανόμενης ποσότητας λεπτομερών δεδομένων σχετικά με την επίδοση των μαθητών.

Η *ιατρική φροντίδα* είναι άλλος ένας τομέας που μαρτυρά μια αξιοσημείωτη εφαρμογή των «Big Data». Είναι πλέον παγκοσμίως αποδεκτό πως η χρήση της τεχνολογίας πληροφοριών δύναται να μειώσει το κόστος της ιατρικής φροντίδας καθώς αυξάνει την ποιότητά της. Η ιατρική φροντίδα μπορεί να γίνει περισσότερο ποιοτική δίνοντας επιλέον προσοχή στην πρόληψη, στην εξατομίκευση και βάσει αυτών στην εκτενέστερη συνεχή παρακολούθηση των ασθενών. Για παράδειγμα, η United Health care στις Ηνωμένες Πολιτείες δαπανά μεγάλη προσπάθεια στην «αναγνώριση» των διαθέσεων των πελατών όπως καταγράφονται από φωνητικά αρχεία. Η εν λόγω εταιρεία συνδιάζει την επεξεργασία της φυσικής γλώσσας με δεδομένα κειμένου προκειμένου να αναγνωρίσει τα αισθήματα και την ικανοποίηση του πελάτη ασθενούς. Αυτό είναι ένα τυπικό παράδειγμα λήψης διαφορετικών «Big Data» και ανάπτυξης αναλυτικών μοντέλων. Η McKinsey &Company είναι παγκόσμια συμβουλευτική εταιρεία. Ασχολείται με την ποιοτική και την ποσοτική ανάλυση δεδομένων, με σκοπό τη λήψη βέλτιστων αποφάσεων. Σύμφωνα με δική της εκτίμηση, μόνο στις Ηνωμένες Πολιτείες γίνεται ετήσια εξοικονόμηση 300 δισεκατομμυρίων δολαρίων.

Η McKinsey&Company υποστηρίζει, επίσης, τη μεγάλη επίδραση των «Big Data» στον εργασιακό τομέα. Είναι γεγονός πως το 2010 ήταν αναγκαίοι στον τομέα της εργασίας 140,000 – 190,000 εργάτες με «αναλυτική» εμπειρία. Επιπροσθέτως, 1.5 εκατομμύριο διευθυντές διαφόρων επιχειρήσεων χρειάστηκε να ενημερωθούν για τα νέα δεδομένα της αγοράς. Αξιοσημείωτο γεγονός είναι πως η πρόσφατη έκθεση PCAST στο Network in gand I TR & D χαρακτήρισε τα «Big Data» ως «ερευνητικό εμπόδιο» το οποίο «μπορεί να επιταχύνει την πρόοδο σε ένα ευρύ φάσμα προτεραιοτήτων». Ακόμη και τα μέσα ενημέρωσης εκτιμούν την αξία τους καθώς είναι αποδεδειγμένη από άρθρα στον Economist [Eco2011], στους New York Times [NYT2012], και στο National Public Radio [NPR2011a, NPR2011b].

Η τελευταία δεκαετία χαρακτηρίζεται από την άνθηση ιστοσελίδων *των μέσων κοινωνικής δικτύωσης*, όπως το Facebook, το LinkedIn και το Twitter. Όλες αυτές οι ιστοσελίδεςδιευκολύνουντο αυξανόμενο φάσμα των ανθρωπίνων αλληλεπιδράσεων το οποίο παρέχει βάσεις «Big Data». Το γεγονός πως σε όλες τις πτυχές της ζωής μας εμφανίζονται πλέον τα μέσα κοινωνικής δικτύωσης, δηλώνεται σαν σύνθεση πολύπλοκων σχέσεων μεταξύ πολλών ατόμων. Είναι κοινώς αποδεκτό πως η έρευνα σ'αυτόν τον τομέα θα βελτιώσει τις γνώσεις μας πάνω στην τοπολογία των κοινωνικών δικτύων και θα ενισχύσει τη μελέτη των ανθρώπινων αλληλεπιδράσεων. Η Google, η Apple και το Facebook μαζί με πολλές άλλες διαδικτυακές εταιρείες που υπάρχουν σήμερα, παρέχουν υπηρεσίες που κυμαίνονται από την ανεύρεση παλιών

φίλων μέχρι τη δυνατότητα να μοιραζόμαστε σκέψεις και προσωπικές δραστηριότητες. Επιπλέον, αν σκεφτούμε πως υπάρχει η δυνατότητα να κοινοποιούμε την τοποθεσία στην οποία βρισκόμαστε ανά πάσα στιγμή και να αποθηκεύουμε τα αρχεία μας στο GoogleDrive, μπορούμε να αντιληφθούμε την τεράστια έκταση δεδομένων που συλλέγονται καθημερινά για καθένα από μας ξεχωριστά. Αυτά τα δεδομένα μπορούν να χρησιμοποιηθούν για την βελτίωση των υπηρεσιών που μας προσφέρονταιμέσω ποικιλίας αλγορίθμων. Όλες αυτές οι αποθήκες δεδομένων που αναφέρθηκαν παραπάνω, μας δελεάζουν προκειμένου να διανείμουμε τα προσωπικά μας δεδομένα σε συλλέκτες δεδομένων σαν αντάλλαγμα για τις υπηρεσίες που μας προσφέρονται. Ο δελεασμός, βέβαια, φαίνεται από τα τεράστια νούμερα χρηστών σε καθένα από τα αναπτυσσόμενα προγράμματα του Internet. Οι θεωρητικές τεχνικές δικτύων παρέχουν αποδοτικά μέσα για την ανάλυση όλων των παραπάνω δεδομένων. Ωστόσο, υπάρχουν περιορισμοί στα υπάρχοντα μοντέλα ανάλυσης τα οποία περιορίζουν την επέκταση αυτών των μεθόδων σε περισσότερο εξελιγμένες εφαρμογές. Παρόλα αυτά, αυτοί οι περιορισμοί, οφείλονται κυρίως στην έλλειψη χωρητικότητας για την επαρκή ανάλυση των ατελών δεδομένων που χαρακτηρίζουν τέτοιες εφαρμογές.

Τέλος, υπάρχουν πολλοί άλλοι τομείς που σχετίζονται άμεσα με την μεγάλη αξία των «Big Data». Στον *πολεοδομικό σχεδιασμό*, χρησιμοποιούνται μέσω της σύνδεσης αξιόπιστων γεωγραφικών δεδομένων. Στις *μεταφορές,* χρησιμοποιούνται μέσω ανάλυσης των λεπτομερών δεδομένων των οδικών δικτύων. Στον *περιβαλλοντικό σχεδιασμό, χ*ρησιμοποιούνται δίκτυα με αισθητήρες που συλλέγουν δεδομένα. Στις *υπολογιστικές επιστήμες*, χρησιμοποιείται μια νέα μεθοδολογία αυξανόμενης δημοφιλίας του δραματικά μειούμενου κόστους απόκτησης πληροφοριών. Στην *ανάλυση συστηματικού οικονομικού ρίσκου,* τα «Big Data» χρησιμοποιούνται μέσω ολοκληρωμένης ανάλυσης ενός δικτύου αντιθέσεων προκειμένου να φανούν οι εξαρτήσεις ανάμεσα σε οικονομικές οντότητες. Στην *εθνική ασφάλεια μέ*σω ανάλυσης κοινωνικών δικτύων και οικονομικών συναλλαγών πιθανών τρομοκρατών. Στην *ασφάλεια υπολογιστών*, τα «Big Data» χρησιμοποιούνται μέσω ανάλυσης συνδεόμενων πληροφοριών γνωστών σαν Security Information and Event Management (SIEM).

1.3 Ανεύρεση και Καταγραφή Δεδομένων

Τα «Big Data» δεν είναι μια καινούρια ιδέα. Ήδη, από την προηγούμενη δεκαετία κι ίσως νωρίτερα, αρκετοί ερευνητές ασχολήθηκαν και μελέτησαν το θέμα του μεγάλου όγκου δεδομένων.Ένα από τα βασικά θέματα που τους απασχόλησε ήταν ο τρόπος ανεύρεσης και καταγραφής δεδομένων. Μέσω επιστημονικών πειραμάτων και προσομοιώσεων προβλημάτων διαπίστωσαν πως είναιδυνατό να λαμβάνουν petabytes δεδομένων καθημερινά. Τα περισσότερα από αυτά τα δεδομένα δεν έχουν ενδιαφέρον και γι’ αυτό χρειάζεται έρευνα στο επιστημονικό πεδίο που αφορά στη μείωση των δεδομένων. Η μείωση αυτή είναι απαραίτητη έτσι ώστε οι χρήστες των

δεδομένων να μπορούν να τα διαχειριστούν χωρίς να χάνουν ουσιαστικές πληροφορίες. Θα πρέπει, επίσης να σημειωθεί πως απαιτούνται «on-line» τεχνικές ανάλυσης διότι δεν είναι δυνατόν πρώτα να αποθηκεύσουμε τα δεδομένα και στη συνέχεια να μειώσουμε το μέγεθός τους. Η πρωταρχική πρόκληση, λοιπόν, είναι να βρεθούν τα κατάλληλα φίλτρα προκειμένου να φιλτραριστούν τα «εισερχόμενα» δεδομένα και να μην χαθούν χρήσιμες πληροφορίες. Η δεύτερη μεγάλη πρόκληση είναι η αυτόματη παραγωγή των «metadata», δηλαδή των δεδομένων που θα χρησιμοποιηθούν προκειμένου να περιγραφείποιά δεδομένα συμμετέχουν, εγγράφονται και με ποιό τρόπο γίνεται η μέτρησή τους. Για παράδειγμα, σε διάφορα επιστημονικά πειράματα, οι σημαντικές λεπτομέρειες που αφορούν σε συγκεκριμένες πειραματικές συνθήκες, απαιτείται να ερμηνεύουν σωστά τα αποτελέσματα. Επιπροσθέτως, κρίνεται απαραίτητο, τα «metadata» να καταγράφονται με δεδομένα από παρατήρηση. Τα συστήματα προσκτήσεως «metadata», μειώνουν το βάρος του επιστήμονα – μελετητή, στην καταγραφή αυτών των δεδομένων.

Ιδιαίτερα σημαντικές έννοιες που εμπλέκονται στην ανεύρεση και την εγγραφήτων «Big Data», είναι ο όγκος *(volume), η προέλευση (provenance),η αλήθεια (veracity), η ταχύτητα (velocity)* και *η ποικιλία (variety)* αυτών των δεδομένων. Η έννοια του όγκου των δεδομένων έχει ιδιαίτερα μεγάλη σημασία σήμερα εξαιτίας της ραγδαίας διαθεσιμότητας βάσεων δεδομένων κλίμακας terabyte ή ακόμη και petabyte. Οι βάσεις αυτές προκύπτουν μέσω επιστημονικών προσομοιώσεων και πειραμάτων, οι οποίες μπορεί να αφορούν σε δεδομένα που προκύπτουν από επιχειρηματικές συναλλαγές ή ψηφιακά αποτυπώματα μεμονωμένων ατόμων. Η ραγδαία αύξηση του όγκου των δεδομένων σε διάφορους τομείς, είτε αναφερόμαστε σε επιχειρήσεις είτε σε ανθρωπιστικές επιστήμες, παρουσιάζει αλλαγές στην κλίμακα και στην *προέλευση (provenance)* των δεδομένων. Όσον αφορά στην προέλευση των δεδομένων, θα πρέπει να σημειωθεί πως η καταγραφή πληροφοριών σχετικά με τα δεδομένα κατά τη γέννησή τους, είναι άχρηστη αν αυτές οι πληροφορίες δεν μπορούν να μεταφραστούν μέσω της πορείας ανάλυσης. Με την κατάλληλη προέλευση, γίνεται εύκολη η αναγνώριση όλης της μετέπειτα επεξεργασίας των δεδομένων. Παρόλα αυτά, απαιτείται επιπλέον έρευνα τόσο στον τομέα της παραγωγής κατάλληλων «metadata», όσο και συστημάτων δεδομένων που κρατούν την προέλευση τους κατά τη διάρκεια της διαδικασίας ανάλυσης. Ο όρος *αλήθεια (veracity)*, αναφέρεται στην ποιότητα των δεδομένων λαμβάνοντας υπόψη την πολυπλοκότητά τους, τις χαμένες τιμές, το θόρυβο και την «μετατόπιση» του συνόλου δεδομένων. Η έννοια της μετατόπισης είναι περισσότερο προφανής στην περίπτωση των «Big Data» καθώς τα αφανή δεδομένα ίσως παρουσιάζουν μια διανομή που δεν είναι ορατή στα χρησιμοποιούμενα δεδομένα. Αυτό το πρόβλημα είναι άμεσα συνδεδεμένο με την έννοια της *ταχύτητας (velocity)*και παρουσιάζει την πρόκληση δημιουργίας αλγορίθμων που μπορούν να αντιμετωπίσουν ‘αναταράξεις’ στην διανομή των δεδομένων. Αξίζει να σημειωθεί πως η δημιουργία τέτοιων αλγορίθμων είναι ένας καθιερωμένος ερευνητικός τομέας στο πεδίο απόκτησης δεδομένων και λαμβάνει τη μορφή απόκτησης πληροφοριών μέσω δεδομένων συνεχούς ροής («streaming data»). Το θέμα της *ποικιλίας (variety)* είναι αναμφισβήτητα μοναδικό και πολύ ενδιαφέρον.

Η εισροή πολυμορφικών δεδομένων, όπως τα μέσα κοινωνικής δικτύωσης, οι φωτογραφίες, τα video, σε συνδιασμό με τα δεδομένα συγκεκριμένης μορφής, παρέχει συνεχώς νέες ευκαιρίες για απόκτηση «Big Data».

1.4 Επεξεργασία και Ανάλυση Δεδομένων

Στις περισσότερες περιπτώσεις οι πληροφορίες που συλλέγονται από τα δεδομένα δεν βρίσκονται στην κατάλληλη μορφή για ανάλυση. Είναι λογικό, λοιπόν, πως δεν μπορούμε να αναμένουμε αποδοτική ανάλυση των δεδομένων αν προηγουμένως δεν έχουν πάρει συγκεκριμένη μορφή. Σ' αυτό το στάδιο χρειάζεται μια διαδικασία εξαγωγής πληροφοριών που αναδεικνύει τις απαιτούμενες πληροφορίες από τις πηγές δεδομένων και τις εκφράζει σε μια δομημένη μορφή κατάλληλη προς ανάλυση. Μια τέτοια 'εξαγωγή' πληροφοριών είναι σημαντικά εξαρτημένη από την εκάστοτε εφαρμογή. Αυτό σημαίνει ότι υπάρχει μεγάλη διαφορά στην λήψη πληροφιών από μια εικόνα των άστρων και από μια απεικόνιση μαγνητικής τομογραφίας.

Δεδομένης της ανομοιογένειας των εισερχόμενων δεδομένων, δεν είναι αρκετή η απλή καταγραφή τους σε μια βάση δεδομένων. Αρκεί να σκεφτούμε απλά τα δεδομένα που μπορούμε να εξάγουμε από μια σειρά επιστημονικών πειραμάτων. Αν έχουμε απλά μερικά σύνολα δεδομένων, αυτό δεν εξασφαλίζει ότι είναι δυνατή η ανεύρεση των πληροφοριών που χρειαζόμαστε. Πιθανώς, κάποια επαρκή «metadata» δίνουν κάποια ελπίδα, αλλά εξακολουθούν να υπάρχουν εμπόδια εξαιτίας των διαφορών στις πειραματικές λεπτομέρειες και στην δομή της καταγραφής των δεδομένων.

Η ανάλυση δεδομένων δημιουργεί αρκετά περισσότερες προκλήσεις σε σχέση με τον απλό εντοπισμό, τον προσδιορισμό, την κατανόηση και την επίκληση δεδομένων. Προκειμένου να έχουμε μια αποδοτική ανάλυση δεδομένων μεγάλης κλίμακας, όλη η ανάλυσή τους θα πρέπει να διενεργηθεί με έναν εντελώς αυτοματοποιημένο τρόπο. Για να επιτευχθεί αυτός ο τρόπος, θεωρείται απαραίτητη η έκφραση των διαφορών στη δομή των δεδομένων, στη μορίες. Λαμβάνοντας υπόψη τους σκοπούς για τους οποίους θα χρησιμοποιηθεί μια βάση δεδομένων, κάθε σχεδιασμός έχει πλεονεκτήματα και μειονεκτήματα σε σχέση με άλλους σχεδιασμούς βάσεων. Ο σχεδιασμός βάσεων δεδομένων αποτελεί σήμερα έναν ιδιαίτερα ενδιαφέρον και σημαντικό επιστημονικό τομέα. Καθώς οι απαιτήσεις διαχείρισης δεδομένων αυξάνονται καθημερινά, κρίνεται απαραίτητη η συνεχής εξέλιξη του σχεδιασμού τέτοιων βάσεων. Προς αυτήν την κατεύθυνση θα βοηθούσε, πιθανώς, η επινόηση νέων εργαλείων σχεδίασης που θα έδιναν πιο αποτελεσματική ανάλυση σε λιγότερο χρόνο. Επίσης ενδιαφέρουσα ιδέα θα ήταν η παραίτηση από τους ήδη υπάρχοντες σχεδιασμούς με σκοπό την ανάπτυξη νέων τεχνικών ώστε να γίνεται αποτελεσματική χρήση των βάσεων δεδομένων απουσίας έξυπνου σχεδιασμού βάσης.

Οι υπάρχουσες μέθοδοι εξέτασης «Big Data» είναι ριζικά διαφορετικές σε σχέση με τις παραδοσιακές στατιστικές αναλύσεις μικρών δειγμάτων δεδομένων. Τα «Big

Data» είναι συχνά ανομοιογενή και αναξιόπιστα. Παρόλα αυτά, ακόμη και μ'αυτά τα χαρακτηριστικά, μπορούν να είναι περισσότερο πολύτιμα σε σχέση με μικρά δείγματα διότι οι γενικές στατιστικές συνήθως αποκαλύπτουν περισσότερο αξιόπιστα μονοπάτια ανάλυσης και κρυμμένη γνώση. Επιπροσθέτως, τα δίκτυα των «Big Data» που συνδέουν ανοιμοιογενείς πληροφορίες, έχουν την δυνατότητα να βοηθήσουν στην αντιστάθμιση χαμένων δεδομένων καθώς διακρίνονται από πλεονασμό πληροφοριών. Με τέτοια δίκτυα διευκολύνεται επίσης η διασταύρωση συγκρουόμενων περιπτώσεων δεδομένων και η επικύρωση αξιόπιστων σχέσεων μεταξύ τους.

1.5 Ερμηνεία Αποτελεσμάτων

Η ύπαρξη της δυνατότητας ανάλυσης των «Big Data» έχει πολύ μικρή αξία αν οι χρήστες των δεδομένων δεν μπορούν να κατανοήσουν την ανάλυση αυτή. Η ερμηνεία των αποτελεσμάτων μιας ανάλυσης δεδομένων είναι μια δύσκολη διαδικασία, η οποία συχνά απαιτεί τη μελέτη όλων των υποθέσεων που έχουν γίνει κατά την ανάλυση. Βασικό στοιχείο που προκύπτει σαν εμπόδιο στην προσπάθεια ερμηνείας, είναι το γεγονός πως υπάρχουν πολλές πηγές σφαλμάτων. Αρκεί να σκεφτούμε πως τα υπολογιστικά συστήματα μπορεί να έχουν σφάλματα, τα μοντέλα ανάλυσης που χρησιμοποιούμε μπορεί να βασίζονται σε υποθέσεις και τα τελικά αποτελέσματα να προκύπτουν από λανθασμένα δεδομένα. Εξαιτίας όλων των παραπάνω, ο ερευνητής δεν θα πρέπει να αφήσει εξ' ολοκλήρου τη μελέτη του πάνω στο υπολογιστικό σύστημα που χρησιμοποιεί. Καθήκον του είναι να προσπαθήσει να κατανοήσει και να επικυρώσει την αλήθεια των αποτελεσμάτων που προκύπτουν από το σύστημα. Βέβαια, η διαδικασία αυτή δεν είναι απλή εξαιτίας της πολυπλοκότητας των «Big Data».

Είναι σπάνιο, λοιπόν, να είναι αρκετά μόνο τα αποτελέσματα της ανάλυσης μιας βάσης δεδομένων. Συχνά χρειάζονται συμπληρωματικές πληροφορίες οι οποίες επεξηγούν τον τρόπο με τον οποίο αντλήθηκε ένα αποτέλεσμα και πάνω σε ποιά δεδομένα βασίστηκε. Αυτές οι επιπλέον πληροφορίες που χρησιμοποιούνται αποτελούν την προέλευση των δεδομένων που προκύπτουν σαν αποτέλεσμα της ανάλυσης. Αναζητώντας τον καλύτερο τρόπο για την απόκτηση, την αποθήκευση και την εξέταση της προέλευσης των δεδομένων, σε συνδιασμό με τεχνικές κατάλληλες να συλλάβουν επαρκή «metadata», είναι δυνατή η δημιουργία της κατάλληλης υποδομής. Μια κατάλληλη υποδομή θα έδινε τη δυνατότητα στους χρήστες της να ερμηνεύσουν ορθά τα αναλυτικά αποτελέσματα αλλά και να επαναλάβουν την ανάλυση με διαφορετικές υποθέσεις, παραμέτρους και σύνολα δεδομένων.

Επιπροσθέτως, ο χρήστης των δεδομένων, θα πρέπει να έχει την ικανότητα αναγνώρισης της προέλευσης των δεδομένων προκειμένου να κατανοήσει πλήρως τις πληροφορίες που λαμβάνει από τη βάση. Αυτό σημαίνει ότι δεν αρκεί να έχει πρόσβαση στα αποτελέσματα, αλλά θα πρέπει να κατανοεί για ποιό λόγο λαμβάνει τα

συγκεκριμένα αποτελέσματα. Βέβαια, πολλές φορές αυτή η διαδικασία είναι δύσκολη για πολλούς χρήστες. Εναλλακτικά, με σκοπό να είναι περισσότερο προσβάσιμη η ερμηνεία των αποτελεσμάτων, θα μπορούσε να δίνεται η δυνατότητα στο χρήστη να παρεμβαίνει σε κάποια από τα στάδια της ανάλυσης, κάνοντας ίσως μικρές αλλαγές στη ροή της διαδικασίας ή στις χρησιμοποιούμενες παραμέτρους. Ύστερα, οι χρήστες μπορούν να έχουν μια άποψη αυτών των σταδιακών αλλαγών. Μ' αυτόν τον τρόπο, οι χρήστες αποκτούν μια διαισθητική άποψη της ανάλυσης και επιπλέον επικυρώνουν ότι τα αποτελέσματα της ανάλυσης είναι αυτά που αναμένονταν. Τέλος, εννοείται πως για να επιτευχθούν τα παραπάνω, θα πρέπει το χρησιμοποιούμενο υπολογιστικό σύστημα να παρέχει τις κατάλληλες εγκαταστάσεις ώστε να είναι εύκολος ο προσδιορισμός της ανάλυσης από το χρήστη.

1.6 Μηχανική Μάθηση με «Big Data»

Η Μηχανική Μάθηση ή όπως είναι ευρέως γνωστή «Machine Learning», βρίσκεται ανάμεσα στις κυριότερες τεχνικές για την ανάλυση δεδομένων. Βάσει αυτού του γεγονότος αξίζει να σημειωθούν μερικά από τα πλεονεκτήματα των «Big Data» στην μηχανική μάθηση.

Τα τελευταία χρόνια παρατηρήθηκε από τους ερευνητές πως το μέγεθος των διαφόρων προς μελέτη δεδομένων αυξάνεται ραγδαία με αποτέλεσμα να καθίσταται ιδιαίτερα δύσκολη η αποθήκευσή τους κι ακόμη περισσότερο δύσκολη, η σάρωσή τους πάνω από μια φορές. Παρόλα αυτά γινόταν πάντα προσπάθεια διαχείρισης του ολοένα αυξανόμενου μεγέθους τους και θεωρούνταν πως οι κύριοι υπολογισμοί από την ανάλυσή τους μπορούσαν να αποθηκευτούν στην μνήμη. Μέσα από διαδοχικές μελέτες, όμως, άρχισαν να προκύπτουν διάφορα ερωτήματα. Η δυνατότητα της μοναδικής σάρωσης δεδομένων να προσφέρει τις απαραίτητες πληροφορίες τέθηκε υπό αμφισβήτηση. Επιπλέον, οι ερευνητές άρχισαν να διερωτώνται αν είναι μεγάλη η απαίτηση αποθήκευσης των δεδομέων και κατά πόσο αυτή εξαρτάται από το μέγεθός τους. Έτσι οδηγήθηκαν στην ανάπτυξη των αλγόριθμων «one–pass learning». Πρόκειται για αλγόριθμους που απαιτούν μόνο μια σάρωση των δεδομένων με την ελάχιστη απαίτηση αποθήκευσής τους και είναι ανεξάρτητοι του μεγέθους των δεδομένων. Αυτοί οι αλγόριθμοι είναι ιδιαίτερα σημαντικοί γιατί σε πολλές εφαρμογές των «Big Data», τα δεδομένα δεν είναι απλά πολλά, αλλά συσσωρεύονται συνεχώς. Ως εκ τούτου, είναι αδύνατη η γνώση του μεγέθους της βάσης δεδομένων.

Μια από τις κυριότερες ανησυχίες τόσο στον σχεδιασμό των αλγορίθμων μηχανικής μάθησης όσο και στην εφαρμογή τους, είναι ο κίνδυνος της υπερ-τοποθέτησης (over fitting) δεδομένων στο σύστημα ανάλυσης. Αυτή η κατάσταση οδήγησε στη ροπή της μελέτης προς απλά μοντέλα που χρησιμοποιούν λιγότερες παραμέτρους. Η χρήση των «Big Data», φαίνεται πως μπορεί να διευκολύνει την κατάσταση καθώς προσφέρει όλο και περισσότερα διαθέσιμα στοιχεία προς μελέτη, σαν απάντηση στον κίνδυνο της υπερ-τοποθέτησης. Επιπλέον, οι περιορισμοί που επιφέρουν οι

λιγότερες χρησιμοποιούμενες παράμετροι, ίσως μπορούν να καμφθούν με τα «Big Data». Χάρη σε αυτά τα δεδομένα μεγάλου μεγέθους, είναι δυνατή, πλέον, η δημιουργία μοντέλων με δισεκατομμύρια παραμέτρους και η εφαρμογή τους μέσω πανίσχυρων υπολογιστικών εγκαταστάσεων.

Ένα εξαιρετικό χαρακτηριστικό των «Big Data», είναι το γεγονός πως περιέχουν μεγάλο πλήθος πληροφοριών που μπορούν να δώσουν απάντηση σε πολλά ερωτήματα. Βασικό στοιχείο, όμως, είναι η δυνατότητα εκτίμησης της πραγματικής αξίας αυτών των πληροφοριών. Λύση σ' αυτό το πρόβλημα αποτελεί η στροφή στα τεστ στατιστικών υποθέσεων. Τα στατιστικά τεστ μπορούν να επικυρώσουν δύο πράγματα. Πρώτον, επικυρώνουν πως η μελέτη δεδομένων που έγινε ήταν ακριβώς αυτή που χρειαζόταν ανά περίπτωση και δεύτερον, πως τα αποτελέσματα της μελέτης δεν επηρεάστηκαν από πιθανές 'ανωμαλίες' στα δεδομένα. Παρά το γεγονός ότι τα στατιστικά τεστ έχουν μελετηθεί για αιώνες και έχουν χρησιμοποιηθεί για δεκαετίες στην μηχανική μάθηση, ο σχεδιασμός και η ανάπτυξη επαρκών στατιστικών μεθόδων δεν είναι απλός.

Επιπροσθέτως, πολλές φορές τα «Big Data» υπάρχουν 'χωρισμένα' σε τμήματα. Αυτό σημαίνει πως διαφορετικά κομμάτια των συνολικών δεδομένων βρίσκονται στα χέρια διαφορετικών ιδιοκτητών, χωρίς να κατέχει κανείς το σύνολο των δεδομένων. Αυτό συμβαίνει συχνά στην περίπτωση κατά την οποία κάποιες πηγές είναι καίριες για τους αναλυτικούς στόχους, ενώ κάποιες άλλες έχουν λιγότερη σημασία. Με αυτή τη βάση προκύπτουν ερωτήματα που έχουν τεθεί υπό μελέτη τα τελευταία χρόνια. Πολλοί ερευνητές ξεκινώντας από την υπόθεση πως οι διαφορετικοί ιδιοκτήτες μπορούν να δώσουν πρόσβαση αναλυτών στα δεδομένα τους με διαφορετικά όμως δικαιώματα, μελετούν αν υπάρχει η δυνατότητα ανάμειξης των πηγών δεδομένων χωρίς να υπάρχει πρόσβαση στο σύνολο των δεδομένων.

Συν τοις άλλοις, διαφορετικοί χρήστες συνήθως έχουν διαφορετικές απαιτήσεις από τα ίδια δεδομένα. Η κατασκεύη όμως ενός μοντέλου για καθεμία από τις ποικίλλες απαιτήσεις ξεχωριστά, εμποδίζεται από τα φορτία του υπολογισμού και της αποθήκευσης των δεδομένων. Πρόσφατα μελέτες έχουν στραφεί στην προσπάθεια δημιουργίας ενός γενικού μοντέλου που θα μπορεί να προσαρμοστεί σε διαφορετικές απαιτήσεις με ελάχιστες τροποποιήσεις κάθε φορά.

Κεφάλαιο 2

Σχεδιασμός Πειραμάτων

2.1 Πείραμα

Ένα πείραμα είναι μια διαδικασία που διεξάγεται με στόχο την επικύρωση ή τη διάψευση της εγκυρότητας μιας υπόθεσης. Ένα πείραμα λειτουργεί με την βάση της αιτίας – αποτελέσματος αποδεικνύοντας πως το αποτέλεσμα του πειράματος εξαρτάται από τον τρόπο χειραγώγισης ενός συγκεκριμένου παράγοντα. Τα πειράματα ποικίλουν σημαντικά στους στόχους τους αλλά πάντα βασίζονται σε επαναλαμβανόμενες διαδικασίες και στη λογική ανάλυση των αποτελεσμάτων. Τυπικά, τα πειράματα περιλαμβάνουν ελέγχους οι οποίοι είναι σχεδιασμένοι να ελαχιστοποιούν τις επιδράσεις όλων των μεταβλητών εκτός της μοναδικής ανεξάρτητης μεταβλητής. Αυτή η διαδικασία αυξάνει την αξιοπιστία των αποτελεσμάτων, συχνά μέσω σύγκρισης μεταξύ των μετρήσεων των ελέγχων και των υπόλοιπων μετρήσεων. Οι επιστημονικοί έλεγχοι είναι μέρος μιας επιστημονικής μεθόδου. Σε ιδανική κατάσταση, όλες οι μεταβλητές που συμμετέχουν σε ένα πείραμα ελέγχονται και καμία δεν βρίσκεται εκτός ελέγχου. Σε μια τέτοια κατάσταση, αν όλοι οι έλεγχοι λειτουργούν όπως αναμένεται, είναι πιθανό να καταλήξουμε στο γεγονός πως το πείραμα δουλεύει όπως ακριβώς προοριζόταν. Επιπλέον, τα αποτελέσματα αυτού του πειράματος θεωρούνται πλήρως έγκυρα και εξαρτώμενα μόνο από την μεταβλητή που μελετάται.

2.2 Βασικές αρχές σχεδιασμού πειραμάτων

Γενικά, ο σχεδιασμός πειραμάτων ή αλλιώς, πειραματικός σχεδιασμός, είναι ο σχεδιασμός οποιωνδήποτε ‘μεθόδων’ συγκέντρωσης πληροφοριών, όπου υπάρχει μεγάλη ποικιλία, είτε με ύπαρξη πλήρους ελέγχου του πειραματιστή, είτε όχι. Παρόλα αυτά, στη στατιστική ο όρος «σχεδιασμός πειραμάτων», χρησιμοποιείται στην περίπτωση ελεγχόμενων από τον πειραματιστή πειραμάτων. Ο σχεδιασμός πειραμάτων χρησιμοποιείται συχνά στην αξιολόγηση χημικών σκευασμάτων, κατασκευών, συστατικών και διάφορων υλικών. Ο πειραματιστής συνήθως ενδιαφέρεται για την επιρροή κάποιων διαδικασιών ή παρεμβάσεων πάνω σε διάφορα αντικείμενα. Σκοπός του είναι να βελτιστοποιήσει μια διαδικασία ή ένα σύστημα μέσω ενός πειράματος και να καταλήξει σε χρήσιμα αποτελέσματα σχετικά με τη συμπεριφορά του εξεταζόμενου αντικειμένου. Αναλογιζόμενοι το κόστος ενός μοναδικού πειράματος, στόχος είναι πάντα η ελαχιστοποίηση του αριθμού των διεξαγόμενων πειραμάτων. Με τον σχεδιασμό πειραμάτων, αυτός ο αριθμός διατηρείται σε χαμηλά επίπεδα, όσο το δυνατόν περισσότερο και επιλέγεται ο συνδιασμός των παραγόντων που θα αποδώσει τις περισσότερες πληροφορίες.

Προκειμένου να γίνει περισσότερο αντιληπτή η βασική ιδέα του σχεδιασμού πειραμάτων, θα πρέπει να εξετάσουμε τον τρόπο με τον οποίο γινόταν παραδοσιακά αυτός ο σχεδιασμός. Η αρχική προσέγγιση είναι να γίνει αλλαγή της τιμής ενός από τους παράγοντες του πειράματος κάθε φορά, μέχρι το σημείο όπου δεν παρατηρείται επιπλέον βελτίωση στην απόκριση. Είναι σαφές πως είναι αρκετά δύσκολος ο

εντόπισμος αυτής της κρίσιμης τιμής για τον κάθε παράγοντα από τον πειραματιστή. Σε αυτή την περίπτωση, λοιπόν, ο πειραματιστής δημιουργεί ένα σύνολο πειραμάτων γύρω από ένα κεντρικό σημείο (‘center point’). Βάσει αυτού, ο σχεδιασμός πειραμάτων καθοδηγείται απότην οργάνωση μιας συμμετρικής διανομής των πειραμάτων γύρω από ένα κεντρικό σημείο. Δεδομένου ενός συγκεκριμένου εύρους των παραγόντων που θα επηρεάσουν το πείραμα, είναι εύκολος ο υπολογισμός αυτού του κεντρικού σημείου.

2.3 Μαθηματικά Μοντέλα

Η βάση του σχεδιασμού πειραμάτων είναι μια προσέγγιση του πραγματικού προβλήματος με τη βοήθεια ενός μαθηματικού μοντέλου. Προκειμένου να προσομοιωθούν τα σημαντικά χαρακτηριστικά του μελετούμενου συστήματος, χρησιμοποιούνται οι αποκρίσεις και οι παράγοντες. Η απόκριση δίνει στον πειραματιστή τις απαραίτητες πληροφορίες για το σύστημα προς μελέτη και οι παράγοντες χρησιμοποιούνται για τη διαχείριση του συστήματος. Οι συνηθέστεροι παράγοντες μπορούν να δεχθούν δύο ή περισσότερες τιμές και να διακριθούν σε κατηγορίες βάσει διαφόρων κριτηρίων. Έτσι, υπάρχουν οι ελέγξιμοι, οι μη ελέγξιμοι, οι ποσοτικοί και οι ποιοτικοί παράγοντες.

Ένα μοντέλο φυσικά δεν είναι ποτέ τέλειο, αλλά βοηθά σημαντικά στην προσομοίωση της πολυπλοκότητας ενός προβλήματος μέσω μιας διαχειρίσιμης εξίσωσης. Το απλούστερο από τα υπάρχοντα μοντέλα είνα το γραμμικό. Σε αυτό το μοντέλο, οι g παράγοντες που επηρεάζουν το πρόβλημα προσομοιώνονται ως $x_1, .., x_g$. Αυτοί οι παράγοντες επηρεάζουν την απόκριση y σύμφωνα με την παρακάτω σχέση:

$$y = \beta_0 + \beta_1 x_1 + \ ... + \beta_g x_g + \varepsilon \qquad (2.1).$$

Σε αυτή την περίπτωση οι συντελεστές β_0 , β_1 , ..., β_g είναι οι συντελεστές παλινδρόμησης και το «ε» αντιπροσωπεύει το ‘τυχαίο’μέρος του μοντέλου, το οποίο θεωρείται ότι γενικά έχει μέση τιμή μηδέν και διακύμανση σ^2. Υπάρχει η δυνατότητα επέκτασης σε μια σχέση με N πολλαπλές αποκρίσεις. Έτσι, καταλήγουμε στη σχέση:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_g x_{ig} + \varepsilon_i \qquad (2.2)$$

όπου το y_i ανταποκρίνεται στηνi-οστή απόκριση με παράγοντες x_{i1}, x_{i2}, ….,x_{ig}. Οι N σχέσεις που δημιουργούνται μπορούν να γραφούν στη μορφή πίνακα ως εξής:

$$Y = X\beta + \varepsilon \qquad (2.3)$$

όπου ο πίνακας X διαστάσεων N × (g + 1), περιέχει όλους τους παράγοντες για τις αποκρίσεις. Τα Y και ε είναι διανύσματα N×1. Οι συντελεστές παλινδρόμησης β αποτελούν την άγνωστη παράμετρο στο μοντέλο. Συγκεντρωτικά τα παραπάνω μπορούν να γραφούν με την ακόλουθη μορφή πινάκων:

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}, X = \begin{bmatrix} 1 & \cdots & x_{1g} \\ \vdots & \ddots & \vdots \\ 1 & \cdots & x_{Ng} \end{bmatrix}, \beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_N \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_N \end{bmatrix} \quad (2.4)$$

Εκτός από το παραπάνω γραμμικό μοντέλο, υπάρχουν κι άλλα μαθηματικά μοντέλα, όπως τα μοντέλα αλληλεπιδράσεων (interaction models) και τα τετραγωνικά μοντέλα (quadratic models).

Τα μοντέλα αλληλεπιδράσεων είναι περισσότερο πολύπλοκα σε σχέση με τα γραμμικά μοντέλα, αλλά χρησιμοποιούνται για τους ίδιους πειραματικούς σκοπούς με τα γραμμικά. Αυτού του τύπου τα μοντέλα περιέχουν τους ίδιους όρους με τα γραμμικά αλλά με πρόσθετους όρους αλληλεπίδρασης. Ένας όρος αλληλεπίδρασης προκύπτει από το συνδιασμό δύο παραγόντων x_i και x_j, με συντελεστή σύνδεσης β_{ij}. Η σχέση (2.5) που ακολουθεί δίνει ένα παράδειγμα του μοντέλου αλληλεπίδρασης με τρεις παράγοντες:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 + \varepsilon \quad (2.5).$$

Το τετραγωνικό μοντέλο (quadratic models) είναι το τρίτο συνηθέστερο χρησιμοποιούμενο μαθηματικό μοντέλο στο σχεδιασμό πειραμάτων. Αυτό το μοντέλο επεκτείνει το μοντέλο αλληλεπιδράσεων με πρόσθετους τετραγωνικούς όρους για κάθε παράγοντα. Ένας τετραγωνικός όρος είναι το τετράγωνο ενός παράγοντα x_i με το συντελεστή τον β_{ii}. Τα τετραγωνικά μοντέλα είναι πιο περίπλοκα σε σχέση με τους προηγούμενους δύο τύπους μοντέλων που περιγράφηκαν προηγουμένωνωςκαι χρησιμοποιούνται για διαδικασίες βελτιστοποίησης. Η σχέση (2.6) που ακολουθεί δίνει ένα παράδειγμα του τετραγωνικού μοντέλου με τρεις παράγοντες:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{11} {x_1}^2 + \beta_{22} {x_2}^2 + \beta_{33} {x_3}^2 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 + \varepsilon \quad (2.6)$$

2.4 Βασικοί Πειραματικοί Στόχοι

Ο βασικός στόχος της χρήσης των πειραματικών σχεδιασμών είναι η ικανοποίηση κάποιων βασικών πειραματικών στόχων. Βάσει αυτών, προκύπτουν τρεις βασικοί τύποι σχεδιασμών.

Σχεδιασμοί Κρησαρίσματος(«Screening designs»)

Ο όρος «screening design» αναφέρεται σε έναν πειραματικό σχεδιασμό που προορίζεται να ανακαλύψει τους λιγότερο σημαντικούς παράγοντες ανάμεσα σε μια λίστα παραγόντων. Συγκεκριμένα, πρωταρχικός στόχος αυτών των σχεδιασμών είναι η συγκεκριμενοποίηση των κύριων επιδράσεων και η διερεύνηση των αλλαγών στην απόκριση καθώς οι παράγοντες λαμβάνουν διάφορες τιμές. Όσον αφορά στις αλληλεπιδράσεις μεταξύ των διαφόρων παραγόντων, μπαίνουν σε φθίνουσα σειρά. Εξαιτίας των παραπάνω, τέτοιοι σχεδιασμοί χρησιμοποιούνται στην περίπτωση μεγάλου αριθμού παραγόντων και παρουσιάζονται στο αρχικό στάδιο της ερευνητικής μελέτης του πειραματιστή καθώς απαιτείται να χαρακτηρίσει την πειραματική διαδικασία. Η ταυτοποίηση των κρίσιμων για την διαδικασία, παραγόντων, είναι ιδιαίτερα σημαντική για την μετέπειτα βελτίωση του πειράματος διότι μόνο ένα υποσύνολο των παραγόντων θα χρησιμοποιηθεί τελικά. Συνηθέστερα, οι σχεδιασμοί τέτοιου τύπου είναι αναλυτικής τάξης resolution III ή resolution IV. Άλλη γνωστή οικογένεια σχεδιασμών διαλογής είναι το σύνολο των «Plackett-Burman» σχεδιασμών. Αξίζει να σημειωθεί πως οι «screening designs» είναι οι οικονομικότεροισχεδιασμοι που επικεντρώνουν στην απόφαση της σημασίας των περισσότερων κύριων επιδράσεων.

Σχεδιασμός Βελτιστοποίησης («Optimization design»)

Μετά το στάδιο της διαλογής, είναι σύνηθες να γίνεται βελτιστοποίηση. Αυτού του τύπου ο σχεδιασμός δίνει στον πειραματιστή λεπτομερείς πληροφορίες σχετικά με την επιρροή των παραγόντων και ξεδιαλύνει τον συνδιασμό των παραγόντων που οδηγεί στην καλύτερη απόκριση. Με άλλα λόγια, ο σχεδιασμός αυτός βοηθά στην ανεύρεση του βέλτιστου πειραματικού σημείου, προβλέποντας τις τιμές της απόκρισης για όλους τους πιθανούς συνδιασμούς παραγόντων (Montgomery 1991).

Εξέταση Ανθεκτικότητας («Robustness Test»)

Αυτός ο σχεδιασμός είναι ο τελευταίος που δημιουργείται πριν ολοκληρωθεί η πειραματική διαδικασία. Στόχος του είναι ο προσδιορισμός του τρόπου προσαρμογής των παραγόντων προκειμένου να είναι εγγυημένη η ανθεκτικότητα. Σε αυτή την περίπτωση, η ανθεκτικότητα δηλώνει ότι μικρές διακυμάνσεις των παραγόντων δεν επηρεάζουν την απόκριση με σαφή τρόπο. Αν μια διαδικασία δεν ικανοποιεί το τεστ ανθεκτικότητας, τα όρια της πειραματικής διαδικασίας θα πρέπει να αλλάξουν.

2.5 Στατιστικοί Πειραματικοί Σχεδιασμοί

Υπάρχουν τρεις βασικοί τύποι στατιστικών σχεδιασμών για συνηθισμένες πειραματικές περιοχές, οι οποίοι έχουν διαφορετικά χαρακτηριστικά και εφαρμογές σε διαφορετικές ερευνητικές περιοχές.

2.5.1 Παραγοντικοί Σχεδιασμοί (Full Factorial Designs)

Οι παραγοντικοί σχεδιασμοί είναι πολύ σημαντική μέθοδος προκειμένου να διευκρινιστούν οι επιδράσεις πολλαπλών μεταβλητών σε μια απόκριση. Με σκοπό να δώσουμε έναν ορισμό στους παραγοντικούς σχεδιασμούς θα μπορούσαμε να πούμε

«*Με τον όρο παραγοντικοί σχεδιασμοί εννοούμε ότι σε κάθε πλήρη δοκιμή ή επανάληψη του πειράματος, εξετάζονται όλοι οι δυνατοί συνδυασμοί των επιπέδων των παραγόντων».*

Παραδοσιακά, ένα πείραμα σχεδιάζεται ώστε να εξετάζεται η επίδραση μιας μεταβλητής πάνω σε μία απόκριση. Ένα πείραμα που βασίζεται σε έναν πλήρη παραγοντικό σχεδιασμό (full factorial experiment), περιλαμβάνει δύο ή περισσότερους παράγοντες ο καθένας από τους οποίους παίρνει διακεκριμένες τιμές ή αλλιώς ‘επίπεδα’ (levels). Οι πειραματικές μονάδες των παραγόντων αυτού του πειράματος λαμβάνουν υπόψη όλους τους δυνατούς συνδιασμούς των επιπέδων όλων των παραγόντων. Ένας πλήρης παραγοντικός σχεδιασμος (full factorial design) μπορεί επίσης να ονομαστεί ‘πλήρως διασταυρωμένος σχεδιασμός (fully crossed design)’. Ένας σχεδιασμός αυτού του είδους είναι ιδιαίτερα αποδοτικός καθώς δίνει στον ερευνητή- πειραματιστή πολλά πλεονεκτήματα. Ο Fisher ήταν ο πρώτος που απέδειξε την πληθώρα των πλεονεκτημάτων του παραγοντικού σχεδιασμού, συνδυάζοντας την μελέτη των πολλαπλών μεταβλητών στο ίδιο παραγοντικό πείραμα.Τοπρώτο σημαντικό όφελος της χρήσης ενός παραγοντικού σχεδιασμού είναι το γεγονός πως είναι δυνατόν να μειωθεί αισθητά ο αριθμός των πειραματικών εκτελέσεων που πρέπει να κάνει ο ερευνητής-πειραματιστής, προκειμένου να μελετήσει πολλαπλούς παράγοντες ταυτόχρονα. Επιπλέον, σημαντικό στοιχείο για ένα τέτοιο σχεδιασμό είναι το γεγονός πως μπορεί να μελετηθεί η επίδραση κάθε ενός ανεξάρτητου παράγοντα πάνω στην μεταβλητή της απόκρισης (response variable) . Οι επιδράσεις αυτές ονομάζονται ‘κύριες επιδράσεις’ και είναι διαδεδομένες με τον όρο ‘main effects’. Το χαρακτηριστικό, όμως, που ξεχωρίζει τον παραγοντικό σχεδιασμό είναι η δυνατότητα μελέτης επιρροής της αλληλεπίδρασης των διαφόρων παραγόντων πάνω στην μεταβλητή απόκρισης. Η μελέτη αυτών των επιδράσεων είναι απαραίτητη όταν χρειάζεται να ληφθούν υπόψη όλες οι δυνατές επιρροές για να επεξηγηθεί η προκύπτουσα μεταβλητή απόκρισης. Αυτές οι επιδράσεις ονομάζονται ‘επιρροές αλληλεπίδρασης’ και είναι γνωστές με τον όρο ‘interaction effects’. Για την πλειονότητα των παραγοντικώνπειραμάτων, ο κάθε παράγοντας έχει μόνο δύο επίπεδα. Για παράδειγμα, με δύο παράγοντες ο καθένας από τους οποίους έχει δύο επίπεδα, συνολικά προκύπτει η μελέτη τεσσάρων συνδιασμών. Σ’ αυτή την

περίπτωση έχουμε έναν 2^2παραγοντικό σχεδιασμό όπως θα αναλυθεί αργότερα. Εάν, ο αριθμός των συνδυασμών που προκύπτουν σε ένα τέτοιο πείραμα είναι τόσο μεγάλος ώστε να είναι πρακτικά αδύνατο να γίνει αποδοτική μελέτη, τότε χρησιμοποιείται ένας 'κλασματικός παραγοντικός σχεδιασμός' (fractional factorial design). Με τη βοήθεια ενός τέτοιου σχεδιασμού ένα μεγάλο μέρος των συνδιασμών (συνήθως οι μισοί τουλάχιστον από αυτούς) παραλείπονται. Παρά τα πλεονεκτήματα που αναφέρθηκαν παραπάνω, ένας παραγοντικός σχεδιασμός έχει και μειονεκτήματα. Αυτού του είδους ο σχεδιασμός παρουσιάζει δυσκολία στην επίτευξη πραγματικών αριθμητικών τιμών, εξαιτίας της αναγκαστικής μελέτης της παλινδρόμησης. Συνεπώς, ο παραγοντικός σχεδιασμός δίνει σαν αποτέλεσμα μόνο συγγενείς τιμές απόκρισης. Παρ' όλα αυτά, ο παραγοντικός σχεδιασμός είναι μια πολύ χρήσιμη πειραματική μέθοδος που χρησιμοποιείται τόσο σε εργαστηριακέςόσο και σε βιομηχανικές εφαρμογές. Η ευρεία χρήση του εξηγείται από το γεγονός ότι εξετάζει όλες τις δυνατές συνθήκες. Αξίζει να σημειωθεί πως ο παραγοντικός σχεδιασμός συνήθως χρησιμοποιείται για μικρό αριθμό παραγόντων με λίγες στάθμες ο καθένας. Αυτό συμβαίνει διότι αυτός ο σχεδιασμός οδηγεί σε μεγάλο αριθμό δοκιμών γεγονός που τον καθιστά ακριβό και χρονοβόρο σ'αυτή την περίπτωση. Επιπλέον, οι παραγοντικοί σχεδιασμοί δουλεύουν καλύτερα και είναι περισσότερο αποδοτικοί στην περίπτωση όπου έχουμε ισχυρές αλληλεπιδράσεις ανάμεσα στις μεταβλητές και κάθε μία από αυτές συνεισφέρει σημαντικά στην απόκριση.

Παραγοντικοί Σχεδιασμοί με δύο Παράγοντες

Η πλειονότητα των πειραμάτων στοχεύει συνήθως στην μελέτη των επιδράσεων δύο ή περισσότερων παραγόντων στο εξαγόμενο αποτέλεσμα. Σε αυτές τις περιπτώσεις, ως καταλληλότερο μέσο μελέτης ενδείκνυται ο παραγοντικός σχεδιασμός καθώς είναι ιδιαίτερα αποδοτικός σε τέτοιου είδους πειράματα. Οι απλούστεροι τύποι παραγοντικών σχεδιασμών είναι αυτοί με δύο παράγοντες ή αγωγές. Υπάρχουν aεπίπεδα του παράγοντα A και b επίπεδα του παράγοντα B και αυτά είναι ταξινομημένα σε έναν παραγοντικό σχεδιασμό. Δηλαδή, κάθε επανάληψη του πειράματος περιλαμβάνει όλους τους a b συνδιασμούς αγωγών. Γενικά μπορούμε να πούμε ότι χρησιμοποιούνταιn επαναλήψεις. Προκειμένου να γίνει κατανοητή η διαδικασία ενός παραγοντικού πειράματος με δύο παράγοντες θα περιγραφεί παρακάτω ένα παράδειγμα τέτοιου είδους σχεδιασμού.

2.5.2 Κλασματικοί Παραγοντικοί Σχεδιασμοί (Fractional Factorial Designs)

Μελετώντας την προηγούμενη ενότητα είναι φανερό πως οι παραγοντικοί σχεδιασμοί έχουν πολλά πλεονεκτήματα. Πρώτον, τέτοιου είδους σχεδιασμοί είναι ιδιαίτερα αποδοτικοί σε σύγκριση με πειράματα ενός παράγοντα κάθε φορά. Δεύτερον, οι παραγοντικοί σχεδιασμοί είναι απαραίτητοι όταν υπάρχουν αλληλεπιδράσεις προκειμένου να αποφύγουμε λανθασμένα αποτελέσματα. Τρίτον, επιτρέπουν την εκτίμηση των επιδράσεων ενός παράγοντα σε αρκετά επίπεδα των άλλων

παραγόντων, παρέχοντας έγκυρα αποτελέσματα πάνω σε δεδομένο αριθμό πειραματικών συνθηκών. Για όλους τους παραπάνω λόγους, οι παραγοντικοί σχεδιασμοί χρησιμοποιούνται ευρύτατα σε πολλά πειράματα. Είναι ιδιαίτερα διαδεδομένη η χρήση τους σε περιπτώσεις πειραματικών διαδικασιών όπου μας ενδιαφέρει η μελέτη πολλαπλών παραγόντων καθώς και η μεταξύ τους αλληλεπίδραση. Σημαντικό γεγονός αποτελούν αρκετές ιδιαίτερες περιπτώσεις του γενικού παραγοντικού σχεδιασμού που έχουν αξιοσημείωτη βαρύτητα. Τέτοιες είναι οι περιπτώσεις σχεδιασμών που χρησιμοποιούνται κυρίως σε ερευνητικές εργασίες και αποτελούν επίσης τη βάση άλλων σχεδιασμών με μεγάλη πρακτική αξία. Ξεχωριστή θέση ανάμεσα σ' αυτές τις περιπτώσεις πειραμάτων κατέχει η περίπτωση στην οποία έχουμε k παράγοντες καθένα σε δύο μόνο στάθμες. Οι στάθμες αυτές είναι δυνατό να είναι ποσοτικές (π.χ θερμοκρασία, κιλά, χρόνος, απόσταση, όγκος κ.τ.λ) ή ποιοτικές (π.χ 'υψηλή' στάθμη ενός παράγοντα, 'μέτρια' απόδοση κ.τ.λ.). Μια πλήρης επανάληψη ενός τέτοιου συνδιασμού απαιτεί 2*2...*2 = παρατηρήσεις και γι' αυτό το λόγο λέγεται 2^k σχεδιασμός. Ένας τέτοιος σχεδιασμός είναι ιδιαίτερα χρήσιμος στα πρωταρχικά στάδια μιας πειραματικής διαδικασίας όπου είναι απαραίτητο να εξεταστούν πολλοί παράγοντες ταυτόχρονα. Αυτού του είδους ο σχεδιασμός μας δίνει τη δυνατότητα να μελετήσουμε όλους τους παράγοντες με τον μικρότερο αριθμό εκτελέσεων (runs). Προκειμένου να γίνει πλήρως κατανοητός ένας 2^k παραγοντικός σχεδιασμός, θα πρέπει να λάβουμε υπόψη μας πως και οι k παράγοντες του πειράματος θεωρούνται αμετάβλητοι κατά τη διάρκεια της πειραματικής μελέτης. Επιπλέον, οι σχεδιασμοί αυτοί είναι τελείως τυχαιοποιημένοι και διατηρούνται όλες οι υποθέσεις της κανονικότητας. Ας μελετήσουμε για παράδειγμα, την περίπτωση ενός παραδείγματος με δύο παράγοντες Α και Β καθέναν σε δύο στάθμες. Σ' αυτή την περίπτωση λοιπόν, όπου έχουμε έναν 2^2 παραγοντικό σχεδιασμό, οι στάθμες των δύο παραγόντων μπορούν να ονομαστούν ως «χαμηλή» και «υψηλή». Στον 2^2 σχεδιασμό η χαμηλή και υψηλή στάθμη των Α και Β συμβολίζονται με '-' και '+' αντίστοιχα. Έστω ότι θέλουμε να μελετήσουμε την επίδραση της αντίστασης και της ποιότητας ενός καταλύτη στον χρόνο αντίδρασης μιας χημικής διαδικασίας. Θεωρούμε πως ο παράγοντας «Α» αντικατοπτρίζει τη συγκέντρωση της αντίστασης με δύο στάθμες, ενώ ο παράγοντας «Β» συμβολίζει τον καταλύτη με τη χαμηλή στάθμη να συμβολίζει τη χρήση μιας μόνο θήκης ενώ η υψηλή τη χρήση δύο θηκών.

Τα «Α» και «Β» αναφέρονται στις επιδράσεις των παραγόντων Α και Β ενώ με τον συμβολισμό «ΑΒ» αναφερόμαστε στην αλληλεπίδραση ΑΒ. Οι τέσσερις συνδιασμοί αγωγών στον παραπάνω σχεδιασμό παριστάνονται με μικρά γράμματα. Πιο συγκεκριμένα, η υψηλή στάθμη καθενός παράγοντα στον συνδιασμό αγωγής συμβολίζεται με το αντίστοιχο μικρό γράμμα ενώ η χαμηλή στάθμη ενός παράγοντα συμβολίζεται με την απουσία του αντίστοιχου γράμματος. Ο συμβολισμός «(1)» συνήθως χρησιμοποιείται για να δηλώσουμε ότι και οι δύο παράγοντες βρίσκονται στη χαμηλή στάθμη.Είναι ένας συμβολισμός που χρησιμοποιείται σε όλους τους 2^k σχεδιασμούς. Μπορούμε να προσδιορίσουμε τη μέση επίδραση των διαφόρων παραγόντων σαν την αλλαγή πάνω στην απόκριση από την αλλαγή στη στάθμη αυτού

του παράγοντα. Η μέση επίδραση προσδιορίζεται κατά μέσο όρο πάνω στις στάθμες του άλλου παράγοντα. Η επίδραση του Α στη χαμηλή στάθμη είναι : [a – (1)] / n και η επίδραση του Α στη υψηλή στάθμη του Β είναι :[ab – b] / n. Ο μέσος όρος των δύο τελευταίων ποσοτήτων δίνεται από την σχέση : $A = \frac{1}{2n}$ { [ab – b] + [a – (1)]}= $\frac{1}{2n}$ [ab – b+ a – (1)].

Η επίδραση του Β στη χαμηλή στάθμη του Α είναι: [b – (1)] / n. Η αντίστοιχη επίδραση πάνω στην υψηλή στάθμη του παράγοντα Β είναι : [ab – a] / n. Έτσι, ο μέσος όρος των δύο τελευταίων ποσοτήτων δίνεται από τη σχέση: $B = \frac{1}{2n}$ { [ab – a] + [b – (1)]}= $\frac{1}{2n}$ [ab + b - a – (1)]. Η αλληλεπίδραση AB ορίζεται σαν τη μέση διαφορά μεταξύ της επίδρασης του Α στην υψηλή στάθμη του Β και της επίδρασης του Α στην χαμηλή στάθμη του Β. Η αλληλεπίδραση AB δίνεται από την παρακάτω σχέση: AB = $\frac{1}{2n}$ { [ab – b] + [a – (1)]}= $\frac{1}{2n}$ [ab - b - a + (1)]. Στους 2^k παραγοντικούς σχεδιασμούς είναι συνήθης η εξέταση του μεγέθους και της κατεύθυνσης των επιδράσεων των παραγόντων. Μια τέτοια εξέταση είναι σημαντική προκειμένου να προσδιορίσουμε ποιές μεταβλητές έχουν ιδιαίτερη επιρροή στο αποτέλεσμα του πειράματος και ποιές όχι. Η μέθοδος μέσω της οποίας κάνουμε τη συγκεκριμένη μελέτη είναι η ανάλυση διασποράς. Κατά τη διάρκεια της ανάλυσης διασποράς με k = 2 χρησιμοποιούμε αθροίσματα τετραγώνων των Α, Β, ΑΒ που δίνονται από τις παρακάτω σχέσεις:

$$SSA = \frac{[ab - b + a - (1)]^2}{n*4}$$

$$SSB = \frac{[ab + b - a - (1)]^2}{n*4}$$

$$SSAB = \frac{[ab - b - a + (1)]^2}{n*4}$$

Μετά το παραπάνω παράδειγμα, μπορούμε να κάνουμε γενίκευση στην περίπτωση ενός 2^k παραγοντικού σχεδιασμού με k παράγοντες καθέναση δύο στάθμες. Το στατιστικό μοντέλο για έναν 2^k σχεδιασμό περιέχει k κύριες επιδράσεις $\binom{k}{2}$ αλληλεπιδράσεις δύο παραγόντων, $\binom{k}{3}$ αλληλεπιδράσεις τριών παραγόντων και μια αλληλεπίδραση k παραγόντων. Οι συνδιασμοί αγωγών μπορούν να γραφούν στην τυπική διάταξη εισάγοντας τους παράγοντες έναν κάθε φορά, με κάθε νέο παράγοντα να συνδιάζεται με όλους όσους προηγούνται από αυτόν. Προκειμένου να εκτιμήσουμε μια επίδραση ή να υπολογίσουμε το άθροισμα τετραγώνων για μια επίδραση, θα πρέπει πρώτα να προσδιορίσουμε την αντίθεση που αντιστοιχεί σ'αυτή την επίδραση. Η σχέση που ισχύει για τις επιδράσεις είναι:

$$Contrast_{AB..K} = (a \pm 1)\,(b \pm 1) \ldots (k \pm 1)$$

Τελικά, έχουμε τη δυνατότητα να υπολογίσουμε και τις αλληλεπίδράσεις αφού υπολογιστούν οι αντιθέσεις για τις αλληλεπιδράσεις. Οι αλληλεπιδράσεις θα υπολογιστούν βάσει των παρακάτω σχέσεων.

$$\text{AB}\ldots\text{K} = \frac{2}{n2^k}\,(\,Contrast_{AB\ldots K})$$

$$SS_{AB..K} = \frac{1}{n2^k}(Contrast_{AB\ldots K})^2$$

Σημειώνεται ότι στους παραπάνω τύπους το n δηλώνει τον αριθμό των επαναλήψεων του πειράματος.

Κλασματικοί Παραγοντικοί Σχεδιασμοί με δύο στάθμες

Οι Κλασματικοί Παραγοντικοί Σχεδιασμοί προέκυψαν από την ανάγκη μελέτης πειραμάτων με αυξημένο αριθμό παραγόντων. Συγκεκριμένα, καθώς ο αριθμός των παραγόντων σ' έναν 2^k σχεδιασμό αυξάνει, αυξάνει και ο αριθμός εκτελέσεων για μαι πλήρη επανάληψη του σχεδιασμού. Μπορούμε σαν παράδειγμα να εξετάσουμε έναν 2^6 σχεδιασμό. Σ΄αυτήν την περίπτωση απαιτούνται 64 εκτελέσεις του σχεδιασμού οι οποίες δίνουν μόνο 6 βαθμούς ελευθερίας για τις κύριες επιδράσεις και 15 για τις αλληλεπιδράσεις δύο παραγόντων. Οι υπόλοιποι 42 βαθμοί ελευθερίας αντιστοιχούν στις αλληλεπιδράσες τριών και περισσότερων παραγόντων. Συνέπεια της παραπάνω κατάστασης ήταν η εκτέλεση ενός μόνο κλάσματος του πλήρους παραγοντικού σχεδιασμού, προκειμένου να προσδιοριστούν οι κύριες επιδράσεις αλλά και οι αλληλεπιδράσεις χαμηλής τάξης. Θα πρέπει να σημειωθεί πως ορισμένες αλληλεπιδράσης υψηλής τάξης θεωρούνται αμελητέες από τον πειραματιστή. Η χρήση των κλασματικών παραγοντικών σχεδιασμών είναι ευρέως διαδεδομένη σε πολυάριθμα πρακτικά προβλήματα. Οι περιπτώσεις όμως στις οποίες είναι απαραίτητοι αυτοί οι σχεδιασμοί είναι σε πειράματα κρησαρίσματος (screening experiments). Βασικό χαρακτηριστικό αυτών των πειραμάτων είναι η συμμετοχή πολλών παραγόντων. Τέτοιου είδους πειράματα εκτελούνται στα αρχικά στάδια μιας έρευνας προκειμένου να διευκρινιστούν, αν υπάρχουν, οι παράγοντες με μεγάλες επιδράσεις στο εξαγόμενο αποτέλεσμα καθώς και εκείνοι με μηδενική επιρροή στην απόκριση. Η επιτυχής χρήση των κλασματικών παραγοντικών σχεδιασμών βασίζεται στις παρακάτω αρχές :

1.Αρχή Σποραδικότητας Επιδράσεων (The sparsity of effects principle)

Στις περιπτώσεις που στη μελέτη μας συμμετέχουν αρκετοί παράγοντες, είναι πιθανό να οδηγείται το σύστημα από κάποιες κύριες επιδράσεις ή από αλληλεπιδράσεις χαμηλής τάξης.

2.Αρχή Προβολικής Ιδιότητας (The projective property)

Οι κλασματικοί παραγοντικοί σχεδιασμοί έχουν τη δυνατότητα να προβάλλονται πάνω σε ισχυρότερους σχεδιασμούς με αντικείμενο τους παράγοντες με σημαντική επιρροή στην απόκριση.

3.Αρχή Ακολουθιακού Πειραματισμού (Sequential experimentation)

Βάσει αυτής της αρχής, μας παρέχεται η δυνατότητα να συνδυάσουμε τις εκτελέσεις δύο ή περισσότερων παραγοντικών σχεδιασμών προκειμένου να συγκεντρώσουμε ακολουθιακά έναν μεγαλύτερο σχεδιασμό. Απώτερος σκοπός αυτής της διαδικασίας είναι η εκτίμηση των επιδράσεων και των αλληλεπιδράσεων των παραγόντων που μας ενδιαφέρουν.

2.6 D – Βέλτιστοι Σχεδιασμοί

2.6.1 Εισαγωγή

Στο σχεδιασμό πειραμάτων, οι optimal (βέλτιστοι) σχεδιασμοί αποτελούν μια κλάση πειραματικών σχεδιασμών που είναι βέλτιστοι σε σχέση με κάποια στατιστικά κριτήρια. Η δημιουργία αυτού του πεδίου της στατιστικής οφείλεται στο Δανό στατιστικολόγο Kirstine Smith. Στο σχεδιασμό πειραμάτων για την εκτίμηση στατιστικών μοντέλων, οι optimalσχεδιασμοί επιτρέπουν την εκτίμηση των πειραματικών παραμέτρων χωρίς 'προκαταλήψεις' και με την ελάχιστη διακύμανση. Ένας σχεδιασμός που δεν είναι «optimal» απαιτεί μεγαλύτερο αριθμό πειραματικών εκτελέσεων για την εκτίμηση των παραμέτρων, με την ίδια ακρίβεια που το κάνει ένας optimal σχεδιασμός. Ο D-optimal σχεδιασμός είναι μια μορφή σχεδιασμού που παρέχεται από υπολογιστικούς αλγόριθμους και είναι ιδιαίτερα χρήσιμοι όταν ο αριθμός των παραμέτρων αυξάνεται. Μειώνοντας σημαντικά τον απαιτούμενο αριθμό πειραματικών επαναλήψεων, μειώνουν παράλληλα και το πειραματικό κόστος. Στον Πίνακα 2.1 φαίνεται ένα παράδειγμα του ελάχιστου αριθμού απαιτούμενων επαναλήψεων για παραγοντικό, για κλασματικό και για έναν D-optimal σχεδιασμό.

Πίνακας 2.1: Ελάχιστος αριθμός επαναλήψεων για screening designs

Factors	Full Factorial	Fractional Factorial	D-Optimal
5	32	16	16
6	64	32	28
7	128	64	35
8	256	64	43
9	512	128	52

Οι συνηθέστεροι λόγοι που οδηγούν στη χρήση D-optimal σχεδιασμών αντί των κλασικών είναι: η ανωμαλία της πειραματικής περιοχής,ηανάγκη να συμπεριληφθούν τα ήδη εκτελεσμένα πειράματα, το γεγονός πως οι ποιοτικοί παράγοντες του

πειράματος έχουν περισσότερες από δύο στάθμες, η ανάγκη μείωσης του αριθμού των πειραματικών εκτελέσεων καθώς και η ανάγκη χρήσης ιδιαίτερων μοντέλων παλινδρόμησης.

2.6.2 Η Προσέγγιση των D-βέλτιστων Σχεδιασμών

Ένας D-βέλτιστος σχεδιασμός βασίζεται σε υπολογιστικούς αλγόριθμους, χρησιμοποιώντας το καλύτερο υποσύνολο όλων των πιθανών πειραμάτων. Βασισμένοι σε ένα επιλεγμένο κριτήριο και ένα δεδομένο αριθμό πειρματικών επαναλήψεων, ο καλύτερος σχεδιασμός δημιουργείται μέσω μιας διαδικασίας επιλογής. Το υποψήφιο προς χρήση σύνολο είναι ένας πίνακας που περιέχει όλα τα πιθανά πειράματα, όπου κάθε σειρά αντιπροσωπεύει ένα πείραμα και κάθε στήλη μια μεταβλητή.Αυτός ο πίνακας έχει N γραμμές και ονομάζεται ξ_N. Ο πίνακας του σχεδιασμού ονομάζεται X, είναι ένας πίνακας διαστάσεων n×p και βασίζεται στο μοντέλο με p συντελεστές. Ο αριθμός των γραμμών n μπορεί να επιλεχθεί από τον πειραματιστή και αντιπροσωπεύει τον αριθμό των πειραμάτων στο σχεδιασμό. Με δεδομένο μοντέλο και έναν υποψήφιο πίνακα, η δόμηση του πίνακα του σχεδιασμού είναι εύκολη υπόθεση. Προκειμένου να αναλύσουμε τον τρόπο δόμησης του πίνακα σχεδιασμού X, θα χρησιμοποιήσουμε την εξίσωση που αντιστοιχεί στο γραμμικό μοντέλο με δύο παράγοντες και ένα πρόσθετο όρο αλληλεπίδρασης:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_{12} + \varepsilon \qquad (2.7).$$

Βάσει της εξίσωσης (2.7) το υποψήφιο σύνολο θα έχει δύο γραμμές και δύο στήλες. Ο ξ_N, θα είναι λοιπόν ο παρακάτω:

$$\xi_4 = \begin{bmatrix} -1 & -1 \\ -1 & 1 \\ 1 & -1 \\ 1 & 1 \end{bmatrix}.$$

Όσον αφορά στον πίνακα του σχεδιασμού, αυτός θα έχει τέσσερις γραμμές, τέσσερις στήλες και θα είναι ο πίνακας X που φαίνεται παρακάτω:

$$\mathrm{X} = \begin{bmatrix} 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & 1 & 1 \end{bmatrix}.$$

Η πρώτη στήλη του πίνακα X αντιπροσωπεύει το σταθερό όρο β_0, της εξίωσης (2.7). Η δεύτερη και η τρίτη στήλη του X, είναι οι όροι του μοντέλου που έχουμε πάρει από τον ξ_4 για τους διερευνώμενους παράγοντες x_1, x_2. Η τελευταία στήλη του X, αντιπροσωπεύει μια αλληλεπίδραση μεταξύ των δύο παραγόντων x_1, x_2. Αυτή η στήλη προκύπτει από τον πολλαπλασιασμό των δύο στηλών του ξ_4. Στην περίπτωση που έχουμε ένα μεγαλύτερο υποψήφιο σύνολο, ο αριθμός των

πιθανών υποσυνόλων των ξ_N αυξάνει και η επιλογή του πίνακα σχεδιασμού βασίζεται σε συγκεκριμένο κριτήριο. Σύμφωνα με τον de Aguiar et al.(1995, p. 202): "Ο καλύτερος συνδιασμός αυτών των σημείων ονομάζεται «βέλτιστος» και ο αντίστοιχος πίνακας σχεδιασμού καλείται «πίνακας βέλτιστου σχεδιασμού» και συμβολίζεται με «X^*»". Προκειμένου να γίνει επιλογή του καλύτερου σχεδιασμού, θα πρέπει να οριστούν δύο διαφορετικοί τύποι πινάκων. Ο πρώτος είναι ο πίνακας πληροφορίας $(X'X)$. Ο πίνακας διασποράς $(X'\,X)^{-1}$ είναι ο αντίστροφος πίνακας του $(X'X)$. Σε αυτό το σημείο, πριν αναλυθεί ο ορισμός του D-optimal σχεδιασμού, αξίζει να αναφερθεί ο ορισμός του ορθογώνιου σχεδιασμού.

Ορθογώνιος σχεδιασμός

Ένας σχεδιασμός με πίνακα σχεδιασμού X καλείται ορθογώνιος αν και μόνο αν ο $X'X$ είναι διαγώνιος πίνακας, τέτοιος ώστε $X_{.i}^{'}X_{.j} = 0, \forall\, i \neq j, i, j \in \{1, .., p+1\}$. Η εκτίμηση των ελαχίστων τετραγώνων των συντελεστών β από το μοντέλο με χαρακτηριστική εξίσωση:

$Y = X\beta + \varepsilon$, δίνεται από τη σχέση $\hat{\beta} = (X'\,X)^{-1}X'Y$. Ο αντίστοιχος πίνακας διακύμανσης είναι:Cov $(\hat{\beta}) = \sigma^2(X'\,X)^{-1}$. Στην περίπτωση ορθογώνιου σχεδιασμού η εκτίμηση της επίδρασης $\hat{\beta}_j = 1/\left\|X_{.j}\right\|_2^2 \cdot X_{.j}{}'Y$, με j = 1,..., p, εξαρτάται μόνο από τον αντίστοιχο παράγοντα $X_{.j}$. Επιπροσθέτως, οι εκτιμήσεις $\hat{\beta}_j$ είναι ανά δύο ασυσχέτιστες. Παρόλα αυτά, η σημασία των παραγόντων δύναται να εκτιμηθεί ανεξάρτητα για τον καθένα από αυτούς. Οι ορθογώνιοι παράγοντες επιρροής $X_{.j}$ δεν είναι απαραίτητα ασυσχέτιστοι επειδή οι μέσες τιμές τους δεν χρειάζεται να είναι μηδενικές.

D –optimal σχεδιασμός

Ένας σχεδιασμός με πίνακα σχεδιασμού X ονομάζεται D – optimal αν και μόνο αν ο $X'\,X$ είναι ομαλός και η det$(X'\,X)$ είναι μέγιστη στο σύνολο όλων των k × (p + 1) πινάκων των σχεδιασμών. Η τιμή D(X) = det$(X'\,X)$ καλείται D – value. Οι D – optimal σχεδιασμοί, γενικά, δεν χρειάζεται να είναι ορθογώνιοι. Η σχέση μεταξύ ορθογωνιότητας και D – optimality δίνεται από το παρακάτω αποτέλεσμα. Έστω πίνακας M. Αν οι διαγώνιες τιμές του πίνακα $M'\,M$ είναι καθορισμένες, η det$(M'\,M)$ είναι μέγιστη αν όλες οι διαγώνιες τιμές είναι μηδενικές. Αν ο τυποποιημένος πίνακας του σχεδιασμού είναι ο $X^* := (1_k, X^*)$, με

$$X^* := \left(\frac{X_{.1} - \bar{X}_{.1}}{\sqrt{\frac{1}{k}\sum_{i=1}^{k}(X_{i1} - \bar{X}_{.1})^2}}, \ldots\ldots, \frac{X_{.p} - \bar{X}_{.p}}{\sqrt{\frac{1}{k}\sum_{i=1}^{k}(X_{ip} - \bar{X}_{.p})^2}}\right),$$

οι διαγώνιες τιμές του πίνακα δεδομένων $X^{*\prime}X^*$ καθορίζονται εφόσον ισχύει $X^{*\prime}{}_{.j} \cdot X^*{}_{.j} = \left\|X^*{}_{.j}\right\|_2^2 = 1$ για j = 1,…, p + 1. Εδώ θα πρέπει να σημειωθεί ότι η D – optimality οδηγεί σε ορθογώνια και ασυσχέτιστα μεταξύ τους διανύσματα – στήλες $X^*{}_{.j}$. Η ποιότητα ενός σχεδιασμού με πίνακα σχεδιασμού X, μπορεί να εκτιμηθεί με

την έννοια της D – efficiency που ορίζεται ως : D_{eff}(X) = $\frac{D(X)^{1/p}}{D(X_{opt})^{1/p}}$, όπου X_{opt} είναι ο πίνακας σχεδιασμού ενός σχεδιασμού με μέγιστη πιθανή D –value. Σε περίπτωση που υπάρχει ένας τυποποιημένος ορθογώνιος σχεδιασμός και είναι D –optimal, η D – value θα είναι D(X^*_{opt}) = k^p. Ο πρωταρχικός σκοπός των D – optimal σχεδιασμών δεν έγκειται στη σχέση του σχεδιασμού με την ορθογωνιότητα, αλλά στην ελαχιστοποίηση της αβεβαιότητας σχετικά με τους άγνωστους συντελεστές β.

2.6.3 Κριτήρια για τον καλύτερο D – βέλτιστο Σχεδιασμό

Προκειμένου να γίνει η επιλογή του καλύτερου D – optimal σχεδιασμού, ο πειραματιστής βασίζεται σε διάφορα κριτήρια. Αξίζει να σημειωθεί πως όλα αυτά τα κριτήρια που θα αναλυθούν παρακάτω, έχουν σαν στόχο τη μεγιστοποίηση του πίνακα πληροφορίας ($X'X$).

D – Optimality

Το κριτήριο της D – optimality είναι το πιο σύνηθες κριτήριο που στοχεύει στη μεγιστοποίηση της $|X'X|$. Αυτό σημαίνει ότι ο πίνακας X^* του βέλτιστου σχεδιασμού, περιέχει τα n πειράματα που μεγιστοποιούν την ορίζουσα του $X'X$. Η σχέση $|X^{*\prime}X^*|$ = $\max_{\xi_n}(|X'X|) = \max_{\xi_n}\frac{1}{|(X'X)^{-1}|}$ δείχνει την επιλογή του κατάλληλου X^* από τους πίνακες όλων των πιθανών σχεδιασμών επιλεγμένων από το ξ_N. Αυτή η σύνδεση ανάμεσα στον πίνακα του σχεδιασμού και την ορίζουσα (Determinant), επεξηγεί την χρήση του γράμματος 'D' στον όρο «D–optimal design». Η μεγιστοποίηση της ορίζουσας του πίνακα πληροφορίας, ισοδυναμεί με την ελαχιστοποίηση της ορίζουσας του πίνακα διασποράς $(X'X)^{-1}$. Η τελευταία ισοδυναμία είναι ιδιαίτερα χρήσιμη προκειμένου να συγκρατηθούν για όσο το δυνατό λιγότερο οι τελευταίοι υπολογισμοί.

A – Optimality

Σύμφωνα με αυτό το κριτήριο ο πίνακας ενός σχεδιασμού θεωρείταιA-optimal, όταν το ίχνος του πίνακα διασποράς $(X'X)^{-1}$ γίνεται ελάχιστο. Σε αυτήτην περίπτωση, το ίχνος του τετραγωνικού πίνακα είναι το άθροισμα των στοιχείων της κύριας διαγωνίου. Η ελαχιστοποίηση του ίχνους του πίνακα διασποράς, ισοδυναμεί με την ελαχιστοποίηση της μέσης διακύμανσης των εκτιμώμενων μεταβλητών. Οι A–optimal σχεδιασμοί χρησιμοποιούνται σπάνια, διότι είναι υπολογιστικά περισσότερο δύσκολο να 'ενημερώνονται' με τα νέα δεδομένα κατά τη διάρκεια της διαδικασίας επιλογής. Η παρακάτω σχέση αντιπροσωπεύει τους στόχους αυτού του κριτηρίου :

Trace $(X^{*'}X^*)^{-1}$ = $\min_{\xi_n}$(trace $(X'X)^{-1}$) , όπουtrace $(X'X)^{-1}$ = $\sum_{i=1}^{p} c_{ii}$

V – Optimality (Average Prediction Variance)

Σύμφωνα με τον de Aguiar et al. (1995, p. 203) 'Η συνάρτηση της διακύμανσης είναι ένα μέτρο της αβεβαιότητας της προβλεπόμενης απόκρισης'. Αυτή η διακύμανση για ένανx_iμπορεί να υπολογιστεί με την εξίσωση : d $(x_i) = x_i'*(X'X)^{-1} * x_i$, όπου το x_i ισούται με ένα διάνυσμα που περιγράφει ένα μοναδικό πείραμα και το x_i' αντιπροσωπεύει τη μετατόπιση του διανύσματος αυτού. Οι στόχοι αυτού του κριτηρίου φαίνονται στην παρακάτω σχέση:

$\frac{1}{n}\sum_{i=1}^{n} x_i' * (X^{*\prime}X^*)^{-1} * x_i = \min_{\xi_n}(\frac{1}{n}\sum_{i=1}^{n} x_i' * (X'X)^{-1} * x_i)$.

G – Optimality (Maximum Prediction Variance)

Αυτό το κριτήριο ασχολείται με τη διακύμανση της πρόγνωσης των υποψήφιων σημείων. Ο πίνακας του βέλτιστου σχεδιασμού, επιλέγεται έτσι ώστε να ελαχιστοποιεί την μέγιστη διακύμανση της πρόγνωσης στο σχεδιασμό. Η σχέση που δείχνει τους στόχους αυτού του κριτηρίου φαίνεται παρακάτω:

$\max\ (x_i' * (X^{*\prime}X^*)^{-1} * x_{i)} = \min_{\xi_n} (\max\ x_i' * (X'X)^{-1} * x_i)$.

G – Efficiency

Στις περισσότερες περιπτώσεις, το G κριτήριο δεν χρησιμοποιείται για να βρει τον καλύτερο σχεδιασμό κατά τη διάρκεια της διαδικασίας επιλογής. Συνήθως εφαρμόζεται για την επιλογή ανάμεσα σε όμοιους σχεδιασμούς οι οποίοι δημιουργήθηκαν με κάποιο άλλο κριτήριο. Η G – efficiency προσδιορίζεται από τη σχέση : G_{eff} = 100% * $(\frac{p}{n* d_{max}(\chi)})$, όπου το p είναι ο αριθμός των όρων ή των συντελεστών του μοντέλου, n είναι ο αριθμός των πειραματικών επαναλήψεων και το $d_{max}(\chi)$ αναφέρεται στη μέγιστη διακύμανση της πρόγνωσης στο μοντέλο με πίνακα σχεδιασμού X. Η G – Efficiency μπορεί να θεωρηθεί ως σύγκριση ενός D–optimal σχεδιασμού με έναν κλασματικό παραγοντικό σχεδιασμό. Κατ' επέκταση, η αποδοτικότητα (efficiency) δίνεται σε ποσοστιαία κλίμακα.

Δείκτης Κατάστασης

Ο δείκτης κατάστασης (Condition Number-CN) είναι ένα εκτιμητικό κριτήριο όπως το κριτήριο της G- Efficiency και χρησιμοποιείται για να εκτιμήσει έναν ήδη δομημένο D-βέλτιστο σχεδιασμό. Το συγκεκριμένο κριτήριο αξιολογεί τη σφαιρικότητα και τη συμμετρία ενός D-βέλτιστου σχεδιασμού υπολογίζοντας το λόγο μεταξύ της μέγιστης και της ελάχιστης τιμής του X. Ένας σχεδιασμός με δείκτη κατάστασης ίσο με τη μονάδα θα είναι ορθογώνιος ενώ ένας αυξημένος δείκτης κατάστασης υποδεικνύει ένα λιγότερο ορθογώνιο σχεδιασμό.

2.7 Αριθμός επαναλήψεων σχεδιασμού

Προκειμένου να προσαρμοστεί το κατάλληλο κριτήριο σχεδιασμού, θα πρέπει να καθοριστεί ο αριθμός των πειραμάτων που θέλουμε να υπάρχουν στο σχεδιασμό. Η επιλογή αυτού του παράγοντα ηείναι ιδιαίτερα σημαντική διότι μια αλλαγή στον αριθμό πραγματοποιήσεων (runs) του σχεδιασμού έχει τη δυνατότητα να αλλάξει τον πίνακα του μοντέλου και να επιλεχθεί, τελικά, άλλος βέλτιστος σχεδιασμός. Δεν υπάρχουν κανόνες που να υποδεικνύουν τον καθορισμό αυτού του αριθμού. Ωστόσο, ο ελάχιστος αριθμός πραγματοποιήσεων είναι άμεσα εξαρτημένος από το μοντέλο. Ένα μοντέλο με p συνιστώσες δύναται να ερευνηθεί μόνο με ένα D–βέλτιστο σχεδιασμό ο οποίος έχει τουλάχιστον p πραγματοποιήσεις. Στις περισσότερες περιπτώσεις φαίνεται ιδιαίτερα χρήσιμη η δημιουργία διαφορετικών σχεδιασμών οι οποίοι διαφέρουν στον αριθμό των πραγματοποιήσεων και συγκρίνουν την αποδοτικότητα των σχεδιασμών. Ένας σχεδιασμός με ελάχιστες περισσότερες ή λιγότερες πραγματοποιήσεις συγκριτικά με τον επιθυμητό αριθμό πραγματοποιήσεων, μπορεί να έχει μεγαλύτερο παράγοντα απόφασης και έτσι να γίνεται ο καλύτερος σχεδιασμός προς παρουσίαση.

2.8 Ο BayesianΜετασχηματισμός

Οι D–βέλτιστοι σχεδιασμοί χρησιμοποιούν διαφορετικά κριτήρια προκειμένου να επιλεγεί ο καλύτερος σχεδιασμός μέσα σε μια δεξαμενή όλων των δυνατών συνδιασμών παραγόντων. Για όλα αυτά τα κριτήρια, η διαδικασία επιλογής είναι ισχυρά εξαρτημένη από το μοντέλο και συνεπώς ο πειραματιστής θα πρέπει να δώσει ιδιαίτερη προσοχή στην επιλογή του μοντέλου. Η αλλαγή μιας ή περισσότερων συνιστωσών δημιουργεί ένα διαφορετικό πίνακα μοντέλου ο οποίος με τη σειρά του οδηγεί σε άλλο βέλτιστο σχεδιασμό.

2.8.1 Η προσθήκη δυναμικών όρων

Η εξάρτηση από το υποτιθέμενο μοντέλο μπορεί να είναι πρόβλημα στην περίπτωση που το μοντέλο δεν έχει επιλεγεί προσεκτικά. Ένας Bayesian μετασχηματισμός μειώνει την επίδραση του μοντέλου με την προσθήκη πρόσθετων δυναμικών όρων. Θεωρώντας ένα μοντέλο με p πρωταρχικούς όρους, προσθέτουμε q δυναμικούς όρους και ο πίνακας X του μοντέλου έχει την ακόλουθη μορφή :

$$X = (X_{pri}/X_{pot}) \qquad (2.8).$$

Οι πρόσθετοι όροι συνήθως δεν περιλαμβάνονται στο μοντέλο και έχουν μεγαλύτερο βαθμό σε σχέση με τους πρωταρχικούς όρους. Θα πρέπει να σημειωθεί πως η επιλογή των δυναμικών όρων εξαρτάται από το πρόβλημα.

2.8.2 Ιεράρχηση των παραγόντων

Προκειμένου να υπάρξει όφελος από τον Bayesian μετασχηματισμό, οι πρωταρχικοί και οι δυναμικοί όροι θα πρέπει να ιεραρχηθούν με διαφορετικούς τρόπους. Καθένας από τους μη σταθερούς πρωταρχικούς όρους έχει εύρος 2 και ποικίλει από -1 έως 1. Υπό φυσιολογικές συνθήκες, αυτοί οι όροι ιεραρχούνται και κεντράρονται με ορθογώνια ιεράρχιση:

$$\max(X_{pri}) = 1, \min(X_{pri}) = -1 \quad (2.9).$$

Για τον Bayesian μετασχηματισμό ορίζεται ότι οι δυναμικοί όροι ακολουθούν την ακόλουθη συνθήκη :

$$\max(X_{pot}) - \min(X_{pot}) = 1 \quad (2.10)$$

όπου η μέγιστη και η ελάχιστη τιμή λαμβάνονται από το σύνολο των υποψήφιων σημείων για τον σχεδιασμό. Η ιεράρχηση των δυναμικών όρων επιτυγχάνεται με την παρουσίαση μιας παλινδρόμησης των δυναμικών όρων πάνω στους πρωταρχικούς χρησιμοποιώντας υποψήφια σημεία. Η διαδικασία αυτή οδηγεί στον υπολογισμό των 'a' και 'Z' (Du Mouchel & Jones, 1994) :

$$\mathrm{a} = (X'_{pri}X_{pri})^{-1} * X'_{pri}X_{pot} \quad (2.11)$$

$$\mathrm{R} = X_{pot} - X_{pri} * \mathrm{a} \quad (2.12)$$

$$\mathrm{Z} = \frac{R}{\max(R) - \min(R)} \quad (2.13)$$

Σε αυτή την περίπτωση, το a είναι ο συντελεστής ελαχίστων τετραγώνων του X_{pot} πάνω στο X_{pri} και το R αναπαριστά το υπόλοιπο αυτής της παλινδρόμησης. Οι χρησιμοποιούμενες μέγιστες και ελάχιστες τιμές λαμβάνονται κατά στήλη από το συνολικό υποψήφιο σύνολο. Ο πίνακας Z που λαμβάνεται σαν αποτέλεσμα είναι η ιεραρχημένη και κεντραρισμένη έκδοση των δυναμικών όρων και χρησιμοποιείται στον πίνακα του μοντέλου στη θέση του X_{pot} :

$$\mathrm{X} = (X_{pri}/Z) \quad (2.14).$$

Προκειμένου να γίνει χρήση του πίνακα μοντέλου του Bayesian μετασχηματισμού με φυσικά κριτήρια D – βέλτιστου σχεδιασμού , χρειάζεται η ενημέρωση του πίνακα διασποράς $(X'X)^{-1}$ προσθέτοντας έναν επιπλέον πίνακα. Αυτός ο πίνακας Q είναι ένας $(p+q) \times (p+q)$ διαγώνιος πίνακας με μηδενικά στη θέση των πρώτων

διαγώνιων στοιχείων και μονάδες στη θέση των τελευταίων q διαγώνιων στοιχείων. Βάσει μιας επιλεγμένης τ τιμής, μετασχηματίζουμε τον πίνακα διασποράς ως εξής :

$$(X'X + \frac{Q}{\tau^2})^{-1} \tag{2.15}$$

Κεφάλαιο 3

Επιλογή Μεταβλητών και Ταξινόμηση

3.1 Εισαγωγή

Το πρόβλημα της επιλογής μεταβλητών (variable selection problem) είναι ένα από τα περισσότερο διαδεδομένα προβλήματα επιλογής μοντέλων σε στατιστικές εφαρμογές.

Η επιλογή μεταβλητών έχει γίνει το επίκεντρο πολλών ερευνών στην περιοχή εφαρμογών, για τις οποίες οι διαθέσιμες βάσεις δεδομένων αποτελούνται από δεκάδες ή εκατοντάδες μεταβλητές. Οι εφαρμογές αυτές περιλαμβάνουν την επεξεργασία κειμένου εγγράφων που παρουσιάζουν ενδιαφέρον, τη γονιδιακή ανάλυση του γενετικού υλικού και τη συνδιαστική χημεία. Η επιλογή των μεταβλητών έχει σαν σκοπό την ανάδειξη του βέλτιστου υποσυνόλου επεξηγηματικών ή προγνωστικών μεταβλητών. Πιο συγκεκριμένα, στόχος είναι η ανάλυση των δεδομένων με τον απλούστερο τρόπο. Συνεπώς, οι περιττές προγνωστικές μεταβλητές θα πρέπει να απομακρυνθούν. Σύμφωνα με τον William of Ockham (1287 – 1347), Άγγλο μοναχό και φιλόσοφο: «Among competing hypotheses, the one with the fewest assumptions should be selected». Σύμφωνα με αυτή την αρχή, μεταξύ διάφορων πιθανών επεξηγήσεων ενός φαινομένου, η πιο απλή από αυτές είναι και η καλύτερη. Αν αυτή η αρχή εφαρμοστεί στην ανάλυση δεδομένων, τότε συμπεραίνουμε πως το απλούστερο μοντέλο δεδομένων είναι και το καλύτερο. Η έρευνα στην επιλογή μεταβλητών ξεκίνησε στις αρχές της δεκαετίας του 1960 με την εργασία των Lewis and Sebestyen το 1962. Η έρευνα πάνω στο θέμα εμπλουτίστηκε αργότερα με την δημοσίευση των ερευνών καθώς και των Kohani and John την ίδια χρονιά. Τα τελευταία χρόνια, δεκάδες επιστημονικές εργασίες ασχολούνται με ερευνητικούς τομείς που περιλαμβάνουν εκατοντάδες ή και χιλιάδες διαθέσιμες μεταβλητές. Ο μεγαλύτερος όγκος της έρευνας συνδέεται με τους τομείς της ιατρικής και της βιολογίας.

Η επιλογή του βέλτιστου υποσυνόλου μεταβλητών για την δόμηση των προγνωστικών μεταβλητών δεν είναι μια απλή υπόθεση, καθώς ο αριθμός των υποσυνόλων που πρέπει να ελεγθούν αυξάνεται εκθετικά με τον αριθμό των υποψήφιων μεταβλητών.Ακόμη και με ένα μέτριο αριθμό υποψήφιων μεταβλητών,δεν είναι δυνατό να αξιολογηθούν όλα τα πιθανά υποσύνολα, γεγονός που δηλώνει πως αποτελεί ένα δύσκολο υπολογιστικό πρόβλημα. Αυτό σημαίνει ότι όταν το μέγεθος του προβλήματος είναι μεγάλο, η εύρεση της βέλτιστης επιλογής πρακτικά δεν είναι εφικτή.

Για την αντιμετώπιση των προβλημάτων επιλογής μεταβλητών έχουν αναπτυχθεί δύο διαφορετικές μεθοδολογικές προσεγγίσεις:

- Οι βέλτιστες ή ακριβείς τεχνικές, οι οποίες μπορούν να εγγυηθούν τη βέλτιστη λύση, αλλά είναι εφαρμόσιμες μόνο σε μικρά σύνολα.
- Οι ευρετικές τεχνικές,οι οποίες μπορούν να βρουν καλές λύσεις σε λογικό χρονικό διάστημα, παρόλο που δεν μπορούν να εγγυηθούν το βέλτιστο αποτέλεσμα.

Μια από τις πιο γνωστές ευρετικές τεχνικές είναι ο αλγόριθμος των Narendra and Fukunaga (1977) , αλλά όπως αποδείχθηκε το 1997 από τη μελέτη των Jain and Zongker, αυτός ο αλγόριθμος δεν είναι πρακτικός για προβλήματα με μεγάλο μέγεθος δεδομένων.

Η επιλογή μεταβλητών παίζει πάρα πολύ σημαντικό ρόλο στην ταξινόμηση (classification). Πριν ξεκινήσει ο σχεδιασμός μιας μεθόδου ταξινόμησης, όταν στην μελέτη εμπλέκονται πολλές μεταβλητές, θα πρέπει να επιλεχθούν μόνο εκείνες οι μεταβλητές που θα δώσουν τις απαιτούμενες πληροφορίες. Αυτό είναι το πρώτο βήμα προς τον περιορισμό του μεγάλου μεγέθους μεταβλητών που οδηγεί στις πιο σημαντικές πληροφορίες. Στη συνέχεια γίνεται επιλογή ενός μόνο υποσυνόλου μεταβλητών που θα χρησιμοποιηθεί στη θέση του αρχικού δυσχρηστού συνόλου. Σύμφωνα με μελέτη του Reunanen (2003) ,ισχύουν τέσσερις βασικές αρχές που αφορούν την ταξινόμηση:

- Είναι πολύ πιο οικονομική η μελέτη ενός μόνο υποσυνόλου μεταβλητών.
- Η ακρίβεια της πρόγνωσης είναι δυνατό να βελτιωθεί αισθητά με τον αποκλεισμό των περιττών ή των ασυσχέτιστων μεταβλητών.
- Είναι ευκολότερο και γρηγορότερο να δομηθούν μεταβλητές πρόγνωσης όταν χρησιμοποιούνται πολύ λιγότερες μεταβλητές.
- Η γνώση των σχετικών μεταβλητών μπορεί να διελευκάνει το πρόβλημα της πρόγνωσης και να οδηγήσει στην δόμηση του τελικού μοντέλου ταξινόμησης.

Ο σκοπός του προβλήματος της ταξινόμησης είναι η κατάταξη δεδομένων που χαρακτηρίζονται από ιδιότητες ή μεταβλητές. Αυτό σημαίνει ότι θα πρέπει να ληφθεί απόφαση σχετικά με την κλάση στην οποία ανήκει κάθε δεδομένο. Με βάση ένα σύνολο δεδομένων, των οποίων γνωρίζουμε την κλάση, σχεδιάζεται ένα σύνολο κανόνων το οποίο γενικεύεται ώστε να κατατάξει το σύνολο των αρχικών δεδομένων με την μέγιστη δυνατή ακρίβεια. Υπάρχουν διάφορες μεθοδολογίες για την αντιμετώπιση του προβλήματος της ταξινόμησης. Μερικές από αυτές τις μεθοδολογίες είναι: classic discriminant analysis, logistic regression, neural networks, decision trees, instance–based learning κ.τ.λ. Οι μέθοδοι της γραμμικής ανάλυσης και της λογιστικής πάλινδρόμησης αναζητούν γραμμικές συναρτήσεις και στην συνέχεια τις χρησιμοποιούν για σκοπούς ταξινόμησης. Η χρήση γραμμικών συναρτήσεων προσφέρει τη δυνατότητα καλύτερης ερμηνείας των αποτελεσμάτων, μέσω ανάλυσης της αξίας των αποκτηθέντων συντελεστών. Φυσικά, δεν ταιριάζει κάθε μέθοδος ταξινόμησης σε αυτόν τον τύπο ανάλυσης αλλά παρόλα αυτά η κλασσική ανάλυση εξακολουθεί να είναι από τις πιο ενδιαφέρουσες μεθοδολογίες.

3.2 Επιλογή υποσυνόλου μεταβλητών

Το πρόβλημα επιλογής μεταβλητών μπορεί να αντιμετωπιστεί με την περιγραφή κάποιων βασικών κατευθύνσεων. Οι κατευθύνσεις αυτές συνοψίζουν τη χρησιμότητα επιλογής υποσυνόλων μεταβλητών οι οποίες μαζί μπορούν να έχουν μεγάλη δύναμη πρόγνωσης σε αντίθεση με την κατάταξη μεταβλητών βάσει της ατομικής τους προγνωστικής δύναμης. Οι βασικές κατευθύνσεις–μέθοδοι διαχωρίζονται σε wrappers μεθόδους, ενσωματωμένες μεθόδους (embedded methods), και φίλτρα (filters). Οι wrappers χρησιμοποιούν τη μηχανή μάθησης που μας ενδιαφέρει σαν μαύρο κουτί προκειμένου να πετύχουν υποσύνολα μεταβλητών σύμφωνα με την προγνωστική τους δύναμη.Οι embedded μέθοδοι παρουσιάζουν την επιλογή μεταβλητών κατά την διαδικασία της επεξεργασίας και συνήθως είναι συγκεκριμένες σύμφωνα με την επιλεγμένη μηχανή μάθησης. Τέλος, τα φίλτρα επιλέγουν υποσύνολα σαν ένα στάδιο προ–επεξεργασίας ανεξάρτητα από τον επιλεγμένο προγνώστη.

3.2.1 Wrapper – embedded μέθοδοι και φίλτρα

Η wrapper μεθοδολογία προσφέρει έναν απλό και δυνατό τρόπο ώστε να διευθετηθεί το πρόβλημα επιλογής μεταβλητών ανεξάρτητα από την επιλεγμένη μηχανή μάθησης. Στην πραγματικότητα, η μηχανή μάθησης θεωρείται μαύρο κουτί και η μέθοδος αυτή χρησιμοποιείται από τα πακέτα λογισμικού της μηχανής. Σε ένα γενικότερο πλαίσιο, αυτή η μεθοδολογία συνίσταται στη χρήση μιας παρουσίασης πρόγνωσης προκειμένου να εκτιμηθεί η σχετική χρησιμότητα των υποσυνόλων των μεταβλητών. Πρακτικά, η μέθοδος χρησιμοποιείται με σκοπό να ερευνηθεί ο χώρος όλων των δυνατών υποσυνόλων, να προσδιοριστεί ο τρόπος εκτίμησης της προγνωστικής δύναμης της μηχανής μάθησης και τέλος να προδιοριστούν οι συγκεκριμένοι προγνώστες που θα χρησιμοποιηθούν. Η συγκεκριμένη μεθοδολογία συχνά κριτικάρεται γιατί φαίνεται να απαιτούνται τεράστια ποσά υπολογισμών αν και δεν είναι απαραίτητο. Βάσει αυτών των μεθοδολογιών μπορούν να σχεδιαστούν ιδιαίτερα αποδοτικές στρατηγικές έρευνας χωρίς αυτό να σημαίνει ότι θυσιάζεται η προγνωστική παρουσίαση. Στην πραγματικότητα, φαίνεται σε κάποιες περιπτώσεις να συμβαίνει το αντίστροφο. Αυτό σημαίνει πως κάποιες στρατηγικές μπορεί να ελαφρύνουν το πρόβλημα της υπερπροσαρμογής ενώ άλλες φαίνεται να είναι ιδιαίτερα υπολογιστικά σύμφερουσες και εύρωστες απέναντι στην υπερπροσαρμογή. Στρατηγικές με τα χαρακτηριστικά των τελευταίων διακρίνονται στη forward selection και τη backward elimination. Στη forward selection, οι μεταβλητές ενσωματώνονται σταδιακά σε όλο και μεγαλύτερα υποσύνολα, ενώ στην backward elimination ξεκινάμε με το σύνολο όλων των μεταβλητών, και σταδιακά εξαλείφουμε τις λιγότερα υποσχόμενες. Και οι δύο μέθοδοι συμπεριλαμβάνονται στις nested subsets. Χρησιμοποιώντας τη μηχανή μάθησης σαν μαύρο κουτί, οι wrappers είναι σημαντικά απλές. Παρόλα αυτά, οι embedded μέθοδοι που ενσωματώνουν την επιλογή των μεταβλητών σαν κομμάτι της επεξεργασίας, μπορεί να είναι πιο

αποτελεσματικές. Αυτές οι μέθοδοι κάνουν καλύτερη χρήση των διαθέσιμων δεδομένων και φτάνουν σε λύση γρηγορότερα.

Όσον αφορά στα φίλτρα, υποστηρίζεται πως είναι πολύ γρηγορότερα σε σχέση με τις wrapper μεθόδους. Ακόμη και πρόσφατα ανεπτυγμένες embedded μέθοδοι ανταγωνίζονται τα φίλτρα σε αυτό το κομμάτι. Πολλοί, επίσης, πιστεύουν πως μερικά φίλτρα παρέχουν γενική επιλογή μεταβλητών,η οποία δεν εξαρτάται από την επιλεγμένη μηχανή μάθησης. Ενδιαφέρουσα άποψη είναι και εκείνη που δηλώνει πως η χρήση φίλτρων μπορεί να χρησιμοποιηθεί σαν στάδιο προ–επεξεργασίας προκειμένου να μειωθούν οι διαστάσεις του χώρου και να ξεπεραστεί το πρόβλημα της υπερπροσαρμογής. Εν προκειμένω, φαίνεται λογικό να χρησιμοποιούμε wrapperή embedded μεθόδους με γραμμικούς προγνώστες σαν φίλτρο και μετά να μπαίνουμε στην επεξεργασία με έναν πιο σύνθετο προγνώστη στις προκύπτουσες μεταβλητές. Η πολυπλοκότητα των γραμμικών φίλτρων μπορεί να επεκταθεί με την προσθήκη στη διαδικασία επιλογής προϊόντων των εισαγώμενων μεταβλητών (μονωνύμων ή πολυωνύμων), διατηρώντας τις μεταβλητές που είναι μέρος όλων των επιλεγμένων μονωνύμων. Σε μερικές περιπτώσεις, μπορεί βέβαια κάποιος να θελήσει να μειώσει την πολυπλοκότητα των γραμμικών φίλτρων προκειμένου να προσπεράσει τα προβληματαυπερπροσαρμογής. Όταν ο αριθμός των παραδειγμάτων είναι μικρός σε σύγκριση με τον αριθμό των μεταβλητών, μπορεί κάποιος να χρειάζεται να προσφύγει στην επιλογή με συντελεστές συσχέτισης.

3.2.2 Nested subset μέθοδοι

Στο χώρο των nested subset μεθόδων συμπεριλαμβάνονται κάποιες embedded μέθοδοι συνδιασμένες με forward selection και backward elimination. Προκειμένου να κατανοήσουμε το ρόλο αυτών των μεθόδων ας ονομάσουμε ‘n’ τον αριθμό των αντικειμενικής συνάρτησης που χρησιμοποιεί τέτοιο υποσύνολο μεταβλητών. Η πρόγνωση της αλλαγής στην αντικειμενική συνάρτηση μπορεί να προκύψει από τον υπολογισμό πεπερασμένων διαφορών. Συγκεκριμένα, η διαφορά μεταξύ F(n) και F(n+1) ή F(n-1) υπολογίζεται για τις μεταβλητές που είναι υποψήφιες για προσθήκη ή εξάλειψη. Η πρόγνωση της αλλαγής μπορεί επίσης να επιτευχθεί με τετραγωνική προσέγγιση της συνάρτησης κόστους. Αυτή η μέθοδος προτάθηκε ώστε να περικοπούν οι συντελεστές βαρύτητας στα νευρωνικά δίκτυα. Μπορεί να χρησιμοποιηθεί για backward elimination, μέσω της περικοπής των εισαγώμενων μεταβλητών βαρύτητας w_i . Πραγματοποιείται επέκταση Taylor δεύτερης τάξης. Στο βέλτιστο της F , ο όρος πρώτης τάξης μπορεί να παραλειφθεί για την μεταβλητή i στη διαφορά

$$DF_i = \frac{1}{2}\frac{\partial^2 F}{\partial w_i^2}(Dw_i)^2 \qquad (3.1).$$

Η αλλαγή στο βάρος $Dw_i = w_i$ανιστοιχεί σε αφαίρεση της μεταβλητής i.

Μερικοί αλγόριθμοι χρησιμοποιούν πεπερασμένες διαφορές διότι οι ακριβείς διαφορές μπορούν να υπολογιστούν αποτελεσματικά, χωρίς τη χρήση νέων μοντέλων για κάθε υποψήφια μεταβλητή. Τέτοια είναι η περίπτωση του γραμμικού μοντέλου ελαχίστων τετραγώνων. Συγκεκριμένα, η διαδικασία ορθογωνιοποίησης Gram–Schmidt επιτρέπει την παρουσίαση επιλογής με forward selection προσθέτοντας σε κάθε βήμα την μεταβλητή που μειώνει περισσότερο το μέσο τετραγωνικό σφάλμα. Για άλλους αλγόριθμους, όπως οι kernel methods, μπορούν να υπολογιστούν προσεγγίσεις της διαφοράς.Οι kernel methods είναι μηχανές μάθησης της μορφής :

$$f(\vec{x}) = \sum_{k=1}^{m} a_k \, K(\overrightarrow{x},\ \vec{x}_k) \qquad (3.2),$$

όπου K είναι η συνάρτηση kernel που μετράει την ομοιότητα μεταξύ $\vec{x}$ και $\vec{x}_k$. Η απόκλιση στην F(n) υπολογίζεται διατηρώντας σταθερές τις a_k τιμές.

Στην περίπτωση γραμμικών προγνωστών $f(\vec{x}) = \vec{w}\vec{x}$ + b, χρησιμοποιείται η διαδικασία OBD (optimum brain damage). Οι δημιουργοί αυτού του αλγορίθμου χρησιμοποιούν $D F_i$στη θέση των συντελεστών βαρύτητας $|w_i|$σαν κριτήριο λήξης του αλγορίθμου. Παρόλα αυτά, για γραμμικούς προγνώστες με μια αντικειμενική συνάρτηση F όπου έχει w_i^2 αντί για w_i, τα δύο κριτήρια είναι ισοδύναμα.

3.3 Μέθοδοι για επιλογή μεταβλητών

Τα QSAR (Quantitive structure–activity relationship) είναι μοντέλα παλινδρόμησης ή ταξινόμησης που χρησιμοποιούνται στη χημεία, τη βιολογία και τη μηχανική. Όπως και τα υπόλοιπα μοντέλα παλινδρόμησης, αυτά τα μοντέλα συσχετίζουν ένα σύνολο μεταβλητών (X) με την ισχύ της μεταβλητής απόκρισης (Y), ενώ τα μοντέλα ταξινόμησης QSAR συσχετίζουν το X με μια αποκλειστική τιμή της μεταβλητής απόκρισης. Στην μοντελοποίηση QSAR, οι μεταβλητές πρόγνωσης αποτελούνται από φυσικοχημικές ιδιότητες ή θεωρητικές μοριακές περιγραφές των χημικών ουσιών. Η μεταβλητή απόκρισης στην μοντελοποίηση QSAR μπορεί να είναι βιολογική δραστηριότητα των χημικών ουσιών. Αυτά τα μοντέλα, αρχικά συνοψίζουν την πιθανή σχέση μεταξύ χημικής δομής και βιολογικής δραστηριότητας σε ένα σύνολο δεδομένων των χημικών ουσιών. Δεύτερον, τα QSAR μοντέλα προβλέπουν τη δραστηριότητα νέων χημικών ουσιών μέσω του παρακάτωμαθηματικού μοντέλου:

Δραστηριότητα = f (φυσικοχημικές ιδιότητες και/ ή δομικές ιδιότητες) + σφάλμα

Το σφάλμα που αναφέρεται παραπάνω περιλαμβάνει το σφάλμα του μοντέλου και τη μεταβλητότητα των παρατηρήσεων.

Με σκοπό να αναπτυχθούν μοντέλα παλινδρόμησης / ταξινόμησης, η QSAR ανάλυση τυπικά χρησιμοποιεί μοριακούς περιγραφείς σαν ανεξάρτητες μεταβλητές. Ο αριθμός των μεταβλητών έχει αυξηθεί ραγδαία και στις μέρες μας χιλιάδες μεταβλητές, ικανές να περιγράψουν διαφορετικές οπτικές ενός μορίου, μπορούν να υπολογιστούν μέσω

λογισμικού. Παρόλα αυτά, κατά τη μοντελοποίηση μιας συγκεκριμένης ιδιότητας μιας βιολογικής δραστηριότητας, είναι λογικό να υποθέσουμε ότι ένας μικρός μόνο αριθμός μεταβλητών, είναι πραγματικά συνδεδεμένος με την πειραματική απόκριση. Σαν συνέπεια, ένα βήμα κλειδί είναι η επιλογή του βέλτιστου υποσυνόλου μεταβλητών για την ανάπτυξη του μοντέλου. Αυτός ακριβώς είναι και ο σκοπός των μεθόδων επιλογής οι οποίες επιτρέπουν την πραγματοποιήση των παρακάτω σκοπών:

- Βελτίωση απόκρισης μέσω χρήσης απλών μεθόδων
- Απόρριψη ασήμαντων επιδράσεων, δηλαδή μείωση θορύβου
- Ανάπτυξη της ικανότητας πρόγνωσης του μοντέλου
- Επιτάχυνση του χρόνου μοντελοποίησης.

Κατά την πάροδο των χρόνων, πολλές διαφορετικές μέθοδοι έχουν παρουσιαστεί, από σχετικά απλές σε πιο πρόσφατες που έχουν εμπνευστεί από διαφορετικά επιστημονικά πεδία, όπως η γενετική. Επιπλέον, μερικές μέθοδοι μπορούν να παρουσιάσουν παλινδρόμηση και επιλογή μεταβλητών ταυτόχρονα.

3.3.1 All subset μοντέλα (ASM)

Η μέθοδος ASM είναι η πιο απλή υπολογιστική μέθοδος. Η μέθοδος αυτή συνίσταται στην παραγωγή όλων των δυνατών συνδιασμών των pμεταβλητών, μεγέθους 1-p, όπου p είναι ο συνολικός αριθμός των μεταβλητών. Αυτή η μέθοδος εγγυάται ότι μπορεί να βρεθεί το καλύτερο υποσύνολο των μεταβλητών. Ο συνολικός αριθμός των συνδιασμών των p μεταβλητών είναι 2^p - 1. Σαν συνέπεια, η μέθοδος αυτή γίνεται ακατάλληλη για μεγάλους αριθμούς μεταβλητών. Το αξιοσημείωτο στην ανάπτυξη απλών μεθόδων, για παράδειγμα μοντέλα που περιλαμβάνουν περιορισμένο αριθμό k μεταβλητών, μπορεί κάποιος να υπολογίσει όλους τους πιθανούς συνδιασμούς των pμεταβλητών μέχρι και ένα ανώτερο όριο k, οι θετικές επιρροές μεταφράζονται ευκολότερα σε λιγότερο χρόνο. Σε αυτή την περίπτωση, των p μεταβλητών, ο συνολικός αριθμός των μοντέλων t, από το μέγεθος 1 έως k, δίνεται από τη σχέση

$$t = \sum_k \frac{p!}{k!(p-k)!} \leq 2^p - 1 \qquad (3.3)$$

Ο συνολικός αριθμός των μοντέλων που δημιουργούνται είναι μικρότερος συγκριτικά με την περίπτωση όπου k = p, αλλά είναι ακόμα πολύ μεγάλος όταν ο αριθμός p των μεταβλητών είναι μεγάλος.

3.3.2 Ακολουθιακή έρευνα (Sequential Search (SS))

Η μέθοδος Sequential Search είναι μια απλή μέθοδος που έχει σαν σκοπό να βρίσκει τα καλύτερα υποσύνολα μεταβλητών για ένα καθορισμένο μέγεθος μοντέλου. Η βασική ιδέα είναι η επαναλαμβανόμενη αντικατάσταση κάθε μεταβλητής με όλες τις μεταβλητές (remaining variables) προκειμένου να διευκρινιστεί η πιθανότητα απόκτησης ενός καλύτερου μοντέλου. Αυτή η διαδικασία διαφέρει από τη μέθοδο των All subset models, επειδή σε αυτή την περίπτωση δεν ελέγχονται όλοι οι πιθανοί συνδιασμοί των p μεταβλητών. Το πλεονέκτημα αυτής της μεθόδου είναι ότι είναι λιγότερο χρονοβόρα. Ο αρχικός πληθυσμός συνήθως δημιουργείται τυχαία, βάσει περιορισμών στον αριθμό των μεταβλητών για κάθε μοντέλο. Όλες οι μεταβλητές αντικαθιστώνται με άλλες κι έτσι το νέο μοντέλο επιλέγεται μόνο αφού αντικατασταθούν όλες οι μεταβλητές και συγκριθούν μεταξύ τους όλα τα αποκτηθέντα μοντέλα.

3.3.3 Stepwise μέθοδοι (SW)

Οι Stepwise methods βρίσκονται ανάμεσα στις πιο γνωστές μεθόδους επιλογής υποσυνόλων. Αυτές οι μέθοδοι βασίζονται σε δύο διαφορετικές στρατηγικές που ονομάζονται Forward Selection (FS) και Backward Elimination (BE). Η Forward Selection ξεκινάει με ένα μοντέλο μηδενικού μεγέθους και αναπτύσσεται μέσω πρόσθεσης μεταβλητών που ικανοποιούν ένα συγκεκριμένο κριτήριο. Τυπικά, η μεταβλτητή που προστίθεται σε κάθε βήμα είναι εκείνη που περιορίζει όσο το δυνατόν περισσότερο το ελάχιστο άθροισμα τετραγώνων (RSS). Το RSS μπορεί να εκτιμηθεί με ένα F-testπου προσδιορίζεται από τη σχέση:

$$F_j^+ = max_j \left[\frac{RSS_{P-RSS_{P+J}}}{S^2_{P+J}}\right] > F_{in} \tag{3.4}$$

Όπου RSS_p και RSS_{p+j}, είναι τα αθροίσματα τετραγώνων των μοντέλων με pκαι p+jμεταβλητές, S^2_{P+j}είναι η διακύμανση του μοντέλου που δομείται με τις μεταβλητές p+j και το F_{in} χρησιμοποιείται ως κριτήριο παύσης της διαδικασίας.

Η Backward Elimination αναπτύσσεται με την αντίθετη οδό. Ξεκινάει με ένα μοντέλο μεγέθους p, όπου p είναι ο συνολικός αριθμός μεταβλητών και εξαλείφει τις ασυσχέτιστες μεταβλητές με μια διαδικασία βήμα-βήμα. Σε κάθε περίπτωση η μεταβλητή που διαγράφεται είναι συνήθως εκείνη που προσφέρει την ελάχιστη αύξηση στο RSS. Όπως και στην περίπτωση της Forward Selection, αυτό μπορεί να εκτιμηθεί με ένα F- test, που προσδιορίζεται από τη σχέση:

$$F_j^- = min_j \left[\frac{RSS_{p-j} - RSS_P}{S^2_P}\right] > F_{out} \tag{3.5}$$

Όπου F_{out} χρησιμοποιείται ως κριτήριο παύσης.

Ο αρχικός λογάριθμος βελτιώθηκε αργότερα από τον Efroymson το 1960 συνδιάζοντας τη Forward Selection και την Backward Elimination. Ξεκινάει με Forward Selection και ύστερα κάθε μεταβλητή (διαφορετική από την πρώτη), προστίθεται στο μοντέλο. Με αυτό τον τρόπο, δημιουργείται ένα μοντέλο προκειμένου να διευκρινιστεί αν καμία από τις επιλεγμένες μεταβλητές μπορεί να αποκλειστεί χωρίς περεταίρω αύξηση του RSS. Αυτή η στρατηγική που βασίζεται σε F-tests είναι εκτός μόδας και πιο πρόσφατες εκδοχές επιλέγουν μεταβλητές που περιορίζουν άλλες συναρτήσεις.

3.3.4 Γενετικοί Αλγόριθμοι

Ο γενετικός αλγόριθμος είναι μια μέθοδος που παρουσιάστηκε για πρώτη φορά το 1961 από τον Bledsoe και δομήθηκε μαθηματικά από τον Holland το 1975. Ο τελευταίος εμπνεύστηκε από τη θεωρία εξέλιξης του Δαρβίνου. Κάθε γονίδιο ανταγωνίζεται τα άλλα, σύμφωνα με τη ιδέα της επιβίωσης του ισχυρότερου. Σύμφωνα με την ορολογία αυτών των αλγορίθμων, κάθε γονίδιο αντιστοιχεί σε μια μεταβλητή και μια αλληλουχία γονιδίων, δηλαδή ένα χρωμόσωμα, σε ένα μοντέλο. Ο πληθυσμός των χρωμοσωμάτων αρχικοποιείται τυχαία και η παρουσία ή η απουσία μιας μεταβλητής κωδικοποιείται από ένα δυαδικό ψηφίο. Τα χρωμοσώματα αξιολογούνται για την ποιότητά τους, σύμφωνα με μια προκαθορισμένη συνάρτηση και ταξινομούνται ανάλογα. Τα ζεύγη χρωμοσωμάτων μπορούν να παράγουν γόνους με διασταύρωση. Η επιλογή των γονικών χρωμοσωμάτων γίνεται τυχαία ή με προτίμηση στα βέλτιστα. Η δεξαμενή των γονιδίων διατηρείται από τους γόνους, ενώ άλλα γονίδια μπορούν να αλλάξουν σύμφωνα με την πιθανότητα διασταύρωσης. Η δεύτερη φάση της εξέλιξης, με την οποία καινούρια χρωμοσώματα μπορούννα δημιουργηθούν, είναι μια διαδικασία μετάλλαξης στην οποία κάθε γονίδιο μπορεί να αλλάξει σύμφωνα με την πιθανότητα μετάλλαξης. Αυτή η πιθανότητα συχνά ορίζεται σε χαμηλή τιμή προκειμένου να αποφευχθούν σημαντικές μετατοπίσεις, που πιθανά θα οδηγήσουν μακρυά από τη βέλτιστη περιοχή. Κάθε φορά, ένα νέο χρωμόσωμα με την καλύτερη απόκριση από τα ήδη υπάρχοντα, μπαίνει στον πληθυσμό και το χειρότερο μοντέλο απορρίπτεται. Με αυτόν τον τρόπο τα χρωμοσώματα ανταγωνίζονται μεταξύ τους και μόνο το ισχυρότερο επιβιώνει. Η φάση της εξέλιξης επαναλαμβάνεται μέχρι να ικανοποιηθεί το κριτήριο παύσης.

3.3.5 Particle Swarm Optimization(PSO)

Είναι μια μέθοδος που εμπνεύστηκε από τη συμπεριφορά σμήνους πουλιών. Σε αντίθεση με τις προηγούμενες μεθόδους, σε αυτή την μέθοδο υπάρχει συνεργασία προκειμένου να επιτευχθεί επίλυση. Η μέθοδος αυτή αρχικά θεωρήθηκε ως μέθοδος βελτιστοποιήσης και αργότερα τροποποιήθηκε για να εφαρμοστεί συγκεκριμένα στην επιλογή μεταβλητών.

Ο αριθμός των συμμετοχόντων είναι τυχαίος και ο πληθυσμός δημιουργείται τυχαία. Τα ‘συστατικά’ αυτής της μεθόδου κινούνται σε δυαδικό χώρο έρευνας και κάθε θέση αντιστοιχεί σε ένα μοντέλο. Στην τροποποιημένη εκδοχή αυτής της μεθόδου, σε κάθε επανάληψη, εξάγεται ένα διάνυσμα ταχύτητας τυχαία, μέσα στο διάστημα [0,1]. Αυτό το διάστημα διαιρείται σε τρία κελιά και η νέα θέση υπολογίζεται βάσει του κελιού στο οποίο ανήκει το διάνυσμα της ταχύτητας, σύμφωνα με τους ακόλουθους τύπους:

$$\text{if } (0 < v_{id} \leq a) \rightarrow x_{id} \text{ (new)} = x_{id} \text{ (old)} \tag{3.6}$$

$$\text{if } (a < v_{id} \leq 0.5\,(1+a)) \rightarrow x_{id} \text{ (new)} = p_{id} \tag{3.7}$$

$$\text{if } (0.5\,(1+a) < v_{id} \leq 1) \rightarrow x_{id} \text{ (new)} = p_{gd} \tag{3.8}$$

όπου v_{id} είναι η ταχύτητα του i–οστού μέρους κατά μήκος της d–οστής διάστασης, p_{id} είναι η τιμή της d–οστής μεταβλητής για το i- οστό μέρος στην προηγούμενη βέλτιστη θέση του και p_{gd} είναι η τιμή της d–οστής μεταβλητής για το g–οστό μέρος που αντιστοιχεί στο καλύτερο μοντέλο που έχει βρεθεί μέχρι στιγμής. Με αυτό τον τρόπο, η κίνηση κάθε μεταβλητής επηρεάζεται από την προηγούμενη βέλτιστη θέση του, p_{id}, και από την καλύτερη καθολική θέση, p_{gd}. Η παράμετρος a, ονομάζεται στατική πιθανότητα και η αρχική της τιμή είναι συνήθως 0.5. Η στατική πιθανότητα παίζει τον εξής ρόλο: όσο μεγαλύτερη είναι η τιμή του a, τόσο μεγαλύτερη είναι η πιθανότητα να προσπεραστούν τα τοπικά βέλτιστα. Από την άλλη μεριά, μια μικρή τιμή του aευνοεί τη μεταβλητή να ακολουθήσει τις δύο πρώτερες καλύτερες καθολικές θέσεις και τον αλγόριθμο να συγκλίνει πιο γρήγορα. Η τροποποιημένη μορφή αυτής της μεθόδου έχει ως σκοπό να έχει μεγαλύτερη ικανότητα αναζήτησης και έτσι μια ενισχυμένη ικανότητα εντοπισμού της τοπικής περιοχής γύρω από τη μεταβλητή. Αντίστοιχα, η παράμετρος έχει οριστεί έτσι ώστε να μειώνει την παραγωγή. Συνεπώς, η στατική πιθανότητα a, συνήθως ξεκινάει με μια τιμή ίση με 0.5 και μειώνεται μέχρι μια τελική τιμή 0.33 κατά τη σύγκλιση.

3.3.6 Ant Colony Optimization (ACO)

Η μέθοδος Ant Colony Optimization (ACO) είναιεμπνευσμένη από τη φύση καθώς πηγή έμπνευσης είναι οι αποικίες μυρμηγκιών. Τα μυρμήγκια καταφέρνουν να βρουν το συντομότερο μονοπάτι που συνδέει τη φωλιά τους με την πηγή φαγητού με εναπόθεση φερομόνης. Καθώς τα μυρμήγκια ταξιδεύουν σε ένα δρόμο από ένα σημείο εκκίνησης μέχρι την πηγή φαγητού εναποθέτουν φερομόνη. Μεταγενέστερα μυρμήγκια θα διαλέξουν γενικά τα μονοπάτια με την περισσότερη φερομόνη και μετά από πολλές δοκιμασίες θα συγκλίνουν σε ένα βέλτιστο μονοπάτι. Η μέθοδος αυτή βρίσκεται μεταξύ GAκαι PSO. Τα μυρμήγκια δημιουργούνται τυχαία με ένα

προκαθορισμένο αριθμό μεταβλητών, μετά αρχίζουν να δομούν ένα μονοπάτι στο χώρο έρευνας σύμφωνα με ένα πιθανολογικό ή μεταβατικό κανόνα. Η κατάθεση φερομόνης θα είναι ανάλογη της ποιότητας της λύσης που μπορούν να βρουν.

Με τ_{ij} (t) συμβολίζεται η πιθανότητα (η δύναμη του ίχνους της φερομόνης) επιλογής στόχου jμετά τον στόχο i, ή η πιθανότητα της επιλογής μιας συγκεκριμένης τιμής jγια παράμετρο iσε γενιά t. Αρχικά, όλες οι πιθανότητες ορίζονται στην ίδια χαμηλή, μη μηδενική τιμή. Αν η μετάβαση μεταξύ συγκεκριμένων στόχων χρησιμοποιείται,το επίπεδο της φερομόνης αυξάνεται με ένα ποσό ανάλογο της συνάρτησης

$f(k)$ σύμφωνα με τη σχέση:

$$\Delta\tau_{i,j}\ (t) = \Delta\tau_{i,j}\ (t) + f(k)^{\beta} \tag{3.9}$$

όπου β μια σταθερά στο εύρος (0,1]. Όταν όλα τα μυρμήγια αξιολογούνται,το επίπεδο φερομόνης αλλάζει βάσει της σχέσης:

$$\tau_{i,j}\ (t) = (1-\rho)\ \tau_{i,j}\ (t-1) + \rho\ \Delta\tau_{i,j}\ (t) \tag{3.10}$$

όπου ρ αντιπροσωπεύει τον ρυθμό εξάτμισης. Δεδομένου J το σύνολο όλων των στόχων, ο επόμενος στόχος επιλέγεται από τις υπόλοιπες εκφράσεις:

j= max $\{\tau_{i,j}\ (t)\}$ if r $\leq r_0$ = j, διαφορετικά, όπου η πιθανότητα επιλογής συγκεκριμένου στόχου j από το σύνολο των διαθέσιμων στόχων J δίνεται από τη σχέση:

$$P_{i,j}\ (t) = \frac{[\tau_{i,j}(t)]^{a}}{\sum_{l\epsilon J_i^k}[\tau_{i,l}(t)]^{a}} \tag{3.11}$$

όπου J_i^k είναι το σύνολο των μυρμηγκιών, rείναι ένας τυχαίος αριθμός μέσα στο σύνολο [0,1] και r_0 είναι threshold τιμή. Αν το r είναι μικρότερο ή ίσο του r_0 , τότε ακολουθείται μια ντετερμινιστική επιλογή, ενώ αν τοείναι μεγαλύτεροαπό το r_0 , γίνεται μια πιθανολογική επιλογή. Το α ελέγχει κατά πόσο το επίπεδο της φερομόνης επηρεάζει την πιθανότητα με την οποία επιλέγεται ένας συγκεκριμένος στόχος. Άλλο ένα χαρακτηριστικό του ACO αλγορίθμου είναι ο αριθμός εξάτμισης ρ της φερομόνης, διαδικασία που οδηγεί στη μείωση της έντασης του ίχνους φερομόνης κατά το πέρασμα του χρόνου. Σε κάθε επανάληψη, δημιουργούνται τα υποσύνολα, η φερομόνη ενημερώνεται, ένα νέο σύνολο μυρμηγκιών δημιουργείται και η διαδικασία

επαναλαμβάνεται μέχρι την ενεργοποίηση ενός κριτηρίου παύσης. Το κριτήριο για την παύση της έρευνας μπορεί να βασιστεί σε προκαθορισμένη συνάρτηση αξιολόγησης.

3.3.7 Least Absolute Shrinkage and Selection Operator (LASSO)

Η μέθοδος LASSO είναι μέθοδος παλινδρόμησης που παρουσιάστηκε από τον R. Tibshinary. Παρόμοια με την μέθοδο Ordinary Least Squares, η LASSO ελαχιστοποιεί το άθροισμα τετραγώνων αλλά θέτει περιορισμό στο άθροισμα των απολύτων τιμών των συντελεστών έτσι ώστε να είναι μικρότερο από μια προκαθορισμένη σταθερά. Αυτός ο επιπλέον περιορισμός στο άθροισμα των απολύτων τιμών μοιάζει πολύ με αυτόν που παρουσιάζεται στην παλινδρόμηση Ridge, όπου ο περιορισμός αναφέρεται στο άθροισμα των τετραγώνων των τιμών των συντελεστών. Αυτή η απλή τροποποίηση επιτρέπει στη μέθοδο LASSO να προσφέρει επιλογή μεταβλητών διότι η ελαχιστοποίηση των συντελεστών είναι τέτοια ώστε κάποιοι συντελεστές μηδενίζονται.Θα μπορούσαμε να πούμε ότι η LASSO είναι η βελτιωμένη έκδοση της Ridge, κι αυτό γιατί η LASSO έχει τα οφέλη της Ridge , αλλά επιπλέον επιτρέπει την επιλογή μεταβλητών, οδηγώντας έτσι σε βελτιωμένη αποκωδικοποίηση των ανεπτυγμένων μοντέλων. Η παράμετρος λ μπορεί να συντονιστεί ώστε να ορίσει το επίπεδο ελαχιστοποίησης.Όσο μεγαλύτερη είναι η τιμή της λ, τόσο περισσότεροι συντελεστές μηδενίζονται. Το επίπεδο ελαχιστοποίησης μπορεί να οριστεί από παράγοντα ελαχιστοποίησης sπου προσδιορίζεται από τη σχέση:

$$s = \frac{\Sigma|\beta|}{\Sigma|\widehat{\beta_{LS}}|} \tag{3.12}$$

όπου $\widehat{\beta_{LS}}$ είναι οι συντελεστές της OLS. Όταν s=1, το επίπεδο ελαχιστοποίησης είναι μηδενικό και τότε η λύση της LASSO αντιστοιχεί στην OLS λύση. Όταν s<1, η LASSO σμικρύνει τους συντελεστές. Για συγκεκριμένες τιμές του s, μερικοί συντελεστές σμικρύνονται ακριβώς μέχρι το μηδέν.

3.3.8 Ελαστικό δίκτυ (Elastic Net)

Η μέθοδος Elastic Net είναι μέθοδος παλινδρόμησης που παρουσιάστηκε από τους Zou και Hastie μόλις το 2005 και συνδιάζει τις μεθόδους LASSOκαι Ridge με τον εξής τρόπο:

$$\lambda\sum_{j=1}^{p}\left((1-a)\left|\beta_j\right|+\ a\beta_j^2\right) \tag{3.13}$$

Ο όρος β_j^2 της μεθόδου Ridge επιτρέπει την ελαχιστοποίηση των συντελεστών, ενώ ο όρος της LASSO μπορεί να σμικρύνει μερικούς συντελεστές μέχρι το μηδέν κι έτσι να πετύχει επιλογή μεταβλητών. Οι δύο όροι μπορούν να οριστούν κατάλληλα από μια παράμετρο α, ανάλογα με το εξεταζόμενο πρόβλημα κάθε φορά. Η μέθοδος αυτή φαίνεται να είναι ιδιαίτερα χρήσιμη όταν το πρόβλημα αφορά σε μεταβλητές με ισχυρή συσχέτιση μεταξύ τους. Σε αυτή την περίπτωση ο όρος του Ridge ελαχιστοποιεί τους συντελεστές των συσχετισμένων μεταβλητών, ενώ ο όρος του LASSO διαλέγει μία ανάμεσα στις συσχετισμένες μεταβλητές και εναποθέτει όλο το βάρος πάνω της.

3.3.9 Σημασία Μεταβλητών σε PLSπροβολές (VIP)

Η μέθοδος Variables Importance on Partial Least Squares (PLS) Projections (VIP) είναι μέθοδος επιλογής μεταβλητών που βασίζεται στην παλινδρόμηση Canonical Powered PLS (CPPLS). Ο CPPLS αλγόριθμος υποθέτει ότι ο χώρος που ορίζεται από τις στήλες του X, έχει έναν υπόχωρο διάστασης M που περιέχει όλες τις σχετικές πληροφορίες για την πρόγνωση του Υπου είναι γνωστός ως σχετικός υπόχωρος. Οι διαφορετικές προσεγγίσεις για επιλογή μεταβλητών βασισμένες σε PLS κινούνται συνήθως σε περιστροφή της πρότυπης λύσης με κατάλληλη χρήση του διανύσματος βαρύτητας του PLS , $\vec{\omega}$, ή του διανύσματος του συντελεστή παλινδρόμησης , $\vec{b}$. Η VIP μέθοδος επιλέγει μεταβλητές με υπολογισμό του VIP αποτελέσματος για κάθε μεταβλητή και αποκλείει όλες τις μεταβλητές VIP με αποτέλεσμα κάτω από μια προκαθορισμένη οριακή τιμή u. Συνήθως δίνεται η τιμή της μονάδας στη u. Όλες οι παράμετροι που παρέχουν μια αύξηση στην ικανότητα πρόγνωσης του μοντέλου, διατηρούνται. Το VIP αποτέλεσμα για την μεταβλητή j ορίζεται ως :

$$VIP_j = \sqrt{\frac{p\sum_{m=1}^{M} w_{mj}^2 ss\,(b_m t_m)}{\sum_{m=1}^{M} ss\,(b_m t_m)}} \tag{3.14}$$

όπου p είναι ο αριθμός των μεταβλητών,M είναι ο αριθμός των διατηρημένων μεταβλητών, w_{mj} είναι το βάρος της PLS της j-οστής μεταβλητής για την m-οστή αδήλωτη μεταβλητή και ss είναι το ποσοστό του y που επεξηγείται από την m–οστή μεταβλητή. Η VIP τιμή ονομάζεται σταθμισμένο άθροισμα των PLS βαρών το οποίο λαμβάνει υπόψη τη διακύμανση κάθε PLS διάστασης. Ο κανόνας «μεγαλύτερος του

ενός» γενικά χρησιμοποιείται ως κριτήριο για την επιλογή μεταβλητών, γιατί η μέση τιμή των τετραγώνων των VIP αποτελεσμάτων ισούται με τη μονάδα.

3.4 Μαθηματική προσέγγιση του προβλήματος επιλογής μεταβλητών

Ας υποθέσουμε πως ονομάζουμε τη μεταβλητή που μας ενδιαφέρει Y και X_1,....X_pθα είναι το σύνολο των ενδεχόμενων επεξηγηματικών μεταβλητών. Το σύνολο αυτό θα είναι ένα διάνυσμα n παρατηρήσεων. Το πρόβλημα της επιλογής μεταβλητών ή επιλογής υποσυνόλου μεταβλητών (subset selection), όπως συχνά αναφέρεται, προκύπτει όταν ο μελετητής επιχειρεί να μοντελοποιήσει τη σχέση μεταξύ Y και συνόλου X_1,....X_p. Το βασικό εμπόδιο σε αυτή την προσπάθεια είναι η αβεβαιότητα σχετικά με το ποιό υποσύνολο μεταβλητών θα πρέπει να χρησιμοποιηθεί. Μια τέτοια περίπτωση έχει ιδιαίτερο ενδιαφέρον όταν το ρείναι μεγάλο και το σύνολο X_1,....X_p θεωρείται ότι περιέχει πολλές περιττές ή ασυσχέτιστες μεταβλητές.

Το πρόβλημα επιλογής μεταβλητών είναι το πιο σύνηθες στο πλαίσιο της γραμμικής παλινδρόμησης όπου το ενδιαφέρον περιορίζεται σε ομαλά γραμμικά μοντέλα. Έστω γ ο δείκτης του υποσυνόλου X_1 ,.... X_p και q_γ το μέγεθος του γ-οστο υποσυνόλου. Το θέμα τότε είναι να γίνει η σωστή επιλογή και να προσαρμοστεί ένα μοντέλο της μορφής Y= $X_\gamma\beta_\gamma$ + ε, όπου X_γ είναι ένας n x q_γ πίνακας του οποίου οι στήλες αντιστοιχούν στο γ-οστό υποσύνολο, β_γ είναι ένα q_γ x1 διάνυσμα των συντελεστών παλινδρόμησης και ε ~ N (0, σ^2I).

Γενικότερα, το πρόβλημα της επιλογής μεταβλητώνείναι μια ειδική περίπτωση του προβλήματος επιλογής μοντέλου, όπου κάθε μοντέλο υπό εξέταση αντιστοιχεί σε διαφορετικό υποσύνολο X_1 ,....X_p. Τυπικά, κάθε μοντέλο εφαρμόζεται σε όλα τα πιθανά υποσύνολα. Αυτό σημαίνει πως, για παράδειγμα, μπορούμε να χρησιμοποιήσουμε μια ευρεία ποικιλία συσχετίσεων με γενικευμένα γραμμικά μοντέλα. Μοντέλα της μορφής g (E (Y)) = a +$X_\gamma\beta_\gamma$, μπορούν να χρησιμοποιηθούν για διάφορες συνδετικές συναρτήσεις g.

Προχωρώντας πέρα από τα συνήθη γραμμικά μοντέλα, ο μελετητής θα μπορούσε να συσχετίσει το Y με τα υποσύνολα X_1,....X_p με μη παραμετρικά μοντέλα που έχουν αναπτυχθεί, όπως το CART ή το MARS.

Οι δομικές εξελίξεις στην επιλογή μεταβλητών φαίνεται να έχουν συμβεί είτε απευθείας μέσα στο πλαίσιο των γραμμικών μοντέλων ή στο πλαίσιο της γενικής επιλογής μοντέλων. Ιστορικά, η εστίαση σε αυτό το θέμα ξεκίνησε με το γραμμικό μοντέλο το 1960 όταν δημιουργήθηκε το πρώτο κύμα σημαντικών εξελίξεων στην υπολογιστική ευχέρεια της πληροφορικής. Το ενδιαφέρον για τα γραμμικά μοντέλα εξακολουθεί να υπάρχει διότι διευκολύνει πολλές φορές την πρόγνωση των κατάλληλων μεταβλητών, αλλά επίσης επειδή πολλά προβλήματα ενδιαφέροντος μπορούν να θεωρηθούν γραμμικά. Για παράδειγμα, για το πρόβλημα εκτίμησης μη

παραμετρικής συνάρτησης , το Y αντιπροσωπεύει τις τιμές της άγνωστης συνάρτησης και το σύνολο X_1 ,.... X_p αντιπροσωπεύει μια γραμμική βάση. Όμως, καθώς η ανάπτυξη στην επιστήμη των υπολογιστών έχει επιτρέψει την χρήση πλουσιότερων κλάσεων μοντέλων, έχει επικρατήσει η διαχείριση του προβλήματος με προσεγγιστική επιλογή γενικών μοντέλων.

3.5 Κριτήρια επιλογής υποσυνόλων μεταβλητών

Βασικό χαρακτηριστικό του προβλήματος επιλογής μεταβλητών είναι το τεράστιο μέγεθός τους. Ακόμη και με μέτριες τιμές για τον αριθμό των μεταβλητών p, τα υπολογιστικά στοιχεία για όλα τα 2^P προκύπτοντα μοντέλα είναι ιδιαίτερα ακριβά και γι' αυτόν το λόγο επιβάλλεται μείωση των διαστάσεων του χώρου μοντέλων. Θεωρώντας σαν βάση το γραμμικό μοντέλο, κάποιες προτάσεις για μείωση των διαστάσεων βασίζονται στο ελάχιστο άθροισμα τετραγώνων που παρέχει μια μερική διάταξη μοντέλων. Οι περισσότεροι από τους προτεινόμενους αλγορίθμους εστιάζουν την προσοχή τους στα καλύτερα υποσύνολα κάθε μεγέθους. Διαφορετικά, η μείωση των διαστάσεων γίνεται με παραλλαγές των stepwise μεθόδων, οι οποίες διαδοχικά προσθέτουν ή διαγράφουν μεταβλητές. Ακόμη και στις μέρες μας, μετά από την ραγδαία ανάπτυξη των υπολογιστικών μεθόδων, αυτοί οι αλγόριθμοι- διαδικασίες παραμένουν σταθερές αξίες στη μείωση των μεταβλητών ενός προβλήματος. Όταν, πλέον, η προσοχή του αναλυτή εστιάζει περισσότερο σε ένα πιο εύχρηστο σύνολο μοντέλων, χρειάζονται κατάλληλα κριτήρια για να επιλεχθεί το βέλτιστο υποσύνολο μεταβλητών. Η πρώτη ανάπτυξη τέτοιων κριτηρίων επιλογής, πάντα στο πλαίσιο του γραμμικού μοντέλου, βασίστηκαν στην προσπάθεια μείωσης του μέσου τετραγωνικού σφάλματος πρόγνωσης. Κάποια άλλα κριτήρια που αντιστοιχούν σε διαφορετικές υποθέσεις σχετικά με το ποιές προγνωστικές τιμές θα χρησιμοποιηθούν, έχουν προταθεί από τον Hocking (1976) και τον Thomson (1978). Ένα από τα πιο γνωστά κριτήρια αυτού του είδους είναι του Mallows σύμφωνα με το οποίο:

$$C_p = (\frac{RSS_\gamma}{\hat{\sigma}^2_{FULL}} + 2q_\gamma - n) \tag{3.15}$$

όπου RSS_γ είναι το ελάχιστο άθροισμα τετραγώνων για το ∞γ-οστό μοντέλο και $\hat{\sigma}^2_{FULL}$ είναι η συνήθης αμερόληπτη εκτίμηση του σ^2 που αναφέρεται σε ολόκληρο το μοντέλο. Ο Mallow, ωθούμενος από την αμερόληπτηεκτίμηση της προγνωστικής ακρίβειας του γ-οστού μοντέλου, πρότεινε τη χρήση τωνC_p γραφημάτων προκειμένου να βοηθήσει στην επιλογή του κατάλληλου υποσυνόλου μεταβλητών.

Δύο άλλα ιδιαίτερα δημοφιλή κριτήρια είναι το Akaike Information Criterion (AIC) και το Bayesian Information Criterion (BIC), τα οποία θα αναλυθούν εκτενέστερα στην επόμενη υποενότητα. Ο Akaike πρότεινε το κριτήριο AIC ωθούμενος από μια θεωρητική πληροφοριακή άποψη ξεκινώντας από την ελαχιστοποίηση της απόστασης των Kullback- Leibler, μεταξύ των κατανομών του y κάτω από το γ-οστό μοντέλο και

το πραγματικό. Επιπλέον, διακρίνεται μια ασυπτωτική ισορροπία του AIC στον Stone (1977). Αντίθετα, ο Schwarz (1978) πρότεινε το BICαπό μια Bayesian οπτική, δείχνοντας ότι υπήρχε ασυμπτωτική ισορροπία καθώς το ntείνει στο άπειρο. Το κριτήριο BIC επικυρώθηκε επίσης, από τη θεωρία κωδικοποίησης του Rissanen (1978). Κατα καιρούς γίνονται πολλές συγκρίσεις των κριτηρίων AIC και BIC βασιζόμενες στην ασυμπτωτική σταθερότητα όταν n→∞. Καθώς αποδεικνύεται, το κριτήριο BIC είναι συνεπές ως προς τη σταθερότητα όταν είναι καθορισμένο το πραγματικό μοντέλο. Αντίθετα, το AIC είναι συνεπές όταν οι διαστάσεις του πραγματικού μοντέλου αυξάνουν μαζί με το n.

Για το γραμμικό μοντέλο, πολλά από τα δημοφιλή κριτήρια επιλογής, είναι ειδικές περιπτώσεις ενός κριτηρίου αθροίσματος τετραγώνων, που παρέχει ένα ενιαίο πλαίσιο συγκρίσεων. Θεωρώντας ότι το σ^2 είναι γνωστό, προκειμένου να αποφευχθούν επιπλοκές, αυτό το γενικό κριτήριο επιλέγει το υποσύνολο που ελαχιστοποιεί το

$$(\frac{RSS_\gamma}{\hat{\sigma}^2} + Fq_\gamma) \tag{3.16}$$

όπου F είναι μια προκαθορισμένη «ποινή διάστασης». Το AIC και το ελάχιστο C_p είναι ουσιαστικά ισοδύναμα, με την αντιστοιχία $F = 2$ και το BIC λειτουργεί θέτοντας $F = \log n$. Επιβάλλοντας μικρότερη ποινή, το AIC και το ελάχιστο C_p θα επιλέξουν μεγαλύτερα μοντέλα σε σχέση με το BIC, εκτός και αν το neίναι πολύ μικρό.

Μελετώντας περισσότερο την επιλογή της F, αυτή επιβάλλεται όταν όλοι οι προγνώστες είναι ορθογώνιοι, στην περίπτωση δηλαδή που οι προγνώστες επιλέγονται με t–στατιστική, με t τέτοιο ώστε t^2 >F. Όταν $X_1, \ldots X_p$ είναι στην πραγματικότητα όλοι ασυσχέτιστοι με το Y, δηλαδή οι συνολικοί συντελεστές παλινδρόμησης του μοντέλου είναι μηδενικοί, τα AIC, C_p είναι ελεύθερα και τείνουν να περιλαμβάνουν ένα μεγάλο ποσοστο των ασυσχέτιστων μεταβλητών. Μια συντηρητική επιλογή για το F προτείνεται από το γεγονός ότι κάτω από το μηδενικό μοντέλο, η αναμενόμενη τιμή του μεγαλύτερου t-στατιστικού τετραγώνου είναι περίπου 2logp όταν το p είναι μεγάλο. Αυτό προτείνει την επιλογή F = 2logp το οποίο προτάθηκε από τους Fosterκαι George (1994). Παρακινημένοι από παρόμοιες σκέψεις, ο Tibshinani (1999) πρόσφατα πρότεινε το κριτήριο «Covariance inflation criterion (CIC)», μια μη παραμετρική μέθοδο επιλογής. Επίσης, άλλη μια πολλά υποσχόμενη προσαρμογή βασισμένη στην έννοια των γενικευμένων βαθμών ελευθερίας είναι αυτή του Ye (1998).

Πολλά άλλα ενδιαφέροντα κριτήρια που αντιστοιχούν σε διαφορετικές επιλογές του προαναφερθέντος F, έχουν προταθεί από τους Hurvitz, Tsai (1989, 1998), Rao, Wu (1989), Wei (1992), Shao (1997).

Ένα από τα μειοντεκτήματα χρήσης σταθερής, καθορισμένης επιλογής F, είναι ότι ευνοούνται κάποια μοντέλα συγκεκριμένου μεγέθους. Συγκεκριμένα αυτό σημαίνει πως για μικρή τιμή του F ευνοούνται μοντέλα μεγάλου μεγέθους, ενώ για μεγάλη τιμή του Fευνοούνται μοντέλα μικρού μεγέθους. Προσαρμοστικές επιλογές του F για να μετριαστεί αυτό το πρόβλημα, έχουν προταθεί από τους Benjamini, Hochberg (1995) και τους George, Foster (2000).

Μια εναλλακτική στην αποσαφήνιση διαφόρων κριτηρίων, είναι μια επιλογή βασισμένη στο προγνωστικό σφάλμα που εκτιμάται μετά από εντατικές υπολογιστικές μεθόδους. Μια ενδιαφέρουσα παραλλαγή είναι η μέθοδος «little boostrap» (Brieman, 1992), το οποίο εκτιμά το προγνωστικό σφάλμα των επιλεγμένων μοντέλων βασιζόμενη στην αναπαραγωγή σύγκρισης δεδομένων.

Άλλο μειονέκτημα των παραδοσιακών μεθόδων επιλογής υποσυνόλων, το οποίο κερδίζει όλο και περισσότερη προσοχή από τους μελετητές, είναι η αστάθεια η σχετική με μικρές παραλλαγές στα δεδομένα. Δύο νέες εναλλακτικές που μετριάζουν κατά μέρος την αστάθεια των γραμμικών μοντέλων είναι η διαδικασία «Non negative garotte» (Brieman, 1995) και η «Lasso», (Tibshinani, 1996). Και οι δύο παραπάνω διαδικασίες αντικαθιστούν το κριτήριο ελαχίστων τετραγώνων περιορίζοντας τα κριτήρια βελτιστοποίησης. Καθώς οι περιορισμοί γίνονται περισσότερο αυστηροί, οι υπολογισμοί εκμηδενίζονται και έτσι το κατάλληλο υποσύνολο αναγνωρίζεται και εκτιμάται κατάλληλα.

3.5.1 Κριτήρια Akaike, AIC, BIC

Ας ξεκινήσουμε θεωρώντας πως τα διαθέσιμα δεδομένα βρίσκονται στη μορφή $(x_i^T, y_i)_{i=1}^n$, όπου $, y_i$ είναι η i- οστή παρατήρηση της αποκρίνουσας μεταβλητής και x_iείναι το διάνυσμα p- διάστασης των σχετιζόμενων μεταβλητών. Οι σχετιζόμενες μεταβλητές, θεωρούνται συνήθως σαν ένα τυχαίο δείγμα του πληθυσμού (X^T, Y), όπου ο υποθετικός μέσος του Y, δεδομένου του X, εξαρτάται από τον γραμμικό προγνώστη $\beta^T X$, όπου $\beta = (\beta_{1,\dots,}\beta_p)^T$. Σε ορισμένα είδη μοντελοποίησης, υποτίθεται συνήθως ότι οι περισσότεροι συντελεστές παλινδρόμησης β_j είναι μηδενικοί. Η επιλογή μεταβλητών έχει σαν στόχο να αναγνωρίσει όλες τις σημαντικές μεταβλητές των οποίων οι συντελεστές παλινδρόμησης δεν εξαφανίζονται και παρέχουν αποτελεσματικές εκτιμήσεις αυτών των συντελεστών.

Προκειμένου να εστιάσουμε στα κριτήρια επιλογής μεταβλητών, θεωρούμε ότι τα δεδομένα παράγονται από την πραγματική συνάρτηση f_{θ_o}, με παράμετρο το διάνυσμα $\theta_o = (\theta_{1,\dots,}\theta_d)^T$. Η συνάρτηση που προαναφέρθηκε, ανήκει στην ευρύτερη οικογένεια μοντέλων της μορφής f_{θ_1}, όπου θ_o είναι ένα υποδιάνυσμα της διανυσματικής (p- διάστασης) παραμέτρου θ_1. Τα προβλήματα που επικεντρώνονται στον σωστό τρόπο εκτίμησης των διαστάσεων ενός μοντέλου αλλά και στην

σύγκριση μοντέλων διαφορετικών διαστάσεων, προκύπτουν σε πολλές στατιστικές εφαρμογές, συμπεριλαμβανομένης της μοντελοποίησης χρονοσειρών.

Ο Akaike προτείνει την επιλογή ενός μοντέλου που ελαχιστοποιεί την απόκλιση των Kullback –Leibler (KL) του χρησιμοποιούμενου μοντέλου από το πραγματικό. Ο Akaike χρησιμοποιεί τον εκτιμητή μέγιστης πιθανότητας (maximum likelihood estimator – MLE) $\hat{\theta} = (\widehat{\theta_1}, \ldots, \widehat{\theta_p})^T$ της διανυσματικής παραμέτρου $\boldsymbol{\theta}$, και δείχνει ότι πάνω από μια σταθερά, η αναμενόμενη απόκλιση KL μπορεί να αναπτυχθεί ασυμπτωτικά όπως φαίνεται ακολούθως:

$$-l_n(\hat{\theta}) + l(\hat{\theta}) = -\, l_n(\hat{\theta}) + \lambda \sum_{j=1}^{p} I(\widehat{\theta_j} \neq 0) \qquad (3.17)$$

Όπου $l_n(\boldsymbol{\theta})$ είναι η συνάρτηση log- πιθανότητας, dim($\boldsymbol{\theta}$) δηλώνει τη διάσταση του μοντέλου και λ=1. Αυτό οδηγεί στο κριτήριο AIC. Ο Schwartz πήρε την Bayesian προσέγγιση με πρωταρχικές επιρροές αυτές που έχουν μη μηδενικές πιθανότητες σε κάποιους χαμηλότερων διαστάσεων υπόχωρους και πρότεινε το κριτήριο BIC με $\lambda = \frac{\log n}{2}$ για την επιλογή μοντέλου.

Τα κριτήρια AIC και BIC προτείνουν μια ενοποιημένη προσέγγιση της επιλογής μοντέλου. Πιο συγκεκριμένα, υποδεικνύουν την επιλογή μιας διανυσματικής παραμέτρου $\boldsymbol{\theta}$ που μεγιστοποιεί την πιθανότητα $\ln(\boldsymbol{\theta}) - \lambda \|\boldsymbol{\theta}\|_0$, όπου η L_0–νόρμα της διανυσματικής παραμέτρου $\boldsymbol{\theta}$ μετράει τον αριθμό των συνιστωσών του $\boldsymbol{\theta}$που δεν εξαφανίζονται. Όσον αφορά στο λ, πρόκειται για μια παράμετρο για την οποία ισχύει $\lambda \geq 0$. Δεδομένου ότι $\|\boldsymbol{\theta}\|_0 = m$, η λύση στην παραπάνω σχέση, είναι το υποσύνολο με τη μέγιστη πιθανότητα ανάμεσα σε όλα τα υποσύνολα μεγέθους m. Εκείνο το μοντέλο μεγέθους m, είναι αυτό που επιλέγεται ώστε να μεγιστοποιήσει ανάμεσα σε p υποσύνολα, τα καλύτερα υποσύνολα μεγέθους m, με $1 \leq m \leq p$.

3.6 Επιλογή μεταβλητών σε χώρους υψηλών διαστάσεων και πολύ υψηλών διαστάσεων

Η ανάλυση δεδομένων σε χώρους υψηλών διαστάσεων έχει γίνει βαθμιαία ιδιαίτερα σημαντική σε διάφορα επιστημονικά πεδία, όπως η μηχανική, οι ανθρωπιστικές επιστήμες, τα οικονομικά και η πληροφορική. Η ανάλυση σε τέτοιους χώρους δεδομένων χαρακτηρίζει ιδιαίτερα πολλά προβληματα στην επιστήμη της στατιστικής .Η έννοια «high dimensionality», χαρακτηρίζει τις περιπτώσεις προβλημάτων όπου οι διαστάσεις του χώρου μεταβλητών αυξάνουν κατά τη διάρκεια της μελέτης. Είναι, όμως, ιδιαίτερα συχνό το φαινόμενο όπου ο αναλυτής έρχεται αντιμέτωπος με χώρους δεδομένων που δεν μπορούν απλά να χαρακτηριστούν ως «high dimensional feature spaces». Στις περιπτώσεις αυτές, οι διαστάσεις του χώρου προς μελέτη, αυξάνουν με μη- πολυωνυμικό ρυθμό καθώς το απλό μέγεθος μεγαλώνει. Οι περιπτώσεις αυτές χαρακτηρίζονται ως προβλήματα πολύ υψηλής διάστασης (« ultra-high dimensional».

Ο Donoho (2000) αποδεικνύει με την εργασία του την ανάγκη για ανάπτυξη στην ανάλυση δεδομένων σε χώρους υψηλών διαστάσεων. Οι Fan και Li (2006), παρουσιάζουν μια κατανοητή επισκόπηση των στατιστικών προκλήσεων με τις υψηλές διαστάσεις σε ένα ευρύ φάσμα θεμάτων. Συγκεκριμένα, αποδεικνύουν πως για ένα πλήθος στατιστικών προβλημάτων, οι παράμετροι του μοντέλου μπορούν να εκτιμηθούν. Επίσης, αποδεικνύουν πως μπορεί να εκτιμηθεί αν το καλύτερο μοντέλο είναι γνωστό εκ των προτέρων, όσο οι διαστάσεις δεν είναι υπερβολικά υψηλές. Οι προκλήσεις που δεν υπάρχουν σε μελέτες μικρότερων διαστάσεων έχουν επαναπροσδιορίσει τη στατιστική σκέψη, την ανάπτυξη διάφορων μεθοδολογιών καθώς και τις θεωρητικές μελέτες.

Οι τρεις σημαντικοί πυλώνες όλων των στατιστικών διαδικασιών, είναι η στατιστική ακρίβεια, η σωστή επεξήγηση του μοντέλου και η υπολογιστική πολυπλοκότητα.Στις παραδοσιακές μελέτες, ο αριθμός των παρατηρήσεων n είναι πολύ μεγαλύτερος από τον αριθμό των μεταβλητών ή των παραμέτρων p. Σε τέτοιες περιπτώσεις, καμία από τις τρεις σημαντικές απόψεις δεν χρειάζεται να θυσιαστεί για την απόδοση των υπόλοιπων. Οι παραδοσιακές μέθοδοι, όμως αντιμετωπίζουν σημαντικές προκλίσεις όταν η διάσταση του χώρου, p, συγκρίνεται με το μέγεθος του δείγματος n. Αυτές οι προκλίσεις περιλαμβάνουν τον τρόπο σχεδιασμού στατιστικών διαδικασιών που είναι περισσότερο αποδοτικές, τον τρόποάντλησης την ασυμπτωτική θεωρία καθώς και τον τρόπο να δημιουργηθούν στατιστικές διαδικασίες που είναι υπολογιστικά αποδοτικές.

Μια περιβόητη δυσκολία της επιλογής σε υψηλών διαστάσεων μοντέλα, προέρχεται από τη συγγραμμικότητα ανάμεσα στους προγνώστες. Η συγγραμμικότητα μπορεί εύκολα να νοθευτεί στη γεωμετρία υψηλών διαστάσεων, γεγονός που μπορεί να μας οδηγήσει σε επιλογή λανθασμένου μοντέλου.

Το γεγονός που δίνει τη δυνατότητα παραγωγής ενός ικανοποιητικού στατιστικού αποτελέσματος σε χώρους υψηλών διαστάσεων είναι η υπόθεση ότι η συνάρτηση παλινδρόμησης βρίσκεται σε χώρο λιγότερων διαστάσεων. Σε τέτοιες περιπτώσεις, οι παράμετροι παλινδρόμησης p-διαστάσεων υποτίθεται ότι είναι αραιές με πολλές μηδενικές συνιστώσες. Οι μη μηδενικές συνιστώσες υποδεικνύουν τις σημαντικές μεταβλητές. Όταν οι παράμετροι που προαναφέραμε είναι αραιές, η επιλογή μεταβλητών μπορεί να βελτιώσει την ακρίβεια της εκτίμησης μέσω ενός αποδοτικού προσδιορισμού του υποσυνόλου που περιέχει όλους τους σημαντικούς προγνώστες. Το σύνολο αυτό μπορεί να αυξήσει την επεξηγηματικότητα του μοντέλου και να βοηθήσει στη μείωση του υπολογιστικού κόστους όταν οι μεταβλητές είναι ιδιαίτερα αραιές.

Η έννοια των ‘αραιών’ μεταβλητών είναι ιδιαίτερα στενή. Θα μπορούσε να γίνει πιο εύκολα αντιληπτή σε μετασχηματισμένους ή διευρυμένους χώρους στοιχείων. Αυτό σημαίνει πως, για παράδειγμα, κάποια πρότερη γνώση θα μπορούσε να μας οδηγήσει στην εφαρμογή κάποιας ομαδοποίησης ή μετασχηματισμού των εισαγώμενων μεταβλητών. Ίσως κάποιος μετασχηματισμός των μεταβλητών είναι κατάλληλος αν ένα σημαντικό μέρος των συσχετισμένων ζευγών είναι μεγάλο. Σε κάποιες

περιπτώσεις, πιθανά ο αναλυτής χρειάζεται να μεγενθύνει το χώρο των δεδομένων προσθέτοντας αλληλεπιδράσεις και όρους υψηλότερων τάξεων προκειμένου να μειωθούν οι διαστάσεις του μοντέλου.

Το ζήτημα των αραιών μεταβλητών προκύπτει σε πολλές επιστημονικές προσπάθειες. Σε ταξινόμηση ασθενειών, για παράδειγμα, θεωρείται γενικά μόνο κάποιες δεκάδες γονιδίων είναι υπεύθυνες για μια ασθένεια. Η επιλογή μόνο κάποιων δεκάδων γονιδίων μπορεί να βοηθήσει όχι μόνο τους αναλυτές της στατιστικής να δομήσουν ένα πιο αποδοτικό κανόνα ταξινόμησης, αλλά και τους βιολόγους να κατανοήσουν τους μοριακούς μηχανισμούς.

Οι κύριοι στόχοι της παλινδρόμησης αλλά και της ταξινόμησης σε χώρους υψηλών διαστάσεων σύμφωνα με τον Bickel (2008) είναι πρωταρχικά η δόμηση μιας όσο γίνεται αποδοτικής μεθόδου προκειμένου να γίνει πρόγνωση μελλοντικών παρατηρήσεων. Έπειτα, στόχος είναι να δημιουργηθεί μια εξελιγμένη μέθοδος πρόγνωσης.

Τις περισσότερες φορές, είναι ιδιαίτερα χρήσιμο να διακριθούν δύο τύποι στατιστικών προσπαθειών σε χώρους υψηλών διαστάσεων. Ο ένας τύπος βασίζεται στην ακρίβεια των εκτιμώμενων παραμέτρων του μοντέλου και ο άλλος βασίζεται στην ακρίβεια της εκτιμώμενης απώλειας του εκτιμώμενου μοντέλου. Η τελευταία ιδιότητα προκύπτει σε προβλήματα μηχανικής μάθησης όπως η ταξινόμηση εγγράφων. Εμφανίζεται, επίσης, σε πλαίσια όπου θέλουμε να διευκρινίσουμε τους σημαντικούς προγνώστες και να χαρακτηρίσουμε την ακριβή συνεισφορά καθενός στην μεταβλητή απόκρισης. Σαν παράδειγμα, θα μπορούσαμε να αναφέρουμε τις ιατρικές μελέτες στις οποίες θα πρέπει να οριστεί η σημασία των παραγόντων ρίσκου προκειμένου να επιτευχθεί ασφαλής πρόγνωση. Πολλές τελευταίες μελέτες έχουν επικεντρωθεί στη συνέπεια των μεθόδων επιλογής μεταβλητών σε χώρους υψηλών διαστάσεων, παραμερίζοντας κατά ένα τρόπο τον χαρακτηρισμό των ασυμπτωτικής συμπεριφοράς των εκτιμώμενων παραμέτρων.

Η επιλογή μεταβλητών σε χώρους υψηλών διαστάσεων περιλαμβάνει πλήθος περιορισμών οι οποίοι αυξάνονται με ραγδαίους ρυθμούς. Για αυτό το λόγο γίνονται συνεχώς προσπάθειες ανάπτυξης νέων μεθόδων επιλογής μεταβλητών. Τα κύρια θεωρητικά ερωτήματα που προκύπτουν συμπεριλαμβάνουν την αποσαφήνιση των διαστατικών ορίων που μπορούν να διαχειριστούν οι διάφορες τεχνικές και τον χαρακτηρισμό της βέλτιστης μεθόδου επιλογής μεταβλητών.

Η επιλογή μεταβλητών σε χώρους ιδιαίτερα υψηλών διαστάσεων έχει γίνει ιδιαίτερα σημαντικό θέμα στον τομέα της στατιστικής και απαιτεί την δημιουργία νέων, ανεπτυγμένων μεθοδολογιών και θεωριών.

Σε μελέτες τέτοιου είδους, επιβάλλεται η αναγνώριση των σημαντικών στοιχείων που συνεισφέρουν τα περισσότερα στην απόκριση και στην αξιοπιστία της πρόγνωσης. Όπως ακριβώς συμβαίνει και σε χώρους υψηλών διαστάσεων, έτσι και σε αυτό το πλαίσιο, ενδιαφέρουν σημαντικά τον αναλυτή το υπολογιστικό κόστος, η στατιστική

ακρίβεια και η δυνατότητα μετάφρασης των αποτελεσμάτων που μας δίνει το επιλεγμένο μοντέλο.

Οι υπάρχουσες τεχνικές επιλογής μεταβλητών, μπορούν να γίνουν χρήσιμες υπολογιστικά για την αντιμετώπιση επιλογής προβλημάτων σε χώρους πολύ υψηλών διαστάσεων. Μια ιδέα είναι η μείωση των διαστάσεων p από μια τεράστια κλίμακα σε μια σχετικά χαμηλότερη, με τη βοήθεια μαις γρήγορης, έμπιστης και αποδοτικής μεθόδου. Στόχος της ενέργειας αυτής θα είναι η εφαρμογή των καλά ανεπτυγμένων τεχνικών επιλογής μεταβλητών σε χώρους μειωμένων διαστάσεων. Ο χειρισμός αυτός είναι ιδιαίτερα αποδοτικός καθώς αντιστοιχεί το κόστος, την ακρίβεια και τη δυνατότητα μετάφρασης των αποτελεσμάτων σε ασυμπτωτική πιθανότητα, όπως διευκρίνησαν οι Fankαι Lv το 2008.

Σαν απόσταγμα των παραπάνωθα μπορούσαμε να καταλήξουμε σε μια μέθοδο δύο κλιμάκων για προβλήματα επιλογής μεταβλητών σε χώρους πολύ υψηλών διαστάσεων. Η μέθοδος θα αποτελείται από μια κλίμακα πρώτης διαλογής μεταβλητών η οποία θα ακολουθείται από μια μέτρια κλίμακα επιλογής. Η ιδέα αυτή προτάθηκε από τους Fan και Lv και σύμφωνα με αυτήν ο αναλυτής έχει τη δυνατότητα να επιλέξει ανάμεσα στις διάφορες μεθόδους διαλογής εφόσον αυτή παρέχει σίγουρη και ακριβή διαλογή.Με παρόμοιο σκεπτικό, ο αναλυτής μπορεί να επιλέξει το επιθυμητό εργαλείο για μια διαλογή μέτριας κλίμακας. Η επιλογή μεταβλητών μέσω υψηλής και μέτριας κλίμακας διαλογής, μπορεί να εφαρμοστεί επαναληπτικά, οδηγώντας σε επαναληπτικού σίγουρου κρησαρίσματος (Iterative Sure Independence Screening, ISIS). Η ανάπτυξη και οι επεκτάσεις τουIterSISπροτείνονται από τους Fan, Samworth και Wu, οι οποίοι επίσης ανέπτυξαν κώδικες του Matlab για να διευκολύνουν την εκτέλεση σε γενικευμένα γραμμικά μοντέλα.

3.7 Επιλογή μεταβλητών για ταξινόμηση

Η επιλογή μεταβλητών παίζει πολύ σημαντικό ρόλο στην ταξινόμηση. Πριν σχεδιαστεί μια μέθοδος ταξινόμησης όταν εμπλέκονται πολλές μεταβλητές, πρέπει να επιλεχθούν μόνο εκείνες οι μεταβλητές που πραγματικά χρειάζονται. Κατά τη διάρκεια της ανάλυσης είναι δυνατό να παρουσιαστούν διάφοροι λόγοι για να επιλεχθεί μόνο ένα υποσύνολο μεταβλητών από όλα τα υποψήφια σύνολα. Ο βασικός λόγος είναι το γεγονός πως είναι οικονομικότερο να αναλυθούν και να ληφθούν υπόψη μόνο μειωμένα σύνολα μεταβλητών. Επίσης, η ακρίβεια της πρόβλεψης μπορεί να βελτιωθεί σημαντικά μέσω του αποκλεισμού των περιττών ή ασυσχέτιστων μεταβλητών. Σημαντικός, επιπλέον, παράγοντας είναι η μεγαλύτερη ευκολία δόμησης του προγνώστη όταν χρησιμοποιούνται λιγότερες εισαγόμενες μεταβλητές. Γνωρίζοντας ποιες μεταβλητές είναι σχετικές μεταξύ τους βρισκόμαστε στη θέση να προσδιορίσουμε πιο εύκολα τη φύση του προβλήματος πρόγνωσης και να αντιληφθούμε καλύτερα το τελικό μοντέλο ταξινόμησης.

Ο σκοπός της ταξινόμησης είναι να ταξινομηθούν παρατηρήσεις που χαρακτηρίζονται από κάποια γνωρίσματα ως μεταβλητές. Αυτό συμβαίνει προκειμένου να αποφασιστεί σε ποιά κλάση ανήκει κάθε παρατήρηση. Βασιζόμενοι σε ένα σύνολο παρατηρήσεων, των οποίων γνωρίζουμε την κλάση, μπορούμε να σχεδιάσουμε ένα σύνολο κανόνων το οποίο γενικεύεται ώστε να ταξινομηθεί το σύνολο των παρατηρήσεων με τη μεγαλύτερη δυνατή ακρίβεια. Υπάρχουν διάφορες μεθοδολογίες για να αντιμετωπιστεί αυτό το πρόβλημα. Οι περισσότερο συχνά εμφανιζόμενες στις διάφορες αναλύσεις- μελέτες είναι η κλασσική διακριτική ανάλυση, η λογιστική παλινδρόμηση και τα νευρωνικά δίκτυα. Οι δύο πρώτες ψάχνουν για γραμμικές συναρτήσεις και στη συνέχεια τις χρησιμοποιούν για σκοπούς ταξινόμησης. Η χρήση συναρτήσεων επιτρέπει την καλύτερη μετάφραση των αποτελεσμάτων. Φυσικά, δεν είναι κάθε μέθοδος ταξινόμησης κατάλληλη για όλους τους τύπους ανάλυσης. Παρόλα αυτά, η κλασσική διακριτική ανάλυση συνεχίζει να είναι αυτή που δίνει τα πιο ενδιαφέροντα αποτελέσματα.

3.8 Μέθοδοι ταξινόμησης μεταβλητών

Θα θεωρήσουμε προβλήματα ταξινόμησης με μεγάλο αριθμό μεταβλητών που μπορεί να είναι ποσοτικοί ή ποιοτικοί. Μας ενδιαφέρει η περίπτωση στην οποία οι ανεξάρτητες μεταβητές μπορεί να είναι ασυσχέτιστες με την ταξινόμηση, αλλά παρόλα αυτά, χρειάζεται να επιλεχθούν χρήσιμες μεταβλητές για την ταξινόμηση. Στη σημερινή τεχνολογία, δεδομένα σε χώρο υψηλών διαστάσεων προκύπτουν σε πολλές επιστημονικές εφαρμογές όπως στη βιολογία, στην αστρονομία, στα οικονομικά και στην επιστήμη των υπολογιστών. Συγκεκριμένα, ο αριθμός των μεταβλητών μπορεί να είναι μεγαλύτερος από το μέγεθος του δείγματος, γεγονός που οδηγεί τη στατιστική ανάλυση μπροστά σε μια πρόκληση. Αν μόνο ορισμένες μεταβλητές είναι χρήσιμες για την ταξινόμηση, η υπερπροσαρμογή είναι ένα πρόβλημα για τις μεθόδους που χρησιμοποιούν όλες τις μεταβλητές. Συνεπώς, η επιλογή μεταβλητών γίνεται ιδιαίτερα κρίσιμη για την στατιστική ανάλυση.

Οι μέθοδοι για την επιλογή μεταβλητών στο πλαίσιο της ταξινόμησης περιλαμβάνουν μεθόδους που ενσωματώνουν την επιλογή μεταβλητών μέσα στη διαδικασία της ταξινόμησης. Αυτή η κλάση περιλαμβάνει την Randomforest, την CART και την GUIDE. Η μέθοδος Random forest αναθέτει ένα βαθμό σημαντικότητας σε καθένα από τις ανεξάρτητες μεταβλητές και έτσι ο αναλυτής μπορεί να παραβλέψει εκείνες τις μεταβλητές των οποίων η σημαντικότητα βρίσκεται κάτω από μια συγκεκριμένη τιμή. Η μέθοδος CART, μετά από ένα ξεκαθάρισμα, θα επιλέξει ένα υποσύνολο μεταβλητών που θα θεωρηθούν ως οι πιο σημαντικές. Η μέθοδος GUIDE είναι μια μέθοδος που βασίζεται σε δεντροδιάγραμμα, η οποία θεωρείται ιδιαίτερα σημαντική σε αμερόληπτη επιλογή μεταβλητών και στην ανίχνευση αλληλεπιδράσεων μεταξύ των μεταβλητών. Άλλες έρευνες περιλαμβάνουν μεθόδους που ενσωματώνουν την επιλογή μεταβλητών, εφαρμόζοντας μεθόδους συρρίκνωσης με L_1 νόρμες ως περιορισμούς στις παραμέτρους. Αποτέλεσμα αυτής της διαδικασίας είναι η

παραγωγή διανυσμάτων ως εκτιμητές των παραμέτρων. Οι Wang, Shen και Zhang ενσωμάτωσαν την επιλογή μεταβλητών στην ταξινόμηση βασιζόμενοι σε μεθόδους μηχανών διανυσμάτων υποστήριξης (Support Vector Machine, SVM). Ο Qiao θεωρεί ότι η επιλογή μεταβλητών βασίζεται στη γραμμική διακριτική ανάλυση. Ένας περιορισμός σε αυτές τις μεθόδους επιλογής μεταβλητών είναι ότι βασίζονται σε μη παραμετρικές μεθόδους. Συνεπώς, ενδέχεται να μην ανταποκρίνονται επαρκώς όταν υπάρχει σύνθετη σύνδεση μεταξύ των ανεξάρτητων μεταβλητών προς ταξινόμηση.

3.8.1 Εισαγωγή στη μέθοδο CATCH

Μια μη παραμετρική μέθοδος για επιλογή μεταβλητών είναι η μέθοδος Categorical Adaptive Tube Covariate Hunting (CATCH). Η μέθοδος αυτή παρουσιάζει έναν αλγόριθμο που μπορεί να χρησιμοποιηθεί προκειμένου να βελτιώσει την παρουσίαση των διαθέσιμων διαδικασιών ταξινόμησης. Η βασική ιδέα αυτής της μεθόδου είναι η δόμηση ενός μη παραμετρικού μέτρου της σχετικής δύναμης μεταξύ κάθε ανεξάρτητης μεταβλητής και της αντίστοιχης κατηγορηματικής απόκρισης. Στη συνέχεια σκοπός είναι η διατήρηση εκείνων των προγνωστών των οποίων η σχέση με την απόκριση είναι μεγαλύτερη από μια προεπιλεγμένη τιμή. Το μη παραμετρικό μέτρο της σημαντικότητας κάθε προγνώστη μπορεί να αντληθεί από την πρώτη μέτρηση του προγνώστη χρησιμοποιώντας τις σχετικές πληροφορίες. Έπειτα, συνδιάζοντας όλες τις σχετικές τιμές σημαντικότητας, προκύπτει μια συνολική τιμή σημαντικότητας.

Πέρα από το προαναφερθέν μη παραμετρικό χαρακτηριστικό, η διαδικασία CATCH έχει επιπλέον μια ιδιότητα. Η μέθοδος αυτή μετράει την σημασία κάθε μεταβλητής εξαρτώμενης από όλες τις άλλες μεταβητές. Ως εκ τούτου, μειώνει τη σύγχυση που ενδέχεται να οδηγήσει σε λανθασμένη επιλογή μεταβλητών συνδεδεμένων με την κατηγορηματική μεταβλητή Y. Αυτό επιτυγχάνεται με περιορισμό όλων των προγνωστών εκτός αυτού στον οποίο επικεντρώνεται το ενδιαφέρον του αναλυτή. Στην περίπτωση που ο αριθμός των ανεξάρτητων μεταβλητών είναι πολύ μεγάλος, αυτό μπορεί να γίνει περιορίζοντας τις κύριες συνιστώσες για μερικούς τύπους αναλύσεων.

Η προσέγγιση της μεθόδου CATCH σχετίζεται με τον αλγόριθμο EARTH, ο οποίος εφαρμόζεται σε μη παραμετρικά προβλήματα παλινδρόμησης με συνεχή μεταβλητή απόκρισης και συνεχείς ανεξάρτητες μεταβλητές.Μετράει την υποθετική σύνδεση μεταξύ μιας ανεξάρτητης μεταβλητήςΙκαι της μεταβλητής απόκρισης. Αυτό εφαρμόζεται σε όλες τις ανεξάρτητες μεταβλητές$\{j\}_{j\neq i}$, περιορίζοντας τις $\{j\}_{j\neq i}$ μεταβλητές σε περιοχές που καλούνται «tubes». Η τοπική τιμή σημαντικότητας βασίζεται σε τοπική γραμμική ή τοπική πολυωνυμική παλινδρόμηση.

Ο αλγόριθμος CATCH μπορεί να χρησιμοποιηθεί σαν ένα βήμα επιλογής μεταβλητών πριν την ταξινόμηση. Οποιαδήποτε μέθοδος ταξινόμησης, κατά προτίμηση οι μη παραμετρικές, μπορούν να χρησιμοποιηθούν αφού ξεχωρίσουμε τις

σημαντικές μεταβλητές. Συγκεκριμένα, οι μέθοδοι SVM και Random forest είναι στατιστικές μέθοδοι ταξινόμησης που μπορούν να χρησιμοποιηθούν με τη μέθοδο CATCH. Κάποιες μελέτες προσομοίωσης δείχνουν ότι όταν το πραγματικό μοντέλο είναι μη γραμμικό κι έτσι υπάρχουν πολλές ασυσχέτιστες μεταβλητές, η χρήση του αλγόριθμου CATCH υποδεικνύει τους ασυσχέτιστες ή αδύναμες μεταβλητές. Αυτή η διαδικασία βελτιώνει ιδιαίτερα την παρουσίαση των μεθόδων SVM και Rand forest.

Ο αλγόριθμος CATCH λειτουργεί με δεδομένα γενικής ταξινόμησης, με αριθμητικούς και κατηγορηματικές ανεξάρτητες μεταβλητές. Επιπλέον, αυτός ο αλγόριθμος είναι ιδιαίτερα αποδοτικός όταν οι ανεξάρτητες μεταβλητέςαλληλεπιδρούν η μία με την άλλη. Είναι, επίσης, ιδιαίτερα σημαντικό πως αυτός ο αλγόριθμος επιτυγχάνει μεγαλύτερη ακρίβεια στην επιλογή μεταβλητών σε σύγκριση με τον CART και τον GUIDE. Αυτό οφείλεται στους περιορισμούς που τίθενται στις τιμές σημαντικότητας των προγνωστών.

3.8.2 Βαθμολόγηση σημασίας για μονοπαραγοντική ταξινόμηση

Έστω $(X^{(i)}, Y^{(i)})$, I = 1,.. ,n ανεξάρτητα και ταυτόσημα κατανεμημένα καθώς (X,Y) ~ P. Αρχικά θεωρούμε την περίπτωση μονοπαραγοντικού X. Στη συνέχεια γίνεται χρήση των μεθόδων που έχουν κατασκευαστεί για την μονοπαραγοντική περίπτωση προκειμένου να δομηθούν οι πολυπαραγοντικές εκδοχές.

Αριθμητικός προγνώστης: Τοπική ενδεχόμενη αποτελεσματικότητα

Θεωρούμε το πρόβλημα ταξινόμησης με αριθμητική (ποσοτική) ανεξάρτητη μεταβλητή. Έστω τότε

$$\Pr(Y = c \mid x) = p_c(x),\ c = 1,\ldots C,\ \sum_{c=1}^{C} p_c(x) \equiv 1 \qquad (3.18)$$

Σε αυτό το σημείο μπορεί να χρησιμοποιηθεί ένα μέτρο που δείχνει πόσο ισχυρά είναι συνδεδεμένα τα X, Y στην περιοχή ενός σταθερού σημείου $x^{(0)}$.Σημεία τέτοια, επιλέγονται τυχαία με τον αλγόριθμο CATCH.Η τοπική αποτελεσματικότητα θα υπολογιστεί για κάθε σημείο χωριστά και στη συνέχεια κατά μέσο όρο. Για σταθερό h> 0, προσδιορίζουμε την περιοχή γύρω από το $x^{(0)}$ ως εξής:

$$N_h(x^{(0)}) = \{x : 0 < \left| x - x^{(0)} \right| \leq h\} \qquad (3.19)$$

Επιπλέον, θεωρούμε πως ο αριθμός των σημείων δεδομένων δηλώνεται από

$$n(x^{(0)}, h) = \sum_{i=1}^{n} I(X^{(i)} \epsilon N_h(x^{(0)})) \qquad (3.20)$$

Για c=1,…,C έστω

$$n_c^-\left(x^{(0)},h\right)=\sum_{c=1}^{C} n_c^-\left(x^{(0)},h\right) \tag{3.21}$$

Και

$$n_c^+\left(x^{(0)},h\right)=\sum_{c=1}^{C} n_c^+\left(x^{(0)},h\right) \tag{3.22}$$

δηλώνει τον αριθμό των παρατηρήσεων $(X^{(i)},Y^{(i)})$ που ικανοποιούν τις συνθήκες

$x^{(0)}-h<X^{(i)}\leq x^{(0)}$και $Y^{(i)}$ = c. Επιπλέον, με $n_c^+\left(x^{(0)},h\right)$ θεωρείται ο αριθμός των $(X^{(i)},Y^{(i)})$ παρατηρήσεων που ικανοποιούν τις σχέσεις

$x^{(0)}<X^{(i)}\leq x^{(0)}+h$ και $Y^{(i)}$ = c. Ισχύουν,επίσης, οι ισότητες :

$n_c\left(x^{(0)},h\right)=n_c^+\left(x^{(0)},h\right)+n_c^-\left(x^{(0)},h\right)$, $n_c^-\left(x^{(0)},h\right)=\sum_{c=1}^{C} n_c^-\left(x^{(0)},h\right)$ και $n_c^+\left(x^{(0)},h\right)=\sum_{c=1}^{C} n_c^+\left(x^{(0)},h\right)$. Ο παρακάτω πίνακας παρουσιάζει τα ενδεχόμενα που προκύπτουν.

Πίνακας 3.1: Πίνακας τοπικών ενδεχομένων

	Y = 1	Y = 2	……	Y = C	Σύνολο
$x^{(0)}-h<X\leq x^{(0)}$	$n_1^-\left(x^{(0)},h\right)$	$n_2^-\left(x^{(0)},h\right)$		$n_C^-\left(x^{(0)},h\right)$	$n^-\left(x^{(0)},h\right)$
$x^{(0)}<X\leq x^{(0)}+h$	$n_1^+\left(x^{(0)},h\right)$	$n_2^+\left(x^{(0)},h\right)$		$n_C^+\left(x^{(0)},h\right)$	$n^+\left(x^{(0)},h\right)$
Σύνολο	$n_1\left(x^{(0)},h\right)$	$n_2\left(x^{(0)},h\right)$		$n_C\left(x^{(0)},h\right)$	$n\left(x^{(0)},h\right)$

Προκειμένου να εκτιμηθεί κατά πόσο συσχετίζεται το Y με τους τοπικούς περιορισμούς του X, θεωρούμε ως ισχύουσα την παρακάτω σχέση:

$$X^2=\sum_{c=1}^{C}\left(\frac{\left(n_c^-\left(x^{(0)},h\right)-E_c^-\left(x^{(0)},h\right)\right)^2}{E_c^-\left(x^{(0)},h\right)}+\frac{\left(n_c^+\left(x^{(0)},h\right)-E_c^+\left(x^{(0)},h\right)\right)^2}{E_c^+\left(x^{(0)},h\right)}\right) \tag{3.23}$$

Όπου $0/0\equiv 0$ και

$$E_c^-\left(x^{(0)},h\right)=\frac{n^-\left(x^{(0)},h\right)\,n_c\left(x^{(0)},h\right)}{n\left(x^{(0)},h\right)} \tag{3.24}$$

$$E_c^+\left(x^{(0)},h\right)=\frac{n^+\left(x^{(0)},h\right)\,n_c\left(x^{(0)},h\right)}{n\left(x^{(0)},h\right)} \tag{3.25}$$

Εξαιτίας του τοπικού περιορισμού στο X, υπάρχει περίπτωση ο πίνακας της τοπικής πιθανοτηταςνα περιέχει μηδενικά κελιά. Παρόλα αυτά, δεν υπάρχει πρόβλημα για την X^2στατιστική. Η X^2 μπορεί να ανιχνεύσει τοπική εξάρτηση μεταξύ των X και Y στην περιοχή του $x^{(0)}$. Όταν όλες οι παρατηρήσεις ανήκουν στην ίδια κατηγορία του Y, διαισθητικά δεν δύναται να ανιχνευθεί καμία εξάρτηση. Σε αυτή την περίπτωση η X^2στατιστική είναι ισοδύναμη με το μηδέν.Θεωρούμε λοιπόν ένα τμήμα $N_h\,(x^{(0)})$ της περιοχής του $x^{(0)}$ και μεγιστοποιούμε το $X^2((x^{(0)},h))$ σε σχέση με το μέγεθος htου τμήματος. Με "plim" δηλώνεται το όριο στην πιθανότητα.

Όσον αφορά στο τοπικό μέτρο συσχέτισης ζ $(x^{(0)})$ αυτό προσδιορίζεται από τις παρακάτω σχέσεις :

$$\zeta\,(x^{(0)},h) = plim_{n\to\infty} n^{-\frac{1}{2}}\sqrt{X^2((x^{(0)},h))}\ ,\ \zeta\,(x^{(0)}) = sup_{h>0}\{\,\zeta\,(x^{(0)},h)\,\} \qquad (3.26).$$

Στην περίπτωση που p_c(x) είναι σταθερή στο x για όλα τα c∈ {1,…,C} για x κοντά στο $x^{(0)}$, τότε, καθώς n→∞, ισχύει ότι $X^2((x^{(0)},h))$ συγκλίνει στην κατανομή χ^2_{C-1}. Σε αυτή την περίπτωση έχουμε ζ $(x^{(0)},h)$ = 0. Από την άλλη μεριά, αν η p_c(x) δεν είναι σταθερή στην περιοχή κοντά στο $x^{(0)}$, τότε η ζ $(x^{(0)},h)$ ορίζει την ασυμπτωτική δύναμη της εναλλακτικής λύσης του κριτηρίου που βασίζεται στο $X^2\left((x^{(0)},h)\right)$. Η συγκεκριμένη εναλλακτική ερευνά αν το X είναι τοπικά ανεξάρτητο του Y και καλείται κριτήριο αποτελεσματικότητας («efficacy test»). Επιπρόσθετα, η έκφραση $\zeta^2(x^{(0)},h)$ δηλώνει την παράμετρο της στατιστική $n^{-1}X^2\left((x^{(0)},h)\right)$. Οι εκτιμητές των ζ $(x^{(0)},h)$ και ζ $(x^{(0)})$ είναι :

$$\hat{\zeta}(x^{(0)},h) = n^{-\frac{1}{2}}\sqrt{X^2((x^{(0)},h))}\ ,\ \hat{\zeta}\,(x^{(0)}) = sup_{h>0}\{\,\hat{\zeta}\ (x^{(0)},h)\,\} \qquad (3.27).$$

Το μέγεθος h της περιοχής έχει επιλεχθεί ως εξής:

$$\hat{h} = \text{argmax}\ \{\hat{\zeta}\,(x^{(0)}) = sup_{h>0}\{\,\hat{\zeta}\ (x^{(0)},h) :\ h \in \{h_1,\dots,h_g\}\} \qquad (3.28)$$

για ένα σύνολο $\{h_{1,\dots,}h_g\}$. Το σύνολο αυτό αντιστοιχεί στην επιλογή ενός h με τη μεγιστοποίηση της εκτιμώμενης δύναμης.

Όσον αφορά στην κατηγορική ανεξάρτητη μεταβλητή, την ενδεχόμενη αποτελεσματικότητα (contingency efficacy), υποθέτουμε και πάλι το σύνολο ($X^{(i)},Y^{(i)}$), i=1,…,n εκτός του X ∈ {1,.., C' }. Έστω $n(c,c')$ αριθμός των παρατηρήσεων που ικανοποιούν τις συνθήκες Y =c και X = c'. Έτσι, θεωρούμε

$$X^2 = \sum_{c=1}^{C}\sum_{c'=1}^{C'}\frac{(n(c,\ c')-E(c,\ c'))^2}{E(c,\ c')^2} \qquad (3.29)$$

και

$$E(c,c') = \frac{\sum_{c=1}^{C} n(c,\ c') \sum_{c'=1}^{C'} n(c,\ c')}{n} \tag{3.30}$$

Σε αυτό το σημείο είναι χρήσιμο να δοθεί ο ορισμός της ενδεχόμενης αποτελεσματικότητας για κατηγορηματική ανεξάρτητη μεταβλητή (contingency efficacy for categorical predictor). Η ενδεχόμενη αποτελεσματικότητα (CE) του Y μιας κατηγορηματικής ανεξάρτητης μεταβλητής X, δίνεται από τη σχέση

$$\zeta = plim_{n\to\infty} n^{-1}\sqrt{X^2} \tag{3.31}$$

Στην περίπτωση της μηδενικής υπόθεσης (null hypothesis) τα X και Y είναι ανεξάρτητα μεταξύ τους και η X^2 συγκλίνει στην κατανομή $\chi^2_{(C-1)(C'-1)}$. Ισχύει, επίσης, ζ = 0. Στη γενική περίπτωση, η ζ^2 είναι η ασυμπτωτική παράμετρος της στατιστικής $n^{-1}X^2$. Η ενδεχόμενη αποτελεσματικότητα ζ καθορίζει την ασυμπτωτική δύναμη του κριτηρίου βάσει της X^2 προκειμένου να ελεγχθεί αν τα X και Y είναι ανεξάρτητα. Χρησιμοποιούμε τη ζ ως μέτρο σημαντικότητας το οποίο δείχνει πόσο ισχυρά εξαρτάται το Y από το X. Ο εκτιμητής της ζ είναι

$$\hat{\zeta} = n^{\frac{-1}{2}}\sqrt{X^2} \tag{3.32}$$

3.9 Επιλογή μεταβλητών για προβλήματα ταξινόμησης με τον αλγόριθμο CATCH

Θεωρούμε μια πολυμεταβλητή X= ($X_{1,\dots,}X_d$), d>1. Προκειμένου να εξεταστεί αν η ανεξάρτητη μεταβλητή X_l συνδέεται με το Y υπό την παρουσία όλων των μεταβλητών, υπολογίζεται η υπό όρους απόδοση σε αυτές τις μεταβλητές καθώς και άνευ περιορισμών απόδοση.

Προκειμένου να ερευνήσουμε την επίδραση του X_l στην τοπική περιοχή του $x^{(0)} = (x_1^{(0)}, \dots . x_d^{(0)})$, κατασκευάζουμε μια περιοχή«σωλήνα»T_l με κέντρο το $x^{(0)}$, στην οποία τα σημεία δεδομένων βρίσκονται σε συγκεκριμένη ακτίνα δ μακρυά από το $x^{(0)}$ σε σχέση με όλες τις μεταβλητές εκτόστουX_l. Στο σημείο αυτό θεωρείται

χρήσιμο να σημειωθεί ο ορισμός της περιοχής «σωλήνα» T_l. Ένας «σωλήνας» T_l, μεγέθους δ (δ>0) για την μεταβλητή X_l με κέντρο το $x^{(0)}$, θεωρείται το σύνολο

$$T_l \equiv T_l\left(x^{(0)}, \delta, D\right) \equiv \left\{x : D\left(x_{-l,} x_{-l}^{(0)} \leq \delta\right)\right\} \tag{3.33}$$

Προσαρμόζουμε το μέγεθος του σωλήνα έτσι ώστε ο αριθμός των σημείων μέσα στον σωλήνα να είναι ένας προκαθορισμένος αριθμός k. Παρατηρείται, γενικά, ότι k≥ 50 είναι μια καλή επιλογή. Να σημειωθεί πως η επιλογή k =n αντιστοιχεί σε οριακή παλινδρόμηση. Στο εσωτερικό του σωλήνα το X_l είναι ανεξάρτητο. Προκειμένου να ερευνήσουμε την εξάρτηση του Y από το X_l, θεωρούμε περιοχές του $x^{(0)}$ κατά μήκος της κατεύθυνσης του X_l μέσα στον σωλήνα. Έστω

$$N_{l,h,\delta,D}(x^{(0)}) = \{x = (x_1, .., x_d)^T \in \mathbf{R}^d : x \in T_l\left(x^{(0)}, \delta, D\right), \left|x_l - x_l^{(0)}\right| \leq h\} \tag{3.34}$$

ένα τμήμα του σωληνα$T_l\left(x^{(0)}, \delta, D\right)$. Για να μελετηθεί πόσο ισχυρά συνδέονται τα X_l, Y μέσα σε αυτό το τμήμα, θεωρούμε

$$X_{l,\delta,D}^2(x^{(0)}, h) \tag{3.35}$$

όπου η (3.35) ορίζεται μέσω της σχέσης (3.34) χρησιμοποιώντας μόνο τις παρατηρήσεις μέσα στο σωλήνα $T_l\left(x^{(0)}, \delta, D\right)$. Όσον αφορά στην τοπική ενδεχόμενη αποτελεσματικότητα της περιοχής του σωλήνα (local contingency tube section efficacy) και την τοπική ενδεχόμενη αποτελεσματικότητα της περιοχής του σωλήνα του Y στο X_l, ορίζονται από τις παρακάτω σχέσεις:

$$\zeta(x^{(0)}, h, l) = plim_{n\to\infty} n^{\frac{-1}{2}} \sqrt{X_{l,\delta,D}^2(x^{(0)}, h)}\ ;\ \zeta_l(x^{(0)}) = sup_{h>0}\{\ \zeta(x^{(0)}, h, l)\ \} \tag{3.36}$$

$$\hat{\zeta}(x^{(0)}, h, l) = \sqrt{X_{l,\delta,D}^2(x^{(0)}, h)};\ \hat{\zeta}(x^{(0)}) = sup_{h>0}\{\hat{\zeta}\ (x^{(0)}, h, l)\ \} \tag{3.37}$$

Για μεγαλύτερη αποσαφήνιση της «local contingency tube efficacy»θεωρούμε ότι Y $\in \{1,2\}$ και την πιθανότητα P (Y = 2| x) = $(x - \frac{1}{2})^2$, $x \in [0,1]$. Έτσι, στη συνέχεια η τοπική αποτελεσματικότητα με μικρό h θα είναι μεγάλη, ενώ για h≥ 1 θα είναι μηδενική.

Στην περίπτωση που το X_l είναι κατηγορηματικό ενώ κάποιοι άλλοι προγνώστες είναι αριθμητικοί, δημιουργούμε σωλήνες βασισμένους στους αριθμητικούς προγνώστες αφήνοντας τις κατηγορηματικές μεταβλητές. Στη συνέχεια, χρησιμοποιούμε τη στατιστικήπροκειμένου να ερευνήσουμε τη δύναμη συσχέτισης μεταξύ των Y και X_l

μέσα στο σωλήνα με τη βοήθεια μόνο των παρατηρήσεων που βρίσκονται μέσα σε αυτόν.

Η τοπική ενδεχόμενη αποτελεσματικότητα σωλήνα του Y στο X_l και η εκτίμησή της δίνονται από τις σχέσεις:

$$\zeta_l(x^{(0)}) = plim_{n\to\infty} n^{\frac{-1}{2}} \sqrt{X^2_{l,\delta,D}(x^{(0)})} \; ; \; \widehat{\zeta_l}((x^{(0)}) = n^{\frac{-1}{2}} \sqrt{X^2_{l,\delta,D}(x^{(0)})} \tag{3.38}$$

Οι εμπειρικές αποδόσεις μετρούν πόσο ισχυρά εξαρτάται το Y από το X_l κοντά στο σημείο $(x^{(0)})$. Για να μετρήσουμε την ολική εξάρτηση του Y από το X_l, δοκιμάζουμε τυχαία M «boostrap» σημείαx^*_a , α=1,...,M με αντικατάσταση από $\{x^{(i)} : 1 \leq i \leq n\}$ και υπολογίζουμε το μέσο όρο των τοπικών αποδόσεων σωλήνα σε αυτά τα σημεία.

Ο βαθμός εμπειρικής σημαντικότητας CATCH (the CATCH empirical importance score) για τη μεταβλητή l είναι:

$$s_l = M^{-1} \sum_{a=1}^{M} \widehat{\zeta_l}(x^*_a) \tag{3.39}$$

όπου $\widehat{\zeta_l}$ ($(x^{(i\alpha)}$) ορίζεται από τις παραπάνω σχέσεις για κατηγορηματικούς και αριθμητικές ανεξάρτητες μεταβλητές, αντίστοιχα.

Προσαρμοστικά διαστήματα σωλήνα (adaptive tubed istances)

Ο Doksum χρησιμοποίησε ένα παράδειγμα προκειμένου να αποδείξει ότι το διάστημα D χρειάζεται να είναι προσαρμοστικό έτσι ώστε οι μεταβλητές να διατηρούνται ισχυρά συνδεδεμένες με το Y. Σκοπός της διατήρησης αυτής της σύνδεσης είναι η απόφυγη επιρροής αδύναμων μεταβλητών. Έστω ότι το Y εξαρτάται από τρεις μεταβλητές X_1 , X_2, X_3 ανάμεσα στα μεγαλύτερα σύνολα προγνωστών. Το μέγεθος της εξάρτησης κυμαίνεται από ασθενής σε ισχυρή, το οποίο θα εκτιμηθεί από την εμπειρική σημαντικότητα. Καθώς ερευνάται η επιρροή της X_2 πάνω στο Y, είναι σημαντικός ο προσδιορισμός του σωλήνα έτσι ώστε να είναι σχετικά μικρότερη η διακύμανση της X_3 σε σχέση με τη διακύμανση της X_1 μέσα στο σωλήνα. Δημιουργείται αυτή η ανάγκη διότι είναι περισσότερο πιθανό η X_3 να επισκιάσει την επιρροή της X_2 σε σχέση με της X_1.

Ορίζουμε, λοιπόν,

$X_{-l} = (X_{1,\ldots,} \; X_{-l}, X_{l+1}, \ldots X_d)^T$να είναι το συμπληρωματικό διάνυσμα του X_l. Έστω D(°, °) να είναιένα διάστημα μέσα στο R^{d-1}.Βάσει του τελευταίου μπορούμε να εισάγουμε ένα διάστημα σωλήνα για την μεταβλητή X_lτης μορφής

$$D_l\,(x_{-l}, x^{(0)}_{-l}) = \sum_{j=1}^{d} w_j \left| x_j - \; x_j^{(0)} \right| \text{I (j}\neq l) \tag{3.40}$$

όπου το ‘βάρος’ w_j ρυθμίζει τη συνεισφορά των ισχυρών μεταβλητών στο διάστημα, έτσι ώστε οι ισχυρές μεταβλητές να είναι περισσότερο περιορισμένες.Ταw_j είναι ανάλογα των διαφορών s_j - s_j'. Για τη δεύτερη επανάληψη, χρησιμοποιούμε τα $s_j = s_j^{(2)}$ και το διάστημα

$$D_l\ (x_{-l}, x_{-l}^{(0)}) = \frac{\sum_{j=1}^{d}(s_j - s_j') + \left|x_j - x_j^{(0)}\right| \text{I (j}\neq l)}{\sum_{j=1}^{d}(s_j - s_j') + \text{I (j}\neq l)} \tag{3.41}$$

όπου $0/0 \equiv 1$ καιs_j' είναι μια οριακή τιμή του s_j κάτω από την υπόθεση άνευ όρων σύνδεσης του X_l με το Y υπολογισμένου με απλή μέθοδο Monte Carlo. Η τελευταία προσαρμογή είναι σημαντική διότι κάποιες ανεξάρτητες μεταβλητές εξ ορισμού έχουν μεγαλύτερα σκορ σημαντικότητας σε σχέση με άλλους όταν είναι όλοι τους ανεξάρτητοι του Y. Για παράδειγμα, ένας κατηγοριματικός προγνώστης με περισσότερα επίπεδα έχει μεγαλύτερο σκορ σημαντικότητας σε σχέση με προγνώστη λιγότερων επιπέδων. Συνεπώς, τα μη προσαρμοσμένα σκορ s_j δεν μπορούν να ποσοτικοποιήσουν με ακρίβεια τη σχετική σημαντικότητα των προγνωστών στο σωλήνα που μελετάται. Στα επόμενα χρησιμοποιούνται οι παραπάνω σχέσεις για την παραγωγή $s_j = s_j^{(3)}$ και χρησιμοποιούμε το $s_j^{(3)}$στην επόμενη επανάληψη και ούτω καθεξής.

Στον παραπάνω ορισμό, χρειάζεται να οριστεί κατάλληλα η απόλυτη διαφορά $\left|x_j - x_j^{(0)}\right|$ για κατηγορηματική μεταβλητή X_j. Επειδή ακριβώς αυτή η μεταβλητή είναι κατηγορηματική, χρειάζεται να γίνει η υπόθεση ότι η απόλυτη διαφορά$\left|x_j - x_j^{(0)}\right|$ είναι σταθερή όταν τα x_j, $x_j^{(0)}$ ανήκουν σε οποιοδήποτε ζεύγος διαφορετικών κατηγοριών. Μια προσέγγιση είναι να ορίσουμε

$\left|x_j - x_j^{(0)}\right| = \infty I\ (x_j \neq x_j^{(0)})$. Σε αυτή την περίπτωση, σε όλα τα σημεία εντός του σωλήνα δίνεται η ίδια τιμή X_j όπως και στο κέντρο του σωλήνα. Αυτή η προσέγγιση είναι ιδιαίτερα ευαίσθητη καθώς υλοποιεί αυστηρά την ιδέα οριοθέτησης του X_j όταν αξιολογείται η επίδραση μιας άλλης ανεξάρτητης μεταβλητής X_l και το X_j δεν συνεισφέρει σε οποιαδήποτε μεταβολη του Y. Ωστόσο, το πρόβλημα με αυτόν τον ορισμό είναι πως όταν ο αριθμός των κατηγοριών του X_j είναι σχετικά μεγάλος σε σύγκριση με το μέγεθος του δείγματος, δεν θα υπάρχουν αρκετά σημεία μέσα στο σωλήνα αν περιορίσουμε το X_j σε μία μόνο κατηγορία. Σε αυτή την περίπτωση, δεν περιορίζουμε το X_j που σημαίνει ότι το X_j μπορεί να λάβει οποιαδήποτε κατηγορία μέσα στο σωλήνα. Τότε ορίζουμε $\left|x_j - x_j^{(0)}\right| = k_0 I\ (x_j \neq x_j^{(0)})$, όπου k_0 είναι μια σταθερά ομαλοποίησης. Η τιμή της σταθεράς k_0 επιλέγεται έτσι ώστε να επιτευχθεί η ισορροπία μεταξύ αριθμητικών και κατηγορηματικών μεταβλητών. Ας υποθέσουμε ότι η X_1 είναι μια αριθμητική μεταβλητή με σταθερή απόκλιση ένα και η X_2 είναι

μιακατηγορηματική μεταβλητή. Για την X_2, θέτουμε $k_0 \equiv$ E ($\left|X_1^{(1)} - X_1^{(2)}\right|$), όπου $X_1^{(1)}$ και $X_1^{(2)}$ είναι δύο διαφορετικές υλοποιήσεις της X_1. Να σημειωθεί ότι χρησιμοποιείται $k_0 = \sqrt{2/\pi}$, αν η X_1 ακολουθεί σταθερή ομαλή κατανομή. Επίσης, η k_0 μπορεί να βασιστεί σε εμπειρικές κατανομές των αριθμητικών μεταβλητών. Βάσει όσων αναφέρθηκαν για την επιλογή του k_0, να σημειωθεί πως χρησιμοποιούμε $k_0 = \infty$ στις προσομοιώσεις και $k_0 = \sqrt{2/\pi}$ στο παράδειγμα πραγματικών δεδομένων.

3.10 Ο αλγόριθμος CATCH

Ο αλγόριθμος επιλογής μεταβλητών για ένα πρόβλημα ταξινόμησης βασίζεται στα σκορ σημαντικότητας και σε μια επαναληπτική διαδικασία.

<u>Ο αλγόριθμος</u>

- **(0) Τυποποίηση:** Τυποποίησηόλων των αριθμητικών προγνωστών προκειμένου να έχουν μέση τιμή 0 και σταθερή απόκλιση δείγματος 1. Για να επιτευχθεί το παραπάνω χρησιμοποιούνται γραμμικοί μετασχηματισμοί.
- **(1) Προετοιμασία των σκορ σημαντικότητας:** Έστω S = $(s_1, \dots, s_d)$ το σύνολο των σκορ σημαντικότητας για τις ανεξάρητες μεταβλητές. Έστω S' = $(s'_{1,\dots}s'_d)$ τέτοιο ώστε s'_lνα είναι ένα οριακό σκορ σημαντικότητας το οποίο αντιστοιχεί στο μοντέλο όπου τα X_l και Y είναι ανεξάρτητα, δεδομένου X_{-l} . Χρησιμοποιείται, επίσης, το S ως (1,..1) και το S' ως (0,...,0). Έστω M ο αριθμος των κέντρων του χρησιμοποιούμενου σωλήνα και m ο αριθμός των σημείων που βρίσκονται μέσα σε κάθε σωλήνα.
- **(2) Αναζήτηση βρόχου σωλήνα:** Για l = 1,..,d εκτελούνται τα βήματα (a), (b), (c) και (d) παρακάτω:
    - (a) **Επιλογή κέντρων σωλήνα:** Για0<b≤ 1, θέτουμε το σύνολο M= [nb] , όπου [ ] είναι η μεγαλύτερη ακέραια συνάρτηση. Θεωρούμε, επίσης, τα 'boostrap 'σημεία$X_1^*, .., X_M^*$ του τυχαίου δείγματος M με αντικατάσταση από n παρατηρήσεις. Θέτουμε k = 1 και εκτελούμε το (b).
    - (b) **Δόμηση σωλήνων:** Για 0<a<1 ,θέτουμε m = [na]. Χρησιμοποιώντας το διάστημα σωληνα D_l που έχει οριστεί παραπάνω, επιλέγεται το μέγεθος a του σωλήνα ώστε να υπάρχουν ακριβώς m≥ 10παρατηρήσεις στο εσωτερικό του σωλήνα για την X_lστο X_k^*.
    - (c) **Υπολογισμός αποτελεσματικότητας:**Αν η X_l είναι αριθμητική μεταβλητή, έστω $\widehat{\zeta_l}(X_k^*)$ ότι είναι η εκτιμώμενη τοπική ενδεχόμενη αποτελεσματικότητα του Y πάνω στο X_l στο σημείο X_k^* μέσα στο

σωλήνα. Αν η X_l είναι μια κατηγορική μεταβλητή έστω ότι $\widehat{\zeta_l}(X_k^*)$ είναι η τοπική ενδεχόμενη αποτελεσματικότητα.

(d) **Οριακές τιμές:** Έστω ότι με Y^* δηλώνεται μια τυχαία μεταβλητή η οποία είναι κατανεμημένη όπως ακριβώς η Y, αλλά είναι ανεξάρτητη της X. Έστω ($Y_0^{(1)},..,Y_0^{(n)}$) να είναι μια τυχαία μετάθεση της ($Y^{(1)},\dots,Y^{(n)}$). Έπειτα, το $Y_0 \equiv (Y_0^{(1)},..,Y_0^{(n)})$ μπορεί να θεωρηθεί σαν πραγματοποίηση της Y^*.Έστω ότι $\widehat{\zeta_l^*}\ (X_k^*)$ είναι η εκτιμώμενη τοπική αποτελεσματικότητα σωλήνα του Y_0 πάνω στο X_l στο σημείο X_k^* που έχει υπολογιστεί όπως στο βήμα (c) με το Y_0 στη θέση του Y.

Αν k<M , θέτουμε k = k +1 και επιστρέφουμε στο βήμα (b). Διαφορετικά, προχωράμε στο βήμα (e).

Ενημέρωση των σκορ σημαντικότητας: Θέτουμε $s_l^{new} = \frac{1}{M}\sum_{k=1}^{M}\widehat{\zeta_l}(X_k^*)$, $s_l'^{new} = \frac{1}{M}\sum_{k=1}^{M}\widehat{\zeta_l^*}\ (X_k^*)$. Έστω $S^{new} = (\ s_1^{new},\dots,s_d^{new}\ )$ και S'^{new} =($s_1'^{new},\dots,s_d'^{new}$) ότι δηλώνουν τις ενημερώσεις των S και S'.

(3) Επαναλήψεις και κανόνας τέλους: Επαναλαμβάνουμε το βήμα (2) iφορές. Σταματάμε όταν η αλλαγή στο S^{new} είναι μικρή. Θα πρέπει να καταγραφούν τα τελευταία Sκαι S', όπως και τα S^{stop}και S'^{stop}, αντίστοιχα.

(4) Βήμα διαγραφής: Χρησιμοποιώντας τα S^{stop} και S'^{stop}, διαγράφουμε το X_l στην περίπτωση που $s_l^{stop} - s_l'^{stop} < \sqrt{2}\,\mathrm{SD}\,(s_l'^{stop})$. Αν το d είναι μικρό, μπορούμε να παράγουμε μερικά σύνολα του S'^{stop} και στη συνέχεια να χρησιμοποιήσουμε την απόκλιση του δείγματος όλων των τιμών σε αυτά τα σύνολα ως SD(s'^{stop}).

Τέλος αλγορίθμου.

Κεφάλαιο 4

D – Βέλτιστοι σχεδιασμοί για επιλογή μεταβλητών

4.1 Εισαγωγή

Προκειμένου να βρεθούν εκείνες οι παρατηρήσεις που δίνουν τις περισσότερες πληροφορίες για την επιλογή μεταβλητών στην ταξινόμηση, αναζητείται ένας D-βέλτιστος σχεδιασμός στο σύνολο των δεδομένων. Οι επιλεγμένες παρατηρήσεις χρησιμοποιούνται σαν βάση για την επιλογή μεταβλητών. Θεωρούμε έναν πίνακα σχεδιασμού $\boldsymbol{X}$ κύριων επιδράσεων. Σε αντίθεση με τον συνηθισμένο πειραματικό σχεδιασμό δεν μπορούν όλοι οι αθαίρετοι nxk πίνακες να επιλεγούν ως σχεδιασμοί αλλά οι $\binom{n}{k}$πιθανοί σχεδιασμοί μεγέθους k (όπου k<<n), μπορούν να αποκτηθούν από τα δεδομένα εκ παρατήρησης. Αυτό σημαίνει πως όλο το σύνολο των δεδομένων μπορεί να θεωρείται σαν σύνολο υποψήφιων σημείων. Ο D- βέλτιστος σχεδιασμός είναι αυτός με την υψηλότερη D–τιμή των $\binom{n}{k}$ πιθανών σχεδιασμών. Για ένα τυποποιημένο πίνακα σχεδιασμού $\boldsymbol{X}^*$ με σταθερές εισαγόμενες καταχωρήσεις δεδομένων k του πίνακα πληροφορίας, η μέγιστη D- τιμή είναι

$$\mathrm{D}\ (\boldsymbol{X}^*_{opt}) = k^p \qquad (4.1).$$

4.2 Έρευνα για D- βέλτιστο σχεδιασμό σε βάσεις δεδομένων

Ο Pumplün et al. προτείνει να διεξαχθεί μια ευρετική έρευνα για D- βέλτιστους σχεδιασμούς μέσω του γενετικού αλγορίθμου που ακολουθεί:

1. Εξωτερικός βρόχος:επανέλαβε 100 φορές
   (α) επέλεξε τυχαία 10 σχεδιασμούς
   (β) εσωτερικός βρόχος : επανέλαβε 10 φορές
   (γ). Υπολόγισε την D – τιμή για κάθε σχεδιασμό
   (δ). Βελτιστοποίησε τοπικά τους καλύτερους σχεδιασμούς με αλλαγή ή διασταύρωση

2. Επέστρεψε τον καλύτερο σχεδιασμό/ το σύνολο των καλύτερων σχεδιασμών σε κάθε επανάληψη.

Ένας σχεδιασμός αποτελείται από k παρατηρήσεις. Πρώτον, ένα σύνολο 10 σχεδιασμών επιλέγεται τυχαία και για κάθε σχεδιασμό υπολογίζεται η D- τιμή. Προκειμένου να βελτιστοποιηθεί το σύνολο των σχεδιασμών, εκείνοι με τις χαμηλότερες D- τιμές αντικαθιστώνται από πρόσφατα κατασκευασμένους σχεδιασμούς που είναι παρόμοιοι με εκείνους που έχουν τις υψηλότερες D– τιμές. Για την δόμηση των σχεδιασμών χρησιμοποιούνται δύο μέθοδοι : η αλλαγή (mutation) και η διασταύρωση (cross – over). Για λεπτομέρειες παραπέμπουμε στους Pumplün et al. (2005α).

Σε μία παραλλαγή αυτού του αλγορίθμου, όχι μόνο επιστρέφεται ο συνολικά καλύτερος σχεδιασμός, αλλά επίσης και ένα σύνολο σχεδιασμών με την μέγιστη D–τιμή. Η τιμή αυτή προκύπτει από κάθε νέο ξεκίνημα του αλγορίθμου, που σημαίνει 100 επαναλήψεις στον εξωτερικό βρόχο.

Ο τυποποιημένος πίνακας σχεδιασμού $\boldsymbol{X}^*$ χρησιμοποιείται για να υπολογιστεί η D-τιμή. Το τελευταίο γίνεται για να βεβαιωθεί ότι η βελτιστοποίηση οδηγεί σε ορθογωνιότητα και σε ασυσχέτιστες επεξηγηματικές μεταβλητές. Η ορθογωνιότητα των σχεδιασμών που προκύπτει από αυτόν τον αλγόριθμο ερευνάται στην μελέτη προσομοίωσης σε επόμενη ενότητα. Ο αντίστοιχος πίνακας σχεδιασμού $\boldsymbol{X}$ χρησιμοποιείται σαν βάση για την επιλογή μεταβλητών ή και την εξάσκηση των μεθόδων ταξινόμησης.

4.3 Επιλογή μεταβλητών

Το πρόβλημα της επιλογής στοιχείων συνίσταται στην εύρεση του υποσυνόλου των u περισσότερο σχετικών μεταβλητών για διάκριση με u<p. Χρησιμοποιούνται δύο διαφορετικές μέθοδοι επιλογής μεταβλητών:

Συσχέτιση (correlation)

Καταρχήν, υπολογίζονται οι εμπειρικοί συντελεστές συσχέτισης $r\boldsymbol{x}_{.j,y}$ των επεξηγηματικών μεταβλητών $\boldsymbol{X}_{.j}$, j = 1…p, και η μεταβλητή κλάσης **y**. Επιλέγονται εκείνες οι μεταβλητές με τις μεγαλύτερες απόλυτες τιμές. Αν οι συσχετίσεις δύο ή περισσότερων προγνωστών είναι ίσες, παρουσιάζεται μια τυχαία επιλογή. Ένα μειονέκτημα αυτής της προσέγγισης μπορεί να είναι ότι οι συντελεστές αυτής της συσχέτισης μετρούν μόνο τις γραμμικές εξαρτήσεις και έτσι μπορεί να χαθούν σημαντικές μεταβλητές.

Δέντρο (tree)

Η επιλογή στοιχείων με βάση τα δέντρα χρησιμοποιεί τον δείκτη «gini». Καταρχήν, γίνεται γνωστό το δέντρο ταξινόμησης. Κάθε φορά που προκύπτει μία μεταβλητή, στο δέντρο ανατίθεται το βάρος 2^{-d}, όπου το d δηλώνει τον αντίστοιχο κόμβο. Η τιμή 2^{-d} επιλέγεται διότι ένα δυαδικό δέντρο μπορεί να έχει το πολύ 2^d κόμβους στο d- οστό επίπεδο. Οι μεταβλητές που προκύπτουν νωρίς στο δέντρο θεωρούνται περισσότερο σημαντικές από αυτές τις μεταβλητές που βρίσκονται κοντά στα φύλλα. Το μέτρο της σημαντικότητας των στοιχείων είναι το συνολικό βάρος κάθε μεταβλητής. Σε περίπτωση που λιγότερες από u μεταβλητές περιλαμβάνονται στο δέντρο, τα υπόλοιπα στοιχεία επιλέγονται τυχαία. Τα δέντρα απόφασης θεωρούνται ασταθή κι αυτό γιατί πολύ μικρές αλλαγές στα δεδομένα μπορεί να οδηγήσουν σε εντελώς διαφορετικά δέντρα απόφασης. Με αυτό τον τρόπο προκύπτουν ξεχωριστές τιμές σημαντικότητας κάθε μεταβλητής.

Ο Pumplünetal. (2005a) πρότεινε δύο διαφορετικά σχέδια επιλογής μεταβλητών. Από την μια μεριά, η επιλογή μεταβλητών γίνεται βάσει του σχεδιασμού με την υψηλότερη D–τιμή, η οποία αποκαλείται «us doptimal». Από την άλλη μεριά, η επιλογή μεταβλητών γίνεται στη βάση ενός συνόλου 100 σχεδιασμών με τις υψηλότερες D- τιμές. Οι τιμές αυτές προκύπτουν από τις 100 επαναλήψεις στον

εξωτερικό βρόχο του γενετικού αλγορίθμου που χρησιμοποιείται. Για κάθε μεταβλητή, η σχετική συχνότητα της επιλογής κάθε σχεδιασμού ξεχωριστά υπολογίζεται και θεωρείται μέτρο της σημαντικότητας των μεταβλητών. Αυτή η διαδικασία καλείται «us doptimal it» και αναμένεται να δώσει πιο σταθερά αποτελέσματα.

4.4 Εκτέλεση

Στον Pumplünetal. (2005α) οι D- βέλτιστοι σχεδιασμοί χρησιμοποιούνται μόνο για επιλογή μεταβλητών, αλλά οι μέθοδοι ταξινόμησης εκτελούνται σε ολόκληρο το σύνολο δεδομένων. Αυτό είναι χρήσιμο προκειμένου να εκτιμηθεί ξεχωριστά αν οι D- βέλτιστοι σχεδιασμοί είναι αποδοτικοί στην περίπτωση επιλογής μεταβλητών. Αλλά σε αυτή την περίπτωση όπου ο αριθμός των μετρήσεων της κάθε κλάσης είναι περιορισμένος,κάποιος θα μπορούσε φυσιολογικά να χρησιμοποιήσει τους D– βέλτιστους σχεδιασμούς και σαν βάση της εκτέλεσης.

Στη βιβλιογραφία υπάρχουν διάφορες προσεγγίσεις που χρησιμοποιούν D-βέλτιστους σχεδιασμούς σε βάσεις δεδομένων. Οι Rupingκαι Weihs (2009) πρότειναν έναν πειραματικό σχεδιασμό που χρησιμοποιεί πυρήνες σε βάσεις δεδομένων και αυτοί οι σχεδιασμοί χρησιμοποιούνται για την εκπαίδευση της μηχανικής μάθησης για προβλήματα παλινδρόμησης. Η προσέγγισή τους αποδίδει ανταγωνιστικούς αλγορίθμους.

Οι Choueiki και Mount – Campbell (1999) χρησιμοποίησαν το κριτήριο D-βελτιστοποίησης για να επιλεγούν δεδομένα εκτέλεσης ενός νευρωνικού δικτύου για μια προσεγγιστική συνάρτηση, σε περιπτώσεις όπου το αποτέλεσμα είναι ακριβό, ριψοκίνδυνο ή χρονοβόρο. Έδειξαν ότι όσο τα δεδομένα προς εκτέλεση επιλέγονται σύμφωνα με το κριτήριο της D- βελτιστοποίησης, το δίκτυο δύναται να κάνει αποδοτικές γενικεύσεις. Το κριτήριο παρουσίασης που χρησιμοποιήθηκε είναι το μέσο τετραγωνικό σφάλμα.

Ο Manolov (1990) πρότεινε έναν διαδοχικό σχεδιασμό των παρατηρήσεων προς εκτέλεση για ταξινόμηση προκειμένου να βρεθούν στοιχεία κοντά στο πραγματικό σύνορο απόφασης. Ένας D-βέλτιστος σχεδιασμός λαμβάνεται σαν αρχικός πειραματικός σχεδιασμός. Χρησιμοποιώντας αυτά τα σημεία, μπορεί να προκύψει μια πρώτη προσέγγιση του φράγματος απόφασης. Βάσει αυτής της προσέγγισης, δημιουργούνται νέα σημεία που, αρχικά, πρέπει να βρίσκονται στην προσέγγιση του φράγματος απόφασης. Επιπλέον, αυτά τα νέα σημεία θα πρέπει μαζί με τα παλιά να αποτελούν ένα D-βέλτιστο σχεδιασμό. Στη συνέχεια,υπολογίζονται τα σημεία που βρίσκονται πιο μακρυά από το φράγμα απόφασης.

Στο δίκτυο των γραμμικών μοντέλων το κριτήριο παρουσίασης και εκτίμησης είναι το μέσο τετραγωνικό σφάλμα. Στην ταξινόμηση τα πράγματα είναι περισσότερο πολύπλοκα. Η παρουσίαση, συνήθως, εκτιμάται με την τάξη του σφάλματος. Για την

εκτίμηση των κανόνων ταξινόμησης υπάρχουν διάφορα κριτήρια, όπως για παράδειγμα το «ML-estimation».

Προκειμένου να επιβεβαιωθεί αν είναι οφέλιμη η χρήση D-βέλτιστων σχεδιασμών στην εκτέλεση των μεθόδων ταξινόμησης, εφαρμόζονται πέντε διαφορετικές μέθοδοι. Οι μέθοδοι αυτοί είναι οι παρακάτω: LDA, QDA, CART, γραμμική SVM και SVM με ακτινική βάση πυρήνα. Η εκτέλεση που βασίζεται σε D-βέλτιστους σχεδιασμούς αναμένεται να δουλέψει κυρίως για την LDA και την γραμμική SVM για τους ακόλουθους λόγους.

Οι βέλτιστοι σχεδιασμοί είναι εξαρτημένοι από το μοντέλο.Το μοντέλο κρησαρίσματος πάνω στο οποίοι βασίζονται οι D–βέλτιστοι σχεδιασμοί επανεμφανίζεται και για τις δύο μεθόδους, τη γραμμική SVM και την LDA, στην περίπτωση δύο κλάσεων με ίσες πρότερες πιθανότητες. Το τελευταίο ισχύει εφόσον ο κανόνας ταξινόμησης μπορεί να γραφτεί ως εξής:

$$\hat{y} = \text{sign}(\beta_0 + \beta_1' x) \tag{4.2}$$

Στην περίπτωση της γραμμικής SVM το κριτήριο εκτίμησης είναι το μέγεθος του περιθωρίου μεταξύ δύο κλάσεων. Στην περίπτωση της LDA για δύο κλάσεις με ίσες πρότερες πιθανότητες ο κανόνας ταξινόμησης δίνεται ως εξής:

$$\hat{y} = \text{sign}\,(-\tfrac{1}{2}\,(\mu_1 + \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2) + (\mu_1 - \mu_2)'\Sigma^{-1}x \tag{4.3}$$

Μπορεί, επίσης να γραφτεί όπως παραπάνω με $\beta_0 := -\tfrac{1}{2}\,(\mu_1 + \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2)$ και

$\beta_1 := \Sigma^{-1}(\mu_1 - \mu_2)$.

Η μέση τάξη και ο πίνακας συνδιακύμανσης υπολογίζονται από το ML- κριτήριο. Αλλά αν υποθέσουμε ότι οι δύο κλάσεις κωδικοποιούνται, για παράδειγμα με -1 και +1, τότε η εκτίμηση του συντελεστή διανύσματος β_1 από τα ελάχιστα τετράγωνα είναι ανάλογη με το $\Sigma^{-1}(\mu_1 - \mu_2)$.

Συνολικά, τα μοντέλα ταξινόμησης για την LDA και τη γραμμική SVM είναι πολύ παρόμοια στη μορφή τους. Επιπλέον, η LDA έχει ισχυρή σύνδεση με την εκτίμηση των ελαχίστων τετραγώνων και έτσι ταιριάζει καλύτερα το δίκτυο των γραμμικών μοντέλων .

4.5 Πειράματα Προσομοίωσης

Η μελέτη προσομοίωσης που περιγράφεται παρακάτω βασίζεται στις προσομοιώσεις του Pumplun (2005α). Χρησιμοποιούνται τα ίδια οκτώ σύνολα δεδομένων όπως οι Pumplünetal. (2005a) . Έξι σύνολα δεδομένων (ισορροπία, στήθος, διαβήτης, ίρις, συκώτι και κρασί)έχουν ληφθεί από την αποθήκη της μηχανής μάθησης UCI (Murphy and Aha, 1994). Το σύνολο δεδομένων ισορροπίας (balance) είναι τεχνητό

όπως και τα υπόλοιπα πέντε, καθώς τα εναπομείναντα σύνολα δεδομένων επιχείρησης και ιατρικής (business and medicine) αποτελούνται από πραγματικά στοιχεία. Όλα τα σύνολα δεδομένα αποτελούν προβλήματα δύο κλάσεων. Σε περίπτωση που ο αριθμός των κλάσεων ήταν μεγαλύτερος, είτε οι παρατηρήσεις κάποιων κλάσεων θα παραλείπονταν, είτε πολύ παρόμοιες κλάσεις (όπως το στο σύνολο δεδομένων της ίριδας τα είδη "virginica" και "versicolor") θα συνδιάζονταν. Ο Πίνακας 4.1 παρακάτω δείχνει τους αριθμούς των παρατηρήσεων n και τις διαστάσεις p των συνόλων δεδομένων.

Πίνακας 4.1 : Αριθμοί παρατηρήσεων και Μεταβλητές

Σύνολο δεδομένων	**n**	**p**
Ισορροπία	576	4
Στήθος	683	9
Διαβήτης	768	8
Ίρις	150	4
Συκώτι	345	6
Κρασί	178	13
Επιχείρηση	157	13
Ιατρική	6610	18

Σε καθένα από τα δέκα σύνολα δεδομένων προς εκτέλεση, ερευνάται η ύπαρξη D-βέλτιστων σχεδιασμών διαφορετικού μεγέθους. Προκειμένου να συγκριθούν τυχαία δείγματα ίδιου μεγέθους, όλα τα σύνολα δεδομένων που εκτελούνται, θεωρούνται βάση για την επιλογή μεταβλητών ή και για την εκτίμηση των μοντέλων ταξινόμησης.

Για την επιλογή μεταβλητών χρησιμοποιήθηκαν σχεδιασμοί μεγέθους $k = p + 1$. Ωστόσο, σημαίνει πολύ λίγες παρατηρήσεις προς εκτέλεση για κάποιες μεθόδους ταξινόμησης, όπως για παράδειγμα για την QDA. Συνεπώς, γίνεται προσπάθεια χρήσης πέντε διαφορετικών αριθμών παρατηρήσεων.

Ονομαστικά έχουμε : $k \in \{p+1, 2(p+1), 0.1n, 0.25n, 0.5n\}$, με το k στρογγυλοποιημένο στους ακέραιους αριθμούς. Η χρήση όλων των παρατηρήσεων προς εκτέλεση για την επιλογή μεταβλητών ή και την εκτίμηση, αντιστοιχεί στη σχέση $k \approx 0.9n$. Ο πίνακας 4.2 παρουσιάζει τον αριθμό των παρατηρήσεων στους σχεδιασμούς για τα οκτώ σύνολα δεδομένων.

Πίνακας 4.2: Αριθμός k των παρατηρήσεων ανά σχεδιασμό (στογγυλοποιημένος σε ακέραιους) .

Σύνολο δεδομένων	**p+1**	**2(p+1)**	**0.1n**	**0.25n**	**0.5n**	**0.9n**
Ισορροπία	5	10	58	144	288	518
Στήθος	10	20	68	171	342	615
Διαβήτης	9	18	77	192	384	691
Ίρις	5	10	15	38	75	135
Συκώτι	7	14	34	86	172	311
Κρασί	14	28	18	44	89	160
Επιχείρηση	14	28	16	39	78	141
Ιατρική	19	38	661	1652	3305	5949

Προκειμένου να γίνει έρευνα για D–βέλτιστους σχεδιασμούς, χρησιμοποιείται ένας γενετικός αλγόριθμος. Παρόμοιες παράμετροι, όπως στην παρούσα έρευνα, επιλέγονται μόνο όταν το μέγεθος του πληθυσμού αυξάνεται από το 10 στο 100:

- Αριθμός επαναλήψεων στον εξωτερικό βρόχο:100,
- Αριθμός επαναλήψεων στον εσωτερικό βρόχο:10,
- Μέγεθος πληθυσμού:100,
- Ποσοστό σχεδιασμών που αντικαθίστανται μέσω διασταύρωσης από τους καλύτερους σχεδιασμούς:0.4,
- Πιθανότητα αλλαγής: 0.01.

Όπως ήδη έχει περιγραφεί, για την επιλογή μεταβλητών χρησιμοποιούνται δύο διαφορετικά κριτήρια:το κριτήριο της συσχέτισης (correlation) και τα δέντρα (tree). Δοκιμάζονται επίσης, τρεις διαφορετικοί αριθμοί u μεταβλητών. Ονομαστικά έχουμε: $v \in \{0.25p, 0.5p, 0.75p\}$, με το u να στρογγυλοποιείται στους ακέραιους αριθμούς. Προκειμένου να γίνει σύγκριση,υπολογίζουμε, επίσης, την τάξη σφάλματος στη βάση όλων των μεταβλητών. Ο Πίνακας 4.3 παρουσιάζει τους αριθμούς όλων των επιλεγμένων μεταβλητών για τα οκτώ σύνολα δεδομένων.

Εξαρτώμενοι από τον επιλεγμένο σχεδιασμό για την εκτέλεση των δεδομένων, λαμβάνουμε διάφορα σενάρια επιλογής μεταβλητών

Πίνακας 4.3: Αριθμός των v επιλεγμένων μεταβλητών (στρογγυλοποιημένος στους ακέραιους)

Σύνολο δεδομένων	0.25p	0.5p	0.75p	p
Ισορροπία	1	2	3	4
Στήθος	2	4	7	9
Διαβήτης	2	4	6	8
Ίρις	1	2	3	4
Συκώτι	2	3	4	6
Κρασί	3	6	10	13
Επιχείρηση	3	6	10	13
Ιατρική	4	9	14	18

- «vs doptimal»:επιλογή μεταβλητών με χρήση του κριτηρίου συσχέτισης (correlation) ή με δέντρο (tree), βασισμένη σε έναν D–βέλτιστο σχεδιασμό μεγέθους k,
- «vs doptimalit»: επιλογή μεταβλητών βασισμένη σε ένα σύνολο 100 περίπου D–βέλτιστων σχεδιασμών μεγέθους k.(Ο αριθμός των σχεδιασμών στο σύνολο, εξαρτάται από τον αριθμό των επαναλήψεων στον εξωτερικό βρόχο του γενετικού αλγορίθμου.)

Επιπλέον, για σκοπούς σύγκρισης, θεωρούνται άλλα δύο σενάρια επιλογής μεταβλητών τα οποία ονομάζονται:

- «no vs»: καμία επιλογή μεταβλητών, που σημαίνει v = p,
- «vs standard» : επιλογή μεταβλητών σε όλο το σύνολο δεδομένων προς εκτέλεση.

Στο σημείο αυτό προσθέτουμε:

- «vs random var» : μεταβλητές επιλέγονται τυχαία (προκειμένου να εκτιμηθεί η ποιότητα των κριτηρίων επιλογής μεταβλητών),
- «vs random obs» : επιλογή μεταβλητών βασισμένη σε τυχαίο δείγμα k παρατηρήσεων προς εκτέλεση (αντίστοιχη της vs doptimal),
- «vs random obs it» : επιλογή μεταβλητών βασισμένη σε ένα σύνολο 100 τυχαίων δειγμάτων μεγέθους k (αντίστοιχη της vs doptimal it.)

Όλα τα παραπάνω σενάρια (εκτός από τις no vs και vs random var), χρησιμοποιούνται σε συνδιασμό με τα δύο κριτήρια της συσχέτισης και των δέντρων.

Εφαρμόζονται πέντε διαφορετικές μέθοδοι ταξινόμησης. Ονομαστικά έχουμε LDA, QDA, CART, γραμμική SVM (SVMDOT) και SVM με ακτινική βάση πυρήνα (SVMRBF).

Στην παρούσα έρευνα, η επιλογή μεταβλητών γίνεται βάσει επιλεγμένων σχεδιασμών, αλλά τα μοντέλα ταξινόμησης εκτιμώνται στη βάση των παρατηρήσεων προς εκτέλεση. Θα πρέπει να σημειωθεί πως το γεγονός αυτό, δεν έχει ιδιαίτερη σημασία στην παρούσα φάση καθώς δεν είναι δυνατό να προσδιοριστεί η ετικέτα (label) κάθε κλάσης όλων των μεταβλητών. Ωστόσο, βοηθά ώστε να εκτιμηθεί αν οι D – βέλτιστοι σχεδιασμοί είναι οφέλιμοι στην επιλογή μεταβλητών. Στην παρούσα εργασία, ερευνάται, επίσης, η καταλληλότητα των D – βέλτιστων σχεδιασμών σαν βάση για εκτέλεση των μεθόδων ταξινόμησης. Με αυτό τον τρόπο, τα μοντέλα ταξινόμησης εκτιμώνται πάνω στους θεμελιώδεις σχεδιασμούς αντί για το σύνολο των χρησιμοποιούμενων παρατηρήσεων. Σύμφωνα, λοιπόν, με τα σενάρια επιλογής μεταβλητών που περιγράφηκαν παραπάνω, έχουμε τις ακόλουθες παραλλαγές:

- «es tall»: τα μοντέλα ταξινόμησης εκτιμώνται στη βάση όλου του συνόλου παρατηρήσεων προς εκτέλεση,
- «est doptimal»: τα μοντέλα ταξινόμησης εκτιμώνται στη βάση του D–βέλτιστου σχεδιασμού,
- «est random obs» : τα μοντέλα ταξινόμησης εκτιμώνται στη βάση ενός τυχαίου δείγματος,
- «est doptimal it»: τα μοντέλα ταξινόμησης εκτιμώνται στη βάση ενός συνόλου 100 D–βέλτιστων σχεδιασμών,
- «est random obs it»: τα μοντέλα ταξινόμησης εκτιμώνται στη βάση ενός συνόλου 100 τυχαίων δειγμάτων.

Χρησιμοποιούμε αυτά τα σενάρια εκτίμησης χωρίς επιλογή μεταβλητών και σε συνδιασμό με τα αντίστοιχα σενάρια επιλογής μεταβλητών, για παράδειγμα η vs doptimal συνδυασμένη με την est doptimal, εν συντομίαν sest doptimal).

Η διαδικασία που περιγράφηκε παραπάνω μπορεί να συνοψιστεί ως ακολούθως:

- 1.Επέλεξε ένα σχεδιασμό μεγέθους k ή ένα σύνολο 100 σχεδιασμών μεγέθους k (με $k \in \{ p+1, 2(p+1), 0.1n, 0.25n, 0.5n \}$ στο i– οστό σύνολο δεδομένων προς εκτέλεση (είτε D–βέλτιστοι σχεδιασμοί, είτε τυχαία δείγματα) ή κράτησε όλες τις παρατηρήσεις προς εκτέλεση ($k \approx 0.9n$).
- 2.Χρησιμοποίησε τα κριτήρια συσχέτισης και δέντρου για να επιλέξεις v μεταβλητές (με $v \in \{0.25p, 0.5p, 0.75p\}$), βάσει των θεμελιωδών σχεδιασμών ή κράτα όλες τις μεταβλητές ($v = p$).
- 3.Εκτίμησε τα μοντέλα ταξινόμησης (LDA, QDA, CART, SVMDOT, SVMRBF) στη βάση των θεμελιωδών σχεδιασμών ή στη βάση όλου του συνόλου δεδομένων προς εκτέλεση.
- 4.Πρόβλεψε την ετικέτα (label) της κλάσης στο i–οστό δοκιμαστικό σύνολο δεδομένων.

Το R λογισμικό (R Development Core Team, 2008) χρησιμοποιείται για τη μελέτη προσομοίωσης. Επίσης, χρησιμοποιούνται τα ακόλουθα πακέτα:

- MASS (Venables and Ripley, 2002) γιαLDA καιQDA
- Rpart (Therneau and Atkinson, 2009) γιαCART
- E1071 (Dimitriadou et al., 2009) γιαSVMDOT καιSVMRBF

4.6 Αποτελέσματα

Αυτή η ενότητα αποτελείται από πέντε μέρη. Στην πρώτη υποενότητα, μελετάται η επιλογή μεταβλητών βασισμένη σε όλο το σύνολο των παρατηρήσεων προς εκτέλεση. Στη συνέχεια, μελετάται ο θεμελιώδης D – βέλτιστος σχεδιασμός καθώς και τυχαίοι σχεδιασμοί σε σχέση με την D–αποδοτικότητα. Μελετώνται, επίσης, συσχετίσεις μεταξύ των επεξηγηματικών μεταβλητών στη δεύτερη υποενότητα. Στην τρίτη υποενότητα εκτιμάται η καταλληλότητα των D – βέλτιστων σχεδιασμών σαν βάση για την επιλογή μεταβλητών. Στην τέταρτη, μελετώνται οι D – βέλτιστοι σχεδιασμοί σαν βάση για την εκτέλεση των μεθόδων ταξινόμησης. Στην τελευταία υποενότητα ερευνάται αν η χρήση των D – βέλτιστων σχεδιασμών είναι οφέλιμη για τον συνδιασμό επιλογής μεταβλητών και εκτέλεσης δεδομένων.

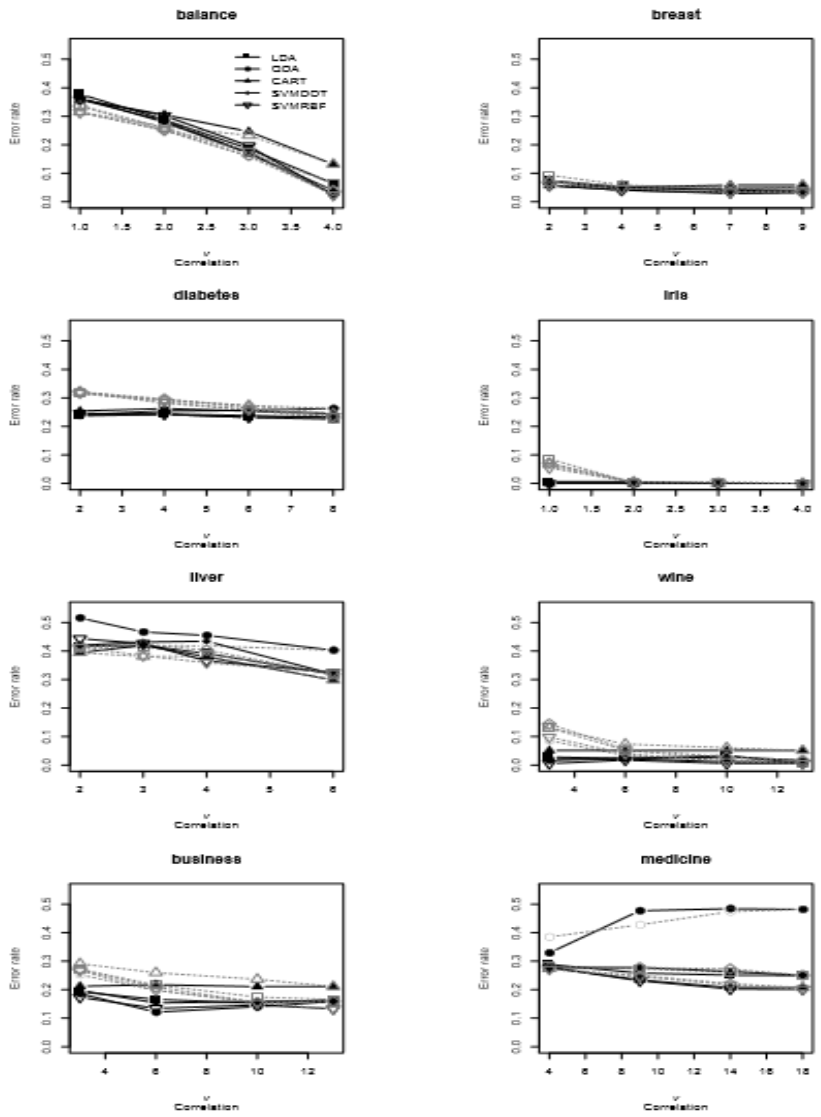
4.7 Επιλογή μεταβλητών και Εκτίμηση βάσει όλων των παρατηρήσεων

Σαν πρώτο βήμα, ερευνάται αν η επιλογή μεταβλητών είναι οφέλιμη για μερικά σύνολα δεδομένων. Δηλαδή, αν η ταξη του σφάλματος μειώνεται ή παραμένει σταθερή αν ο αριθμός των επεξηγηματικών μεταβλητών μειώνεται. Για το σκοπό αυτό, συγκρίνονται οι τάξεις σφάλματος που προκύπτουν από την vs standard σε συνδιασμό με τα κριτήρια συσχέτισης, δέντρων και οι τάξεις σφάλματος βάσει της σύγκρισης όλων των μεταβλητών (no vs). Θα πρέπει να σημειωθεί ότι οι διαστάσεις πολλών συνόλων δεδομένωνείναι μάλλον μικρές. Συνεπώς, δεν είναι δυνατόν να αναμένουμε ότι η επιλογή μεταβλητών θα είναι οφέλιμη για όλα τα σύνολα δεδομένων.

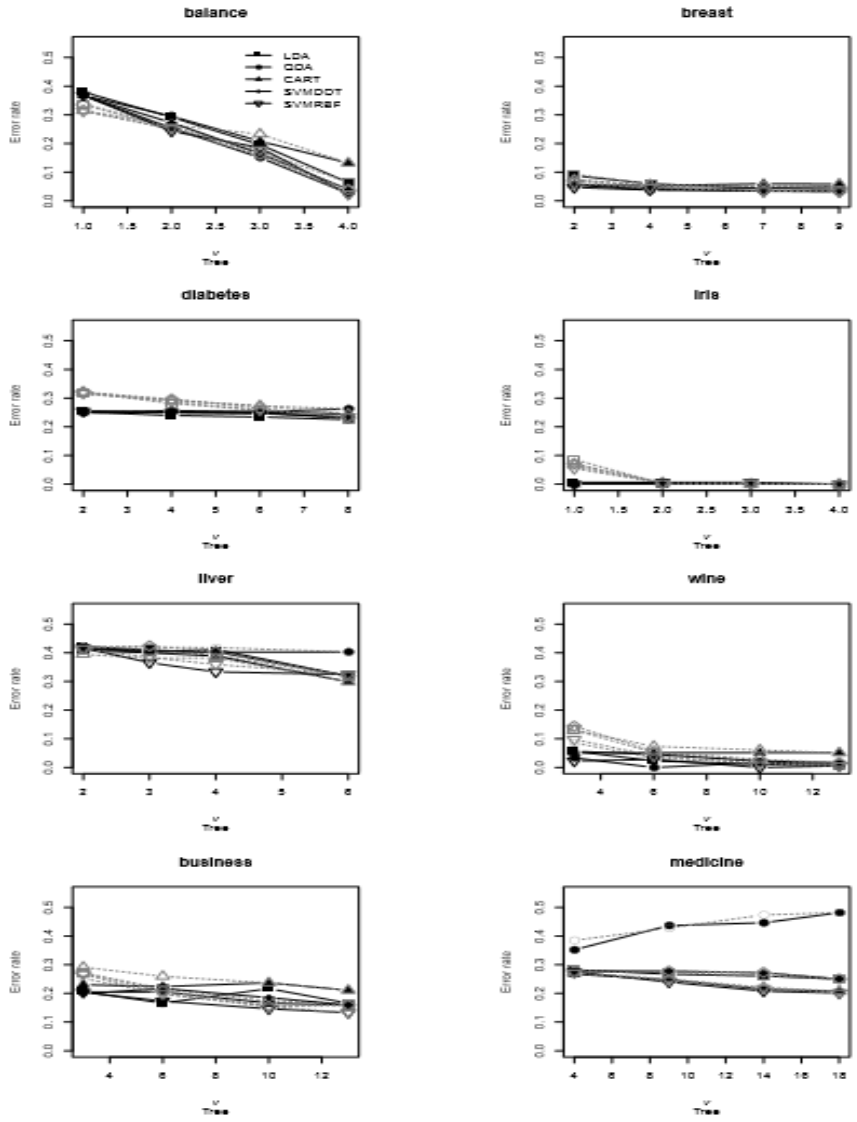
Σε δεύτερο βήμα, θεωρείται η ποιότητα των δύο κριτηρίων επιλογής μεταβλητών. Όπως έχει περιγραφεί παραπάνω, το κριτήριο των δέντρων θεωρείται ασταθές. Από την άλλη μεριά, το κριτήριο της συσχέτισης αφορά μόνο γραμμικές εξαρτήσεις και έτσι ίσως χάνονται σημαντικές μεταβλητές. Προκειμένου να ελεγθεί το τελευταίο, συγκρίνονται οι τάξεις σφάλματος που προκύπτουν από την vs standardβασισμένης στα δύο κριτήρια και από την vs random var όπου οι μεταβλητές επιλέγονται τυχαία.

Τα Σχήματα 4.1 και 4.2 παρουσιάζουν τις τάξεις των επικυρωμένων με διασταύρωση σφαλμάτων για τα οκτώ σύνολα δεδομένων και διαφορετικούς αριθμούς επιλεγμένων μεταβλητών $v \in \{0.25p, 0.5p, 0.75p\}$, τις πέντε μεθόδους ταξινόμησης και δύο κριτήρια επιλογής μεταβλητών. Επιπροσθέτως, η τάξη σφάλματος που προκύπτει από την επιλογή μεταβλητών χαράσσεται με γκρι χρώμα.

Το κρασί και κυρίως η ίρις αποτελούν πολύ απλά προβλήματα ταξινόμησης, καθώς η τάξη σφάλματοςείναι σχεδόν μηδενική. Τα αποτελέσματα που λαμβάνονται πάνω στα σύνολα δεδομένων της ίριδας, δεν θα χρησιμοποιηθούν στα παρακάτω διότι το πρόβλημα ταξινόμησης είναι πάρα πολύ απλό για να ανιχνευθούν μικρές διαφορές ανάμεσα στις μεθόδους ταξινόμησης και επιλογής μεταβλητών. Συχνά, η CART και ηQDA δίνουν ελαφρώς χειρότερη παρουσίαση σε σχέση με τις άλλες μεθόδους ταξινόμησης, ιδιαίτερα η QDA στο σύνολο δεδομένων της ιατρικής. Η επιλογή μεταβλητών μέσω των κριτηρίων της συσχέτισης και των δέντρων, είναι οφέλιμη για τα περισσότερα σύνολα δεδομένων. Μόνο για το σύνολο δεδομένων για το συκώτι και κυρίως για το σύνολο δεδομένων της ισορροπίας, η τάξη σφάλματος μεγαλώνει σημαντικά όταν ο αριθμός των μεταβλητών μειώνεται. Τα κριτήρια της συσχέτισης και των δέντρων αποδίδουν παρόμοια αποτελέσματα.



Σχήμα 4.1: Επικυρωμένα με διασταύρωση σφάλματα που προκύπτουν από την vs standard με το κριτήριο της συσχέτισης (μαύρο), την no vs(μαύρο), και την vs random var (γκρι) για διαφορετικούς αριθμούς μεταβλητών v∈ {0.25p, 0.5p, 0.75p, p}.



Σχήμα 4.2:Επικυρωμένα με διασταύρωση σφάλματα που προκύπτουν από την vs standard με το κριτήριο των δέντρων (μαύρο), την no vs (μαύρο), και την vs random var (γκρι)για διαφορετικούς αριθμούς μεταβλητών v ∈ {0.25p, 0.5p, 0.75p, p}.

Η χρήση των κριτηρίων της συσχέτισης και των δέντρων οδηγεί σε χαμηλότερες τάξεις σφάλματος σε σχέση με την επιλογή τυχαίων μεταβλητών για τα σύνολα των δεδομένων του διαβήτη, της ίριδας, του κρασιού και της επιχείρησης. Πρωτίστως, για μικρούς αριθμούς επιλεγμένων μεταβλητών, τα κριτήρια αποδεικνύονται αποδοτικά. Για την ιατρική και το στήθος λαμβάνονται παρόμοια αποτελέσματα και από τις δύο μεθόδους. Η vs standard έχει μάλλον, χειρότερη παρουσίαση σε σχέση με την τυχαία επιλογή μεταβλητών για τα σύνολα δεδομένων της ισορροπίας και συκωτιού. Θα πρέπει να σημειωθεί ότι για αυτά τα δύο σύνολα δεδομένων η επιλογή μεταβλητών δεν είναι γενικά ιδιαίτερα οφέλιμη.

Προκειμένουνα λάβουμε ένα ενοποιημένο αποτέλεσμα για όλα τα σύνολα δεδομένων, υπολογίζουμε τη μέση σχετική απόκλιση:

$$\text{MRD (μέθοδος 1, μέθοδος 2)} = \frac{1}{m}\sum_{i=1}^{m}\frac{e_1^i - e_2^i}{e_2^i} \quad (4.3)$$

Όπου e_1^i και e_2^i δηλώνουν τη τάξη σφάλματος που λήφθηκε μέσω των μεθόδων 1 και 2 στο i–οστό σύνολο δεδομένων. Εφόσον η διαφορά $e_1^i - e_2^i$διαιρείται με το e_2^i , μια αύξηση μιας χαμηλής τάξης σφάλματος θεωρείται χειρότερη από μια αύξηση μιας υψηλής τάξης σφάλματος. Λαμβάνεται, επίσης, υπόψη ότι μια μείωση χαμηλής τάξης εσφαλμένης ταξινόμησης είναι πιο απίθανο να προκύψει σε σχέση με αντίστοιχη μείωση μιας υψηλής τάξης. Επιπλέον, υπολογίζονται οι σχετικές συχνότητες για τις οποίες η δεύτερη μέθοδος είναι καλύτερη από την πρώτη ως:

$$\frac{1}{m}\sum_{i=1}^{m} I(e_1^i > e_2^i) \quad (4.4)$$

όπου το I δηλώνει τη συνάρτηση – δείκτη.

Ο Πίνακας 4.4 παρουσιάζει τη μέση σχετική απόκλιση MRD (vs standard, no vs) για τα κριτήρια της συσχέτισης και των δέντρων και τις πέντε μεθόδους ταξινόμησης. Τα αποτελέσματα των συνόλων δεδομένων της ίριδας δεν χρησιμοποιούνται.

Η επιλογή μεταβλητών οδηγεί σε αύξηση της τάξης σφάλματος κατά μέσο όρο. Η μέση σχετική απόκλιση MRD (vs standard, no vs) είναι μεγαλύτερη για μικρό αριθμό v επιλεγμένων μεταβλητών. Ο Πίνακας 4.4 βοηθά ιδιαίτερα στην ανίχνευση διαφορών μεταξύ των δύο κριτηρίων επιλογής μεταβλητών. Αν ο αριθμός των μεταβλητών είναι μικρός, τότε το κριτήριο της συσχέτισης φαίνεται να δουλεύει καλύτερα σε σχέση με τα δέντρα. Η τάξη του σφάλματος που προκύπτει από την CART αλλάζει πολύ λίγο όταν ο αριθμός των μεταβλητών μειώνεται. Ο κύριος λόγος για τον οποίο συμβαίνει αυτό είναι ότι συνήθως δεν χρησιμοποιούνται όλες οι μεταβλητές σε ένα συνδιασμόCART–δέντροκαι έτσι η CART επιλέγει μεταβλητές σιωπηλά σε κάθε περίπτωση.

Πίνακας 4.4: Μέση σχετική απόκλιση MRD (vs standard, no vs) της τάξης σφάλματος που προκύπτει από την vs standard και την no vs.

Συσχέτιση	**v**	**LDA**	**QDA**	**CART**	**SVMDOT**	**SVMRBF**	**Σύνολο**
	0.25p	1.50	1.98	0.34	1.42	2.22	1.49
	0.5p	1.05	1.55	0.26	1.01	2.04	1.18
	0.75p	1.02	0.79	0.17	0.61	1.03	0.72
Δέντρο	v	LDA	QDA	CART	SVMDOT	SVMRBF	Σύνολο
	0.25p	2.29	2.18	0.36	1.73	2.69	1.85
	0.5p	1.08	1.10	0.25	1.24	1.98	1.13
	0.75p	0.54	0.74	0.16	0.59	0.83	0.57

Ο Πίνακας 4.5 παρουσιάζει τις σχετικές συχνότητες για τις οποίες η vs standard αποδίδει την ίδια ή χαμηλότερη τάξη σφάλματος σε σχέση με την no vsγια διαφορετικούς αριθμούς επιλεγμένων μεταβλητών και για πέντε μεθόδους ταξινόμησης (αποκλειστικά του συνόλου δεδομένων της ίριδας).

Πίνακας 4.5: Σχετική συχνότητα για την οποία η vs standard δίνει την ίδια ή χαμηλότερη τάξη σφάλματος σε σχέση με την no vs.

Συσχέτιση	**v**	**LDA**	**QDA**	**CART**	**SVMDOT**	**SVMRBF**	**Σύνολο**
	0.25p	0.00	0.29	0.29	0.00	0.14	0.14
	0.5p	0.00	0.57	0.29	0.14	0.00	0.20
	0.75p	0.29	0.57	0.43	0.29	0.43	0.40
Δέντρο	v	LDA	QDA	CART	SVMDOT	SVMRBF	Σύνολο
	0.25p	0.00	0.29	0.29	0.00	0.00	0.11
	0.5p	0.00	0.57	0.29	0.00	0.00	0.17
	0.75p	0.00	0.57	0.14	0.00	0.14	0.17

Χρησιμοποιώντας το κριτήριο συσχέτισης, κατά μέσο όρο, μειώνεται η τάξη σφάλματος ή παραμένει σταθερή συχνότερα με το κριτήριο των δέντρων. Κυρίως η QDA επωφελείται από την επιλογή μεταβλητών.

ΟΠίνακας 4.6 παρουσιάζει τη μέση σχετική απόκλιση MRD (vs random var, vs standard) για σταθερή και τυχαία επιλογή μεταβλητών (αποκλειστικά για το σύνολο δεδομένων της ίριδας). Κατά μέσο όρο, η τυχαία επιλογή μεταβλητών αυξάνει την τάξη σφάλματος που προκύπτει από την vs stasndard, ιδιαίτερα αν ο αριθμός των επιλεγμένων μεταβλητών είναι μικρός.

Πίνακας 4.6: Μέση σχετική απόκλιση MRD (vs random var, vs standard) της τάξης σφάλματος που προκύπτει από την vs standard και την vs random var.

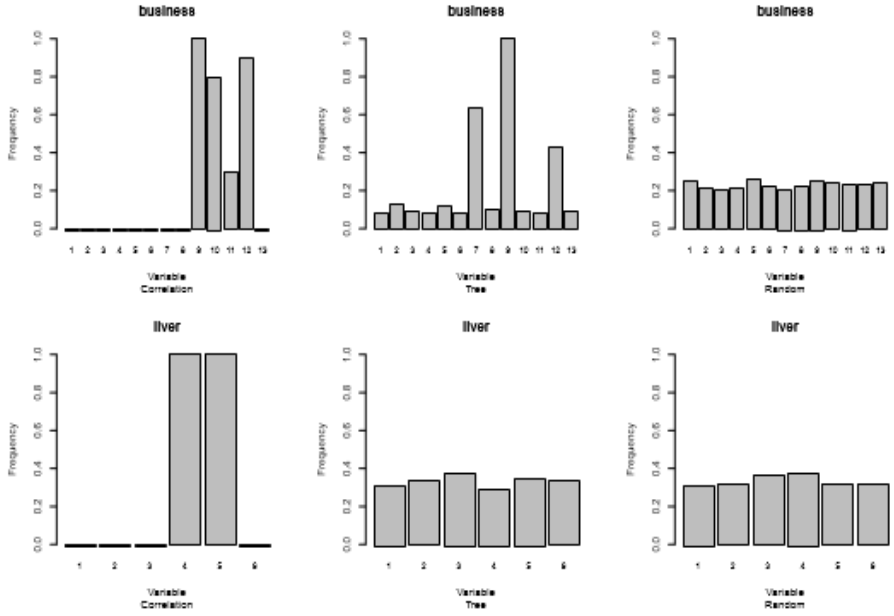
Συσχέτιση	**v**	**LDA**	**QDA**	**CART**	**SVMDOT**	**SVMRBF**	**Σύνολο**
	0.25p	0.64	0.66	0.33	0.88	2.47	1.00
	0.5p	0.26	0.15	0.10	0.28	0.28	0.21
	0.75p	-0.01	0.08	0.04	0.03	0.21	0.07
Δέντρο	v	LDA	QDA	CART	SVMDOT	SVMRBF	Σύνολο
	0.25p	0.26	0.34	0.32	0.33	0.57	0.36
	0.5p	0.24	0.07	0.09	0.04	0.15	0.12
	0.75p	0.08	0.04	0.04	0.02	0.03	0.04

Στον Πίνακα 4.7 παρουσιάζονται οι σχετικές συχνότητες για τις οποίες η vs standard δίνει μικρότερη τάξη σφάλματος σε σχέση με την vs random var. Όπως παρατηρούμε, αυτό συμβαίνει για την πλειονότητα των συνόλων δεδομένων και για τα δύο κριτήρια και σχεδόν για όλους τους συνδιασμούς τωνν με τις μεθόδους ταξινόμησης.

Πίνακας 4.7: Σχετική συχνότητα για την οποία τηνs standard δίνει χαμηλότερη τάξη σφάλματος σε σχέση με την no vs.

Συσχέτιση	**v**	**LDA**	**QDA**	**CART**	**SVMDOT**	**SVMRBF**	**Σύνολο**
	0.25p	0.57	0.71	0.71	0.57	0.57	0.63
	0.5p	0.71	0.57	0.71	0.57	0.71	0.66
	0.75p	0.57	0.57	0.71	0.43	0.71	0.60
Δέντρο	v	LDA	QDA	CART	SVMDOT	SVMRBF	Σύνολο
	0.25p	0.57	0.86	0.71	0.71	0.57	0.69
	0.5p	0.71	0.57	0.71	0.43	1.00	0.69
	0.75p	0.43	0.71	0.57	0.57	0.86	0.63

Τέλος, σε αυτή την υποενότητα ερευνώνται οι συχνότητες των μεταβλητών που που προκύπτουν τόσο από τη χρήση των κριτηρίων συσχέτισης και δέντρων, όσο και αυτών που προκύπτουν από την τυχαία επιλογή μεταβλητών. Όπως έχει ήδη περιγραφεί σε προηγούμενη ενότητα, το κριτήριο της συσχέτισης παρέχει ένα διατεταγμένο σύνολο όλων των μεταβλητών που δεν περιλαμβάνονται στο διαθέσιμο δέντρο. Για αυτόν το λόγο, προκειμένου να φτάσουμε στον επιθυμητό αριθμό v των μεταβλητών, οι υπολειπόμενες μεταβλητές επιλέγονται τυχαία αν δεν υπάρχουν αρκετές μεταβλητές στο δέντρο.



Σχήμα 4.3: Επιλογή συχνοτήτων για τις μεταβλητές, με τη χρήση των κριτηρίων συσχέτισης, δέντρου καθώς και για τυχαία επιλογή μεταβλητών στα σύνολα δεδομένων επιχείρισης και συκωτιού για v = 0.25p.

Το Σχήμα 4.3 παρουσιάζει τη σχετική επιλογή συχνοτήτων για τις μεταβλητές στα σύνολα δεδομένων επιχείρησης και συκωτιού. Ο επιθυμητός αριθμός μεταβλητών είναι v = 0.25p. Αυτό καταλήγει σε 2 μεταβλητές για το σύνολο δεδομένων του συκωτιού και σε 3 μεταβλητές για το σύνολο δεδομένων στην περίπτωση της επιχείρησης. Μέσω των δύο κριτηρίων επιλέγονται διαφορετικά υποσύνολα μεταβλητών. Όπως αναμενόταν, η επιλογή μεταβλητών μέσω του κριτηρίου του δέντρου είναι περισσότερο ασταθές σε σχέση με το κριτήριο της συσχέτισης, ιδιαίτερα στην περίπτωση του συνόλου δεδομένων συκωτιού. Χρησιμοποιώντας σαν παράδειγμα το σύνολο δεδομένων για την επιχείρηση, μπορούμε να παρατηρήσουμε ότι ο συνδιασμόςCART- δέντρα δεν περιλαμβάνει πάντα 3 μεταβλητές. Συνεπώς, κάποιες μεταβλητές επιλέχθηκαν τυχαία.

Γενικά, θα μπορούσε να αναφερθεί ότι η επιλογή μεταβλητών καταλήγει σε μια μικρή μόνο αύξηση της τάξης σφάλματος (εκτός από τα σύνολα δεδομένων για την ισορροπία και το συκώτι). Το κριτήριο συσχέτισης φαίνεται να λειτουργεί καλύτερα σε σχέση με το κριτήριο του δέντρου, ιδιαίτερα για μικρό αριθμό μεταβλητών.

4.8 Σχεδιασμοί

Προκειμένου να εκτιμηθεί η σημαντικότητα κάθε μεταβλητής ξεχωριστά για την ταξινόμηση, ο σκοπός είναι να βρεθούν D–βέλτιστοι σχεδιασμοί οι οποίοι είναι σχεδόν ορθογώνιοι και έτσι οδηγούν σε ασυσχέτιστους προγνώστες. Για αυτόν το λόγο, αρχικά, ερευνώνται οι D -αποδοτικότητες και οι συσχετίσεις μεταξύ των επεξηγηματικών μεταβλητών στα σύνολα δεδομένων. Το τελευταίο βοηθά στο να εκτιμηθεί κατά πόσο είναι πιθανή μια βελτίωση μέσων ενός D–βέλτιστου σχεδιασμού. Δεύτερον, ερευνώνται οι D – αποδοτικότητες και οι συσχετίσεις που προκύπτουν από D–βέλτιστο σχεδιασμό και τυχαία επιλεγμένους σχεδιασμούς.

D–αποδοτικότητα και Συσχετίσεις σε σύνολα δεδομένων

Τα οκτώ σύνολα δεδομένων μπορούν να θεωρηθούν σαν σχεδιασμοί μεγέθους n, αντίστοιχα. Όπως έχει περιγραφεί σε προηγούμενη υποενότητα, τα σύνολα δεδομένων ήταν τυποποιημένα. Συνεπώς, η μέγιστη D – τιμή, αν όλα τα ζεύγη των επεξηγηματικών μεταβλητών ήταν ασυσχέτιστα, είναι

$$\mathrm{D}\ (\boldsymbol{X}^{*}_{opt}) = n^{p} \tag{4.5}$$

Ο Πίνακας 4.8 παρουσιάζει τις D – αποδοτικότητες

$$D_{eff}(\boldsymbol{X}^{*}) = \frac{D(\boldsymbol{X}^{*})^{1/p}}{n} \tag{4.6}$$

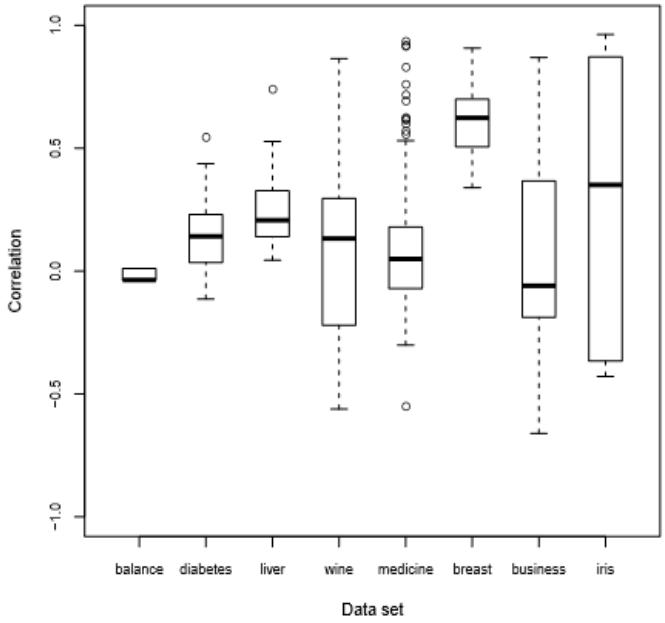
για τα ξεχωριστά σύνολα δεδομένων. Οι επεξηγηματικές μεταβλητές στο σύνολο δεδομένων της ισορροπίας, το οποίο είναι ένα τεχνητό σύνολο δεδομένων, είναι ήδη σχεδόν ορθογώνιες μεταξύ τους και συνεπώς ασυσχέτιστες. Αντίθετα, οι μεταβλητές στο σύνολο δεδομένων για την ίριδα δεν είναι.

Πίνακας 4.8: Οι D – αποδοτικότητες των οκτώ συνόλων που μελετώνται.

Σύνολο δεδομένων X^*	$D_{eff}(X^*)$	$\bar{r}$	s_r
Ισορροπία	0.999	-0.021	0.024
Στήθος	0.458	0.602	0.132
Διαβήτης	0.856	0.147	0.160
Ίρις	0.300	0.290	0.661
Συκώτι	0.785	0.265	0.194
Κρασί	0.555	0.085	0.351
Επιχείρηση	0.438	0.066	0.363
Ιατρική	0.507	0.096	0.255

Ο Πίνακας 4.8 παρουσιάζει τις μέσες τιμές $\bar{r}$ και τις τυπικές αποκλίσεις s_r των εμπειρικών συντελεστών συσχέτισης $r\boldsymbol{x}_{.i}\boldsymbol{x}_{.j}$,ανάμεσα σε όλα τα ζεύγη των επεξηγηματικών μεταβλητών $\boldsymbol{X}_{.i}$και $\boldsymbol{X}_{.j}$, i, j = 1,…, p, i≠j. Εκτός από το σύνολο δεδομένων της ισορροπίας, τα σύνολα δεδομένων της επιχείρησης, του κρασιού και της ιατρικής έχουν, επίσης, μέση συσχέτιση κοντά στο μηδέν. Εξαιτίας, όμως, των υψηλών τυπικών αποκλίσεων οι D – αποδοτικότητες για αυτά τα σύνολα δεδομένων είναι μάλλον χαμηλές.

Το Σχήμα 4.4 παρουσιάζει παράλληλα γραφήματα σε σχήμα κουτιού(boxplots) των συσχετίσεων μεταξύ όλων των ζευγών των επεξηγηματικών μεταβλητών στα οκτώ σύνολα δεδομένων, ταξινομημένα κατά φθίνουσα D–αποδοτικότητα.

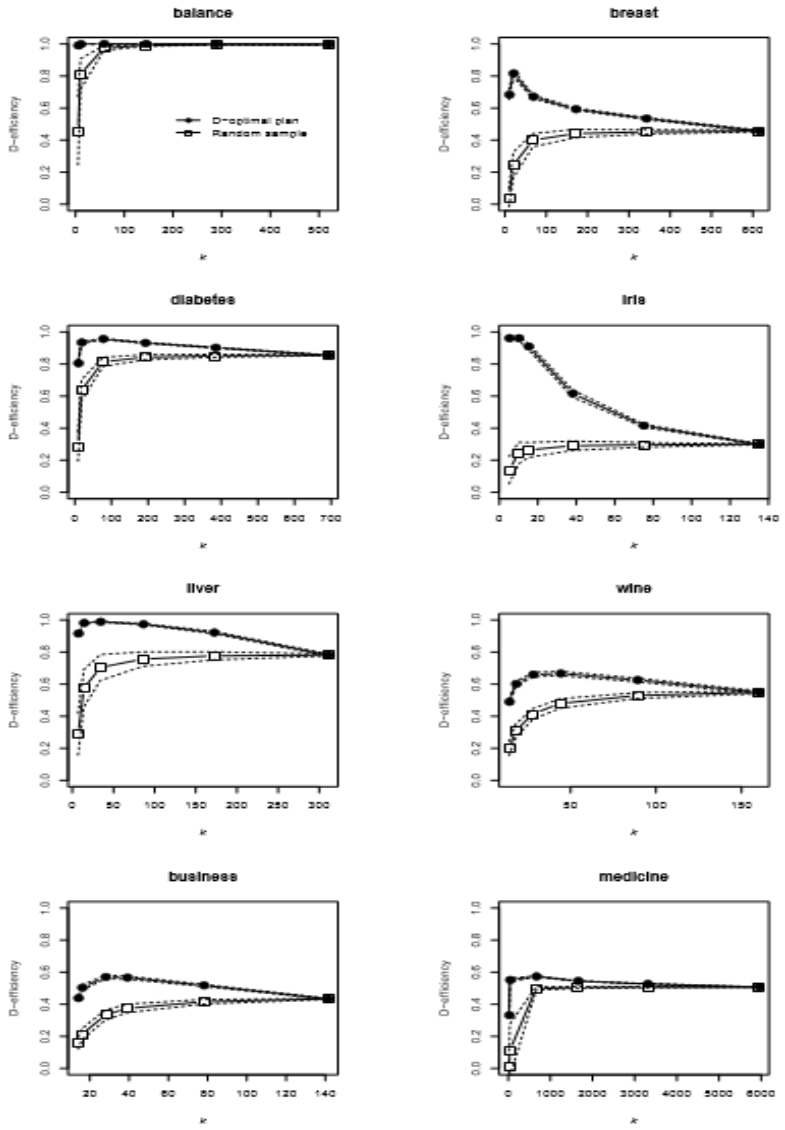


Σχήμα 4.4: Παράλληλα γραφήματα των συσχετίσεων μεταξύ όλων των ζευγών των επεξηγηματικών μεταβλητών, ταξινομημένα σύμφωνα με την D– αποδοτικότητα.

D–αποδοτικότητα και Συσχέτιση των Σχεδιασμών που προκύπτουν

Σε αυτή την παράγραφο, πρώτον,μελετώνται οι D–αποδοτικότητες του D- βέλτιστου σχεδιασμού που προκύπτουν και των τυχαία επιλεγμένων σχεδιασμών, όπως και οι συσχετίσεις μεταξύ των επεξηγηματικών μεταβλητών.

Το γράφημα 4.5 παρουσιάζει τις μέσες τιμές των D- αποδόσεων και τις τυπικές αποκλίσεις των D- βέλτιστων σχεδιασμών, καθώς, επίσης και των τυχαίων δειγμάτων για κάθε αριθμό k των παρατηρήσεων ανά σχεδιασμό για τα οκτώσύνολα δεδομένων. Γενικά, οι D – βέλτιστοι σχεδιασμοί κατέχουν υψηλότερες μέσες D-αποδοτικότητες με μικρότερες διακυμάνσεις σε σχέση με τα τυχαία δείγματα. Οι διαφορές της D- αποδοτικότητας ανάμεσα στον D- βέλτιστο σχεδιασμό και τους τυχαίους σχεδιασμούς είναι μεγαλύτερες όταν ο αριθμός των παρατηρήσεων k στους σχεδιασμούς είναι μικρός. Για τα περισσότερα σύνολα δεδομένων οι D-αποδοτικότητες αυξάνουν με αύξηση του k. Σε αυτές τις περιπτώσεις, συχνά υπάρχει ένας βέλτιστος αριθμός παρατηρήσεων με μέγιστη D–αποδοτικότητα. Επιπλέον, αν το k μεγαλώνει πλησιάζουμε στις πραγματικές D –αποδοτικότητες των συνόλων δεδομένων, με εξαίρεση το σύνολο δεδομένων για την ισορροπία.

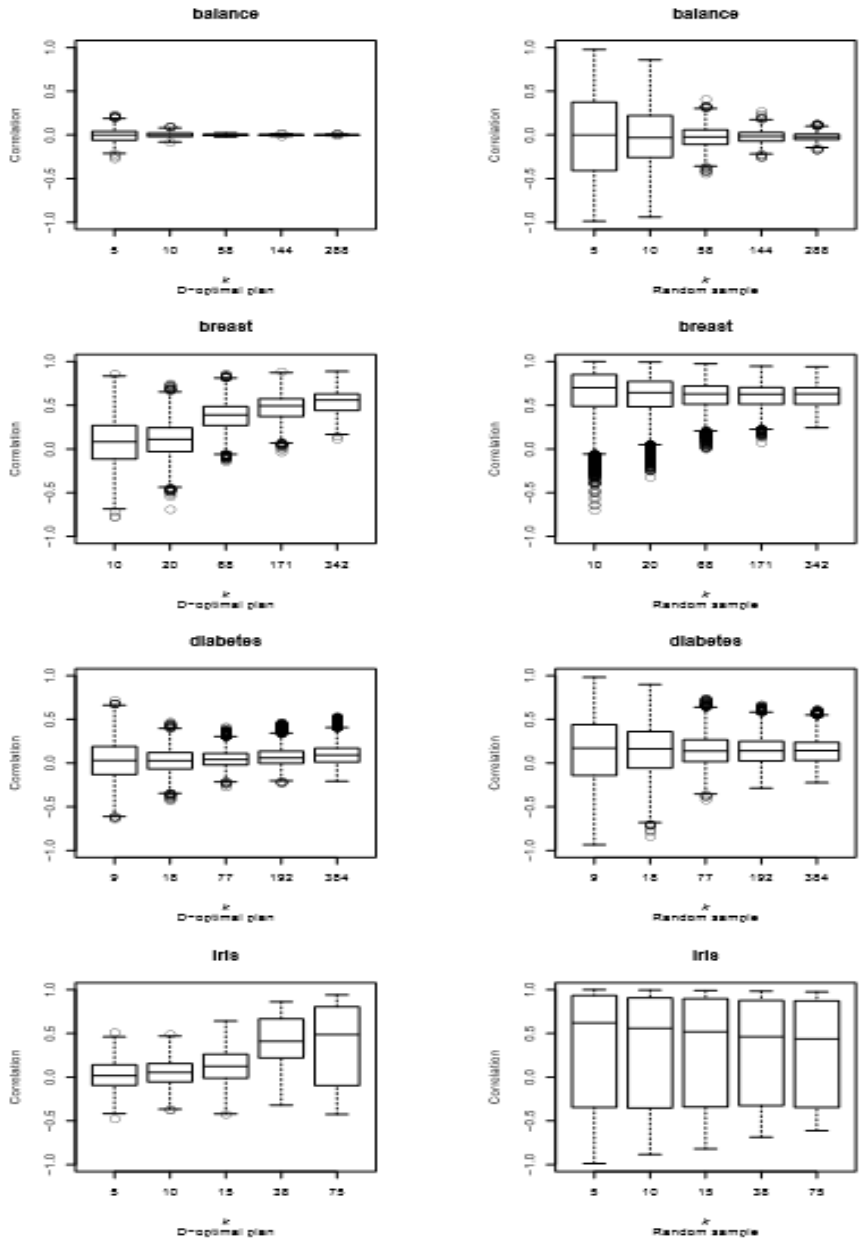


Σχήμα 4.5: Μέσες D – αποδοτικότητες και τυπικές αποκλίσεις του D–βέλτιστου σχεδιασμού 0.25n, 0.5n, 0.9n} των παρατηρήσεων ανά σχεδιασμό.

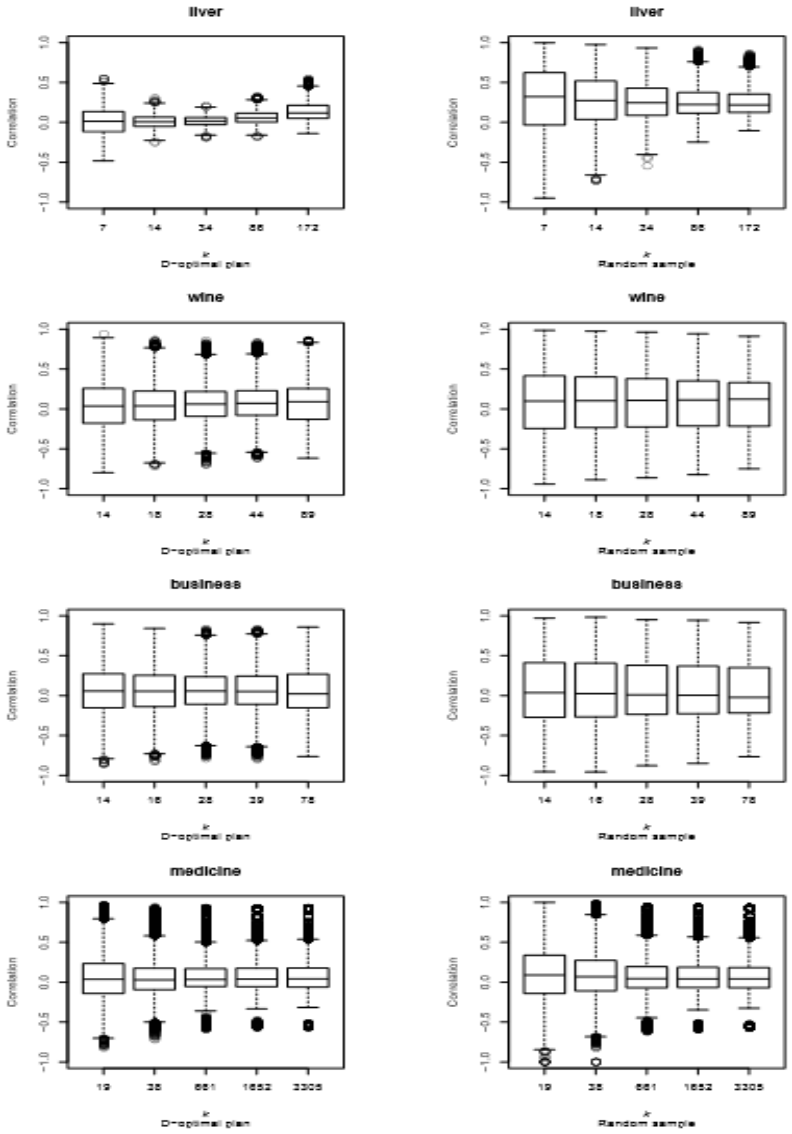
Οι D–αποδοτικότητες των D –βέλτιστων σχεδιασμών είναι πολύ διαφορετικές για ξεχωριστά σύνολα δεδομένων. Για τα σύνολα δεδομένων της ισορροπίας, του διαβήτη, της ίριδας και του συκωτιού, οι D– αποδοτικότητες φτάνουν στο 1, ενώ για παράδειγμα για το σύνολο δεδομένων της ιατρικής, οι αντίστοιχες τιμές είναι χαμηλότερες από 0.6 για όλα τα k. Οι παράμετροι του γενετικού αλγορίθμου επιλέχθηκαν σύμφωνα με τη μέθοδο ad –hoc. Ίσως, μέσω βελτιστοποιημένων παραμέτρων θα μπορούσαν και των τυχαία επιλεγμένων σχεδιασμών για διαφορετικούς αριθμούς k∈ {p+1, 2(p+1), 0.1n,

να επιτευχθούν υψηλότερες τιμές για την D– αποδοτικότητα για μερικά σύνολα δεδομένων. Ωστόσο, βάσει των υποψήφιων σημείων που έχουν αποφασιστεί από κάθε σύνολο ξεχωριστά, ίσως δεν είναι πάντα πιθανό να επιτευχθεί η ορθογωνιότητα.

Τα παράλληλα boxplots των συσχετίσεων μεταξύ των επεξηγηματικών μεταβλητών σε D- βέλτιστους σχεδιασμούς και τυχαία δείγματα, παρουσιάζονται στα γραφήματα 4.6 και 4.7. Στους D- βέλτιστους σχεδιασμούς οι συσχετίσεις μεταξύ των προγνωστών βρίσκονται σημαντικά κοντά στο μηδέν σε σύγκριση με τα τυχαία δείγματα. Επιπλέον, οι διακυμάνσεις των συσχετίσεων είναι πολύ μικρότερες.



Σχήμα 4.6: Συσχετίσεις μεταξύ όλων των ζευγών των επεξηγηματικών μεταβλητών για D–βέλτιστο και τυχαίους σχεδιασμούς για διαφορετικούς αριθμούς $k \in \{p+1, 2(p+1), 0.1n, 0.25n, 0.5n, 0.9n\}$ των παρατηρήσεων ανά σχεδιασμό.



Σχήμα 4.7: Συσχετίσεις μεταξύ όλων των ζευγών των επεξηγηματικών μεταβλητώνγιαD–βέλτιστο και τυχαίους σχεδιασμούς για διαφορετικούς αριθμούς k ∈ {p+1, 2(p+1), 0.1n, 0.25n, 0.5n, 0.9n} των παρατηρήσεων ανά σχεδιασμό.

Όπως έχει περιγραφεί σε προηγούμενη ενότητα, ο Pumplun (2005a) πρότεινε τη χρήση όχι μόνο του σχεδιασμού με την μεγαλύτερη D–τιμή για την επιλογή μεταβλητών ή την εκτέλεση, αλλά και τη χρήση ενός συνόλου 100 σχεδιασμών με τις μεγαλύτερες D-τιμές. Σε αυτή την παράγραφο ερευνάται αν αυτή η προσέγγιση έχει νόημα. Θυμίζουμε ότι βρισκόμαστε σε μια κατάσταση όπου οι ετικέτες των κλάσεων μερικών μόνο παρατηρήσεων προς εκτέλεση μπορούν να προσδιοριστούν. Θα μπορούσε κάποιος να υποθέσει ότι πιθανώς οι 100D–βέλτιστοι σχεδιασμοί σε ένα σύνολο επικαλύπτονται τόσο έντονα ώστε το ποσοστό των παρατηρήσεων για τις οποίες η ετικέτα της κλάσης δεν χρειάζεται να καθοριστεί, δεν είναι μεγάλο.

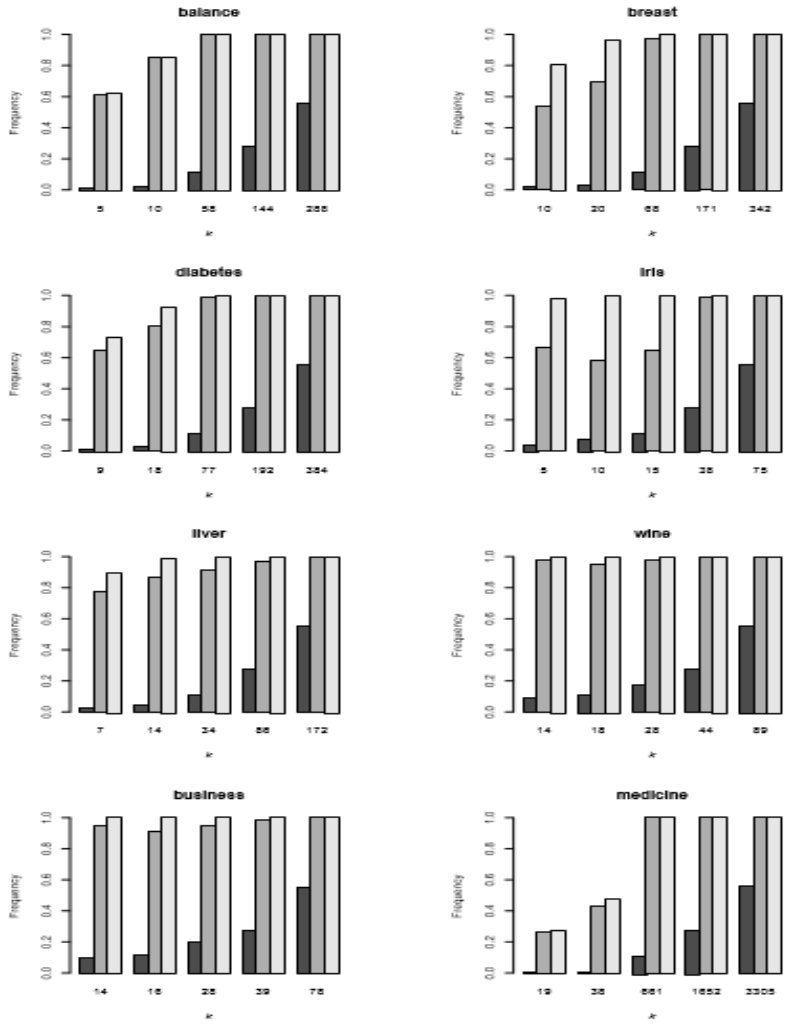
Προκειμένου να ελεγχθεί το τελευταίο, υπολογίζονται τα ποσοστά των παρατηρήσεων που εμπλέκονται στην επιλογή των μεταβλητών ή στην εκτέλεση των μεθόδων ταξινόμησης. Εφόσον η τάξη σφάλματος εκτιμάται μέσω της δεκαπλάσιας διασταυρωμένης επικύρωσης, το μέσο μέγεθος των συνόλων δεδομένων προς εκτέλεση είναι 0.9n. Έτσι, για ένα μόνο D–βέλτιστο σχεδιασμό ή τυχαίο δείγμα, το μέσο ποσοστό των συνόλων παρατηρήσεων που εμπλέκονται στην επιλογή ή στην εκτέλεση είναι $k/0.9n$. Υπολογίζονται, επίσης, τα μέσα ποσοστά των παρατηρήσεων στα σύνολα των D–βέλτιστων σχεδιασμών και στα σύνολα των τυχαία επιλεγμένων σχεδιασμών. Οι μέσες τιμές και στα οκτώ σύνολα δίνονται στον Πίνακα 4.9. Επιπροσθέτως, οι τιμές για μεμονωμένα σύνολα δεδομένων παρουσιάζονται στο Σχήμα 4.8.

Πίνακας 4.9: Μέσες τιμές ποσοστών των παρατηρήσεων που εμπλέκονται στην επιλογή μεταβλητών ή και στην εκτέλεση.

			k		
	p+1	**2(p+1)**	**0.1n**	**0.25n**	**0.5n**
doptimal/random obs	0.04	0.07	0.11	0.28	0.56
doptimal it	0.68	0.77	0.92	0.99	1.00
random obs it	0.79	0.90	1.00	1.00	1.00

Στην περίπτωση που χρησιμοποιείται ένα σύνολο D- βέλτιστων σχεδιασμών, απαιτείται χαμηλότερο ποσοστό παρατηρήσεων σε σύγκριση με την περίπτωση χρήσης τυχαία επιλεγμένων σχεδιασμών. Παρόλα αυτά, τα ποσοστά είναι πολύ μεγαλύτερα. Μόνο για k = p +1 πρέπει να προσδιοριστούν οι ετικέτες της κλάσεις για περισσότερες από τον 50% όλων των παρατηρήσεων. Για k = 0.1n, ήδη, κατά μέσο όρο όλες οι παρατηρήσεις εμφανίζονται σε ένα σύνολο σχεδιασμών. Συνεπώς, το ποσοστό του Pumplun δεν βγάζει νόημα και έτσι στα παρακάτων δεν ακολουθείται αυτή η προσέγγιση.

Γενικά, θα μπορούσε να σημειωθεί ότι η έρευνα για D - βέλτιστους σχεδιασμούς στο σύνολο δεδομένων ήταν επιτυχής. Μέσω του γενετικού αλγορίθμου βρήκαμε D-βέλτιστους σχεδιασμούς με σημαντικά μεγαλύτερη D–αποδοτικότητα σε σύγκριση με τα τυχαία δείγματα και όλα τα σύνολα δεδομένων. Επιπροσθέτως, για μερικά σύνολα δεδομένων επιτεύχθηκε η τιμή 1 για την D–αποδοτικότητα, που σημαίνει ορθογωνιότητα άρα και επιτυχία.



Σχήμα 4.8:Μέσα ποσοστά των παρατηρήσεων που περιέχονται στην επιλογή μεταβλητών ή και στην εκτέλεση (doptimal /randomobs : σκούρο γκρι, random obs it : ανοικτό γκρι).

4.9 D – βέλτιστοι σχεδιασμοί σαν βάση για επιλογή μεταβλητών

Σε αυτή την ενότητα ερευνάται αν οι D- βέλτιστοι σχεδιασμοί είναι οφέλιμοι για την επιλογή μεταβλητών. Προκειμένου να εκτιμηθεί η επίδραση της D –βελτιστοποίησης πάνω στην τάξη σφάλματος ξεχωριστά, διενεργείται μόνο επιλογή μεταβλητών βασισμένη σε D–βέλτιστους σχεδιασμούς. Παρόλα αυτά, χρησιμοποιούνται όλα τα σύνολα δεδομένων για την εκτίμηση των μοντέλων ταξινόμησης.

Εφόσον, τα σωστά υποσύνολα μεταβλητών είναι άγνωστα, η ορθότητα μπορεί να μετρηθεί μόνο μέσω της τάξης σφάλματος. Προκειμένου να εκτιμηθεί αν η επιλογή μεταβλητών σε έναν D–βέλτιστο σχεδιασμό δίνει χαμηλότερη τάξη σφάλματος σε σχέση με τα τυχαία δείγματα, συγκρίνονται τα σφάλματα που προκύπτουν από την vs doptimal και την vs random. Η σύγκριση γίνεται με όρους μέσης σχετικής απόκλισης

MRD (vs random obs, vs doptimal) και σχετικών συχνοτήτων. Συνεπώς, καθώς η τάξη σφάλματος εξαρτάται από πολλούς άλλους διαφορετικούς παράγοντες, όπως το σύνολο των δεδομένων ή τη μέθοδο ταξινόμησης, χρησιμοποιείται ένα γραμμικό μοντέλο και διεξάγεται μια ανάλυση διακύμανσης με σκοπό να εκτιμηθεί ποιοί παράγοντες είναι σημαντικοί όταν η επιλογή μεταβλητών σε D–βέλτιστους σχεδιασμούς δίνει αισθητά χαμηλότερες τάξεις σφάλματος.

Ο Πίνακας 4.10 παρουσιάζει τη μέση σχετική απόκλιση MRD (vs random obs, vs doptimal) για διαφορετικούς αριθμούς παρατηρήσεων ανά σχεδιασμό, αριθμούς μεταβλητών και μεθόδους ταξινόμησης (αποκλειστικά για το σύνολο δεδομένων της ίριδας). Μια τιμή μεγαλύτερη του μηδενός, δηλώνει ότι κατά μέσο όρο οι D-βέλτιστοι σχεδιασμοί καταλήγουν σε χαμηλότερες τάξεις σφάλματος σε σχέση με τα τυχαία δείγματα. Όπως φαίνεται, οι τιμές ποικίλλουν γύρω από το μηδέν και τα ποσοστά των θετικών και των αρνητικών τιμών είναι περίπου ίσα. Έτσι η vs doptimal φαίνεται να μην είναι ιδιαίτερα οφέλιμη για την επιλογή μεταβλητών.

Στον Πίνακα 4.11 δίνονται οι σχετικές συχνότητες για τις οποίες η επιλογή μεταβλητών στους D–βέλτιστους σχεδιασμούς καταλήγει σε χαμηλότερες τάξεις σφάλματος σε σχέση με τα τυχαία δείγματα που δίνονται. Οι τιμές ποικίλλουν γύρω από το 0.5. Συνεπώς, αυτός ο πίνακας δεν παρέχει κάποια ένδειξη για το όφελος της D– βελτιστοποίησης στην επιλογή μεταβλητών.

Πίνακας 4.10: Μέση σχετική απόκλιση MRD (vs random obs, vs doptimal) της τάξης σφάλματος που προκύπτει από την vs doptimal και την vs random obs.

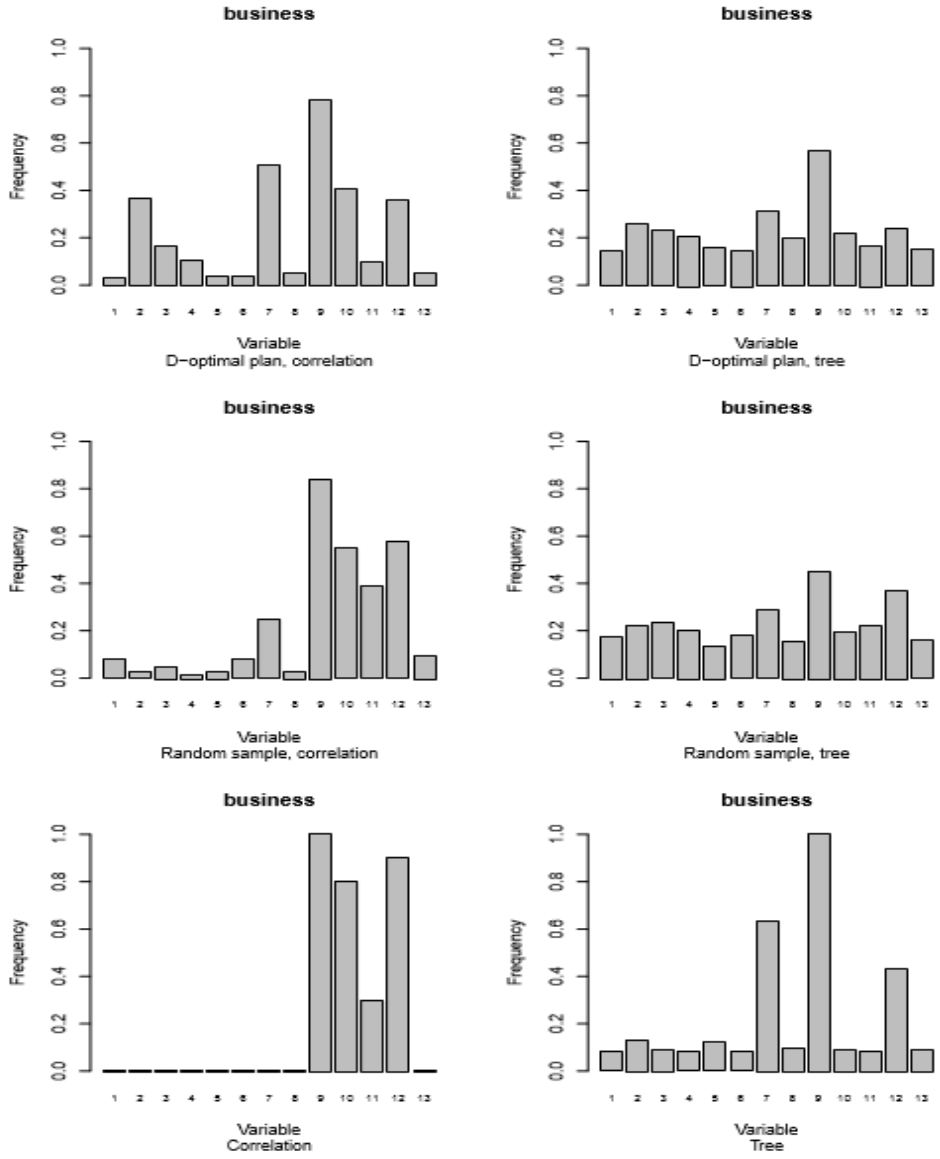
Συσχέτιση	**v**	**LDA**	**QDA**	**CART**	**SVMDOT**	**SVMRBF**	**Σύνολο**
k = p+1	0.25p	0.20	-0.17	0.05	-0.10	0.06	0.01
	0.5p	0.10	-0.01	0.02	-0.07	-0.03	0.00
	0.75p	0.09	0.02	0.02	-0.09	-0.01	0.00
k = 0.1n	0.25p	0.13	-0.09	-0.08	0.09	0.07	0.02
	0.5p	0.10	-0.04	-0.05	-0.06	0.09	0.01
	0.75p	0.09	0.02	0.02	-0.09	-0.01	0.17
k = 0.5n	0.25p	-0.02	0.08	0,02	0.04	0.06	0.03
	0.5p	-0.01	0.02	0.05	-0.01	-0.07	-0.01
	0.75p	-0.03	0.01	0.01	0.04	0.05	0.02
Δέντρο	v	LDA	QDA	CART	SVMDOT	SVMRBF	Σύνολο
k = p+1	0.25p	-0.02	0.02	0.10	-0.09	0.06	0.01
	0.5p	0.12	-0.05	-0.10	0.01	-0.04	-0.01
	0.75p	0.10	-0.03	-0.05	0.11	0.20	0.07
k = 0.1n	0.25p	-0.09	0.02	0.07	-0.01	-0.06	-0.02
	0.5p	0.36	0.13	-0.09	0.02	-0.15	0.05
	0.75p	-0.12	0.10	0.06	-0.01	0.17	0.04
k = 0.5n	0.25p	-0.04	0,10	-0.05	0.05	-0.09	-0.01
	0.5p	0.03	0.18	0.00	-0.05	-0.06	0.02
	0.75p	0.14	-0.10	-0.05	-0.07	0.01	-0.02

Πίνακας 4.11: Σχετική συχνότητα με την οποία η τάξη σφάλματος που προκύπτει από την vs doptimal είναι χαμηλότερη από την τάξη σφάλματος που προκύπτει από την vs random obs.

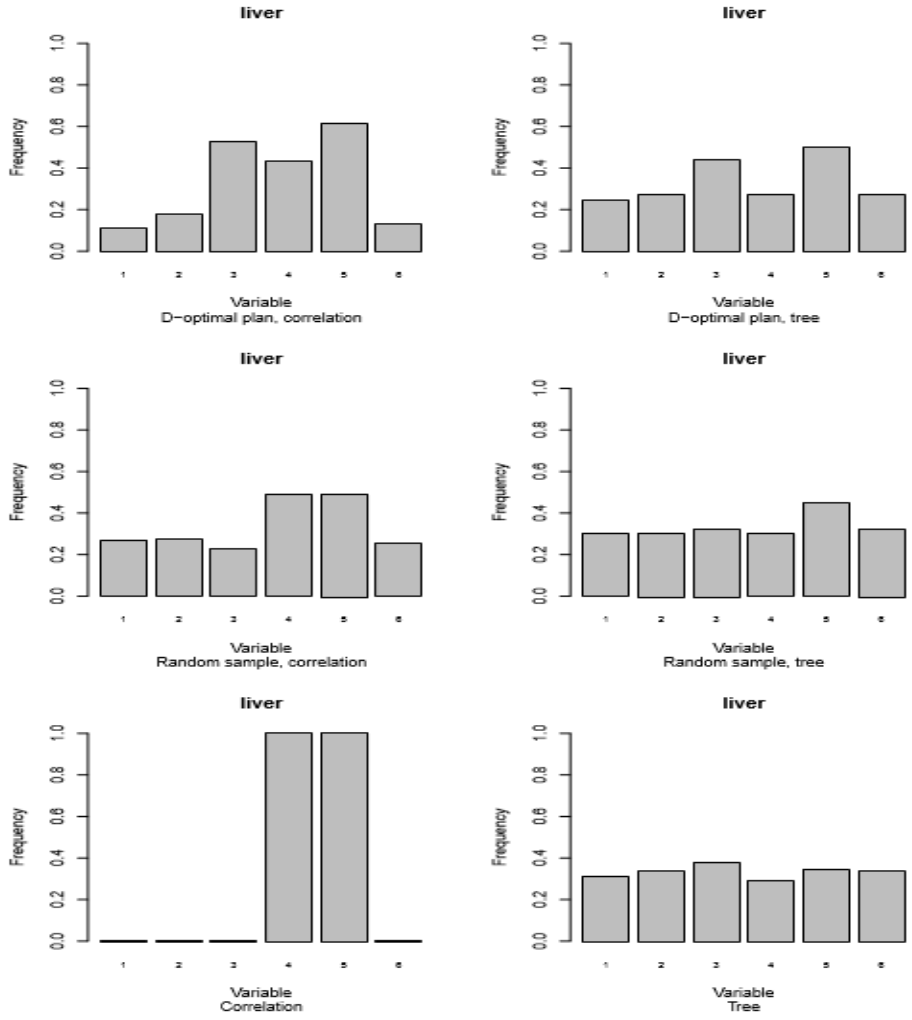
Συσχέτιση	**v**	**LDA**	**QDA**	**CART**	**SVMDOT**	**SVMRBF**	**Σύνολο**
k = p+1	0.25p	0.71	0.29	0.43	0.29	0.43	0.43
	0.5p	0.57	0.57	0.43	0.43	0.14	0.43
	0.75p	0.14	0.14	0.57	0.43	0.29	0.31
k = 0.1n	0.25p	0.43	0.29	0.43	0.43	0.86	0.49
	0.5p	0.71	0.43	0.29	0.29	0.57	0.46
	0.75p	0.43	0.57	0.43	0.71	0.14	0.46
k = 0.5n	0.25p	0.29	0.86	0.57	0.43	0.57	0.54
	0.5p	0.29	0.71	0.43	0.29	0.29	0.40
	0.75p	0.29	0.43	0.29	0.43	0.57	0.40
Δέντρο	v	LDA	QDA	CART	SVMDOT	SVMRBF	Σύνολο
k = p+1	0.25p	0.43	0.71	0.71	0.43	0.71	0.60
	0.5p	0.43	0.14	0.29	0.43	0.43	0.34
	0.75p	0.29	0.43	0.29	0.71	0.86	0.51
k = 0.1n	0.25p	0.43	0.57	0.71	0.14	0.29	0.43
	0.5p	0.86	0.57	0.29	0.29	0.29	0.46
	0.75p	0.14	0.71	0.86	0.29	0.71	0.54
k = 0.5n	0.25p	0.43	0.71	0.43	0.57	0,29	0.49
	0.5p	0.71	0.86	0.29	0.43	0.14	0.49
	0.75p	0.43	0.14	0.14	0.29	0.57	0.31

Χρησιμοποιώντας τα παραδείγματα από τα σύνολα δεδομένων της επιχείρησης και του συκωτιού εκτιμάται η περίπτωση να επιλεχθούν διαφορετικές μεταβλητές στη βάση ενός D- βέλτιστου σχεδιασμού και τυχαίων σχεδιασμών. Και για τα δύο σύνολα δεδομένων η D – αποδοτικότητα των D–βέλτιστων σχεδιασμών είναι πολύ μεγαλύτερη από την αντίστοιχη των τυχαίων δειγμάτων. Η μέση D– αποδοτικότητα των D – βέλτιστων σχεδιασμών στο σύνολο δεδομένων για το συκώτι είναι 1. Έτσι, οι ανεξάρτητες μεταβλητές είναι ασυσχέτιστες.

Τα Σχήματα 4.9 και 4.10 παρουσιάζουν την επιλογή συχνοτήτων των μεταβλητών. Η επιλογή μεταβλητών σε D – βέλτιστους σχεδιασμούς και τυχαία δείγματα είναι λιγότερο σταθερή σε σύγκριση με επιλογή βασισμένη σε όλες τις εκτελούμενες παρατηρήσεις. Το γεγονός αυτό οφείλεται στο μικρό αριθμό των παρατηρήσεων που χρησιμοποιούνται. Οι ίδιες μεταβλητές θεωρούνται περισσότερο σημαντικές και στους δύο τύπους σχεδιασμών καθώς και στο σύνολο όλων των εκτελούμενων παρατηρήσεων. Η επιλογή συχνοτήτων βασισμένη σε όλες τις εκτελούμενες παρατηρήσεις καθώς και στη βάση των τυχαίων δειγμάτων δίνει πολύ παρόμοια αποτελέσματα. Στη βάση των D–βέλτιστων σχεδιασμών επιλέγονται ελαφρώς διαφορετικές μεταβλητές. Έτσι, η D– βελτιστοποίηση παρουσιάζει μια κλίση προς την επιλογή των μεταβλητών , αλλά αυτή η κλίση δεν φαίνεται να είναι οφέλιμη για την τάξη σφάλματος.



Σχήμα 4.9:Επιλογή συχνοτήτων των μεταβλητών με χρήση των κριτηρίων συσχέτισης και δέντρων καθώς και τυχαία επιλογή μεταβλητών στο σύνολο δεδομένων του συνόλου δεδομένων της επιχείρησης για v = 0.25p =3.



Σχήμα 4.10: Επιλογή συχνοτήτων των μεταβλητών χρησιμοποιώντας τα κριτήρια συσχέτισης και δέντρων καθώς και τυχαία επιλογή μεταβλητών στο σύνολο δεδομένων για το συκώτι για v = 0.25p = 2.

Η τάξη σφάλματος εξαρτάται από τον τύπο του σχεδιασμού (D –βέλτιστο ή τυχαίο), τον αριθμό των παρατηρήσεων ανά σχεδιασμό και τον αριθμό των επιλεγμένων μεταβλητών, την επιλογή κριτηρίου, τη μέθοδο ταξινόμησης και το σύνολο δεδομένων. Προκειμένου να εκτιμηθεί ποιοί παράγοντες έχουν επίδραση πάνω στην τάξη σφάλματος, χρησιμοποιείται ένα γραμμικό μοντέλο που περιλαμβάνειέναν όρο διακοπής , κύριες επιδράσεις και αλληλεπιδράσεις δύο παραγόντων. Η επίδραση ενός παράγοντα σημαίνει αύξηση του ποσοστού του σφάλματος και ποσοστό ορθότητας από τον παράγοντα 2.7. Γίνεται, επίσης, κωδικοποίηση της απόκλισης των επιδρώντων παραγόντων από τις μέσες τιμές. Τα τυχαία δείγματα κωδικοποιούνται με -1, ενώ οι D – βέλτιστοι σχεδιασμοί κωδικοποιούνται με +1. Έτσι, αν η εκτιμώμενη επίδραση ενός τύπου παράγοντα του σχεδιασμού είναι αρνητικός αυτό σημαίνει ότι οι D–βέλτιστοι σχεδιασμοί είναι οφέλιμοι εφόσον η τάξη σφάλματος μειώνεται. Η κωδικοποίηση της απόκλισης από τις μέσες τιμές οδηγεί σε ένα διαγώνιο πίνακα πληροφορίας. Έτσι, οι κύριες επιδράσεις και οι αλληλεπιδράσεις δύο παραγόντων δε

συγχέονται. Τα αποτελέσματα από το σύνολο δεδομένων για την ίριδα δεν χρησιμοποιούνται.

Όπως φαίνεται από την ανάλυση του πίνακα διακύμανσης των κύριων μεταβλητών, ο τύπος του σχεδιασμού που χρησιμοποιείται για την επιλογή μεταβλητών δεν έχει επίδραση στην τάξη σφάλματος. Όλοι οι άλλοι παράγοντες που επηρεάζουνιδιαίτερα τα σύνολα δεδομένων και τον αριθμό των επιλεγμένων μεταβλητών, έχουν σημαντική επίδραση.

Πίνακας 4.12: Ανάλυση του πίνακα διακύμανσης των κύριων επιδράσεων.Το R^2 είναι προσαρμοσμένο στην τιμή 0.993.

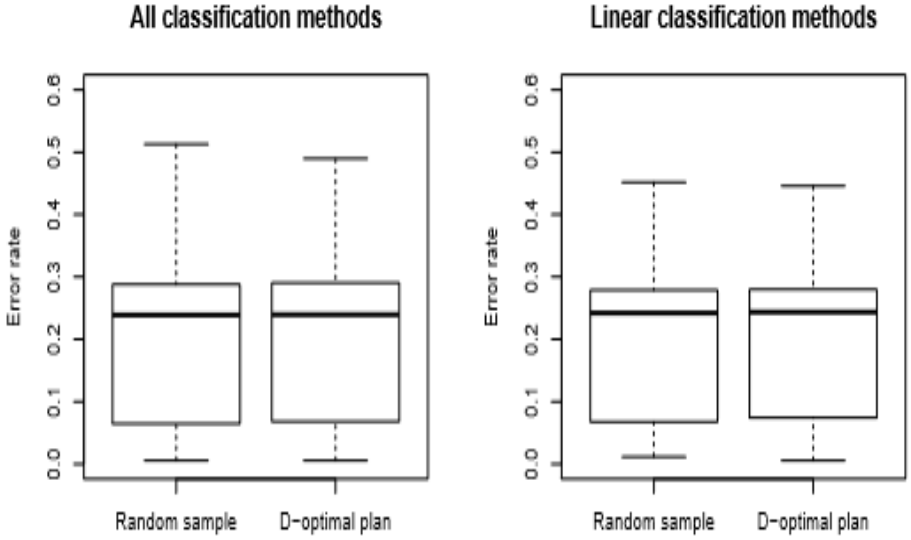
Επιδρών παράγοντας	**Βαθμοί ελευθερίας**	**F - στατιστικό**	**p -τιμή**
Σχεδιασμός	1	0.01	0.91
Αριθμός παρατηρήσεων	4	43.87	0.00
Αριθμός επιλεγμένων μεταβλητών	2	1049.64	0.00
Κριτήριο	1	111.73	0.00
Μέθοδος ταξινόμησης	4	161.38	0.00
Σύνολο δεδομένων	6	11030.22	0.00
Υπόλοιπα	1956		

Επιπροσθέτως, ένα ξεχωριστό μοντέλο έχει προσαρμοστεί για γραμμικές μεθόδους ταξινόμησης (LDA και SVMDOT). Όμως, για γραμμικές μεθόδους ταξινόμησης ο τύπος του σχεδιασμού δεν έχει ιδιαίτερη επίδραση πάνω στην τάξη σφάλματος.

Το Σχήμα 4.11 παρουσιάζει σε boxplots τις τάξεις σφάλματος των D-βέλτιστων σχεδιασμών και τυχαίων σχεδιασμών για όλες τις μεθόδους ταξινόμησης και ιδιαίτερα για γραμμικές μεθόδους ταξινόμησης. Αυτό επικυρώνει τα αποτελέσματα ανάλυσης της διακύμανσης.

Πίνακας 4.13: Ανάλυση του πίνακα διακύμανσης των κύριων επιδράσεων για τις γραμμικές μεθόδους ταξινόμησης LDA και SVMDOT.Το R^2είναι προσαρμοσμένο στην τιμή 0.997.

Επιδρών παράγοντας	**Βαθμοί ελευθερίας**	**F - στατιστικό**	**p -τιμή**
Σχεδιασμός	1	1.19	0.28
Αριθμός παρατηρήσεων	4	28.04	0.00
Αριθμός επιλεγμένων μεταβλητών	2	685.01	0.00
Κριτήριο	1	96.70	0.00
Μέθοδος ταξινόμησης	4	4.6804	0.03
Σύνολο δεδομένων	6	14783.70	0.00
Υπόλοιπα	741		



Σχήμα 4.11: Οι τάξεις σφαλμάτων που προκύπτουν από την επιλογή μεταβλητών πάνω σε τυχαία δείγματα και σε D– βέλτιστους σχεδιασμούς.

Τέλος, χρησιμοποιούνται γραμμικά μοντέλα και διεξάγεται ανάλυση της διακύμανσης για ξεχωριστά σύνολα δεδομένων. Ο τύπος του σχεδιασμού έχει ιδιαίτερη επίδραση για τα σύνολα δεδομένων για το συκώτι και την επιχείρηση. Όμως, τα σημεία των εκτιμώμενων επιδράσεων των παραγόντων είναι διαφορετικά. Αυτό σημαίνει ότι στην περίπτωση του συνόλου δεδομένων για το συκώτι, οι D– βέλτιστοι σχεδιασμοί είναι οφέλιμοι, ενώ για το σύνολο δεδομένων της επιχείρησης λαμβάνονται καλύτερα αποτελέσματα μέσω του τυχαίου σχεδιασμού.

Πίνακας 4.14: Εκτιμώμενες επιδράσεις παραγόντων και F –tests για μεμονωμένα σύνολα δεδομένων.

Σύνολο δεδομένων	Επίδραση του τύπου σχεδιασμού	F -στατιστικό	p - τιμή	R^2
Ισορροπία	0.012	1.06	0.30	0.996
Στήθος	0.022	0.80	0.37	0.999
Διαβήτης	0.043	2.17	0.14	0.996
Συκώτι	-0.128	13.71	0.00	0.958
Κρασί	-0.027	1.09	0.30	0.990
Επιχείρηση	0.058	3.62	0.06	0.990
Ιατρική	0.012	0.85	0.36	0.994

4.10 D- βέλτιστοι σχεδιασμοί σαν βάση για εκτέλεση

Σε αυτή την ενότητα ερευνάται η καταλληλότητα των D–βέλτιστων σχεδιασμών για εκτέλεση των μεθόδων ταξινόμησης. Δε γίνεται καμία επιπρόσθετη επιλογή μεταβλητών.

Οι σχεδιασμοί μεγέθους p +1 και 2(p+1) συχνά περιλαμβάνουν πολύ λίγες παρατηρήσεις για εκτέλεση των μεθόδων ταξινόμησης. Ιδιαίτερα για την QDA εξαιτίας της μοναδικής διακύμανσης εντός κλάσης, οι πίνακες σφάλματος για k= p+1 δεν είναι διαθέσιμοι. Για k = 2(p+1) λαμβάνουμε αποτελέσματα μόνο για τρία σύνολα δεδομένων.

Πίνακας 4.15: Μέση σχετική διακύμανση MRD (est random obs, est doptimal) (αποκλειστικά για το σύνολο δεδομένων της ίριδας)της τάξης σφάλματος που προκύπτει από΄τη est doptimal και την random obs.

k	LDA	QDA	CART	SVMDOT	SVMRBF	Σύνολο
p+1	0.00	-	-0.04	-0.07	-0.28	-0.10
2(p+1)	0.28	0.08	-0.21	0.14	-0.29	-0.01
0.1n	0.17	0.17	0.03	0.06	-0.14	0.05
0.25n	0.11	0.01	-0.02	0.09	-0.06	0.03
0.5n	0.03	0.00	0.00	0.15	0.05	0.05

Πρώτον, συγκρίνονται τα σφάλματα που που προκύπτουν από την est doptimal και την est random. Ο Πίνακας 4.15 παρουσιάζει τη μέση σχετική διακύμανση MRD (est random obs, est doptimal) για διαφορετικούς αριθμούς παρατηρήσεων ανά σχεδιασμό και τις πέντε μεθόδους ταξινόμησης που λαμβάνονται υπόψη. Με εξαίρεση την SVMRBF, οι περισσότερες τιμές στον πίνακα 4.15 είναι θετικές. Αυτό σημαίνει πως η est doptimal καταλήγει σε χαμηλότερες τάξεις σφάλματος σε σχέση με την estrandomobs. Ιδιαίτερα για μέτριες τιμές του k, 0.1n και 0.25n, οι D–βέλτιστοι σχεδιασμοί φαίνονται οφέλιμοι για την εκτέλεση των παρατηρήσεων.

Πίνακας 4.16: Σχετική συχνότητα με την οποία οι μέθοδοι ταξινόμησης βασισμένες σε D–βέλτιστους σχεδιασμούς καταλήγουν σε μικρότερο σφάλμα σε σχέση με τις μεθόδους που βασίζονται σε τυχαίο δείγμα.

k	LDA	QDA	CART	SVMDOT	SVMRBF	Σύνολο
p+1	0.43	-	0.71	0.29	0.14	0.39
2(p+1)	1.00	0.67	0.29	0.43	0.14	0.48
0.1n	1.00	0.60	0.71	0.71	0.43	0.70
0.25n	1.00	0.43	0.43	0.86	0.57	0.66
0.5n	0.86	0.57	0.43	0.71	0.57	0.63

Ο Πίνακας 4.16 ο οποίος παρουσιάζει τις σχετικές συχνότητες για τις οποίες οι D–βέλτιστοι σχεδιασμοί καταλήγουν σε μικρότερο σφάλμα σε σύγκριση με τα τυχαία δείγματα, επιβεβαιώνει αυτές τις παρατηρήσεις. Περισσότερο εποφελείται η LDA, ενώ για την SVMRBF το σφάλμα μάλλον χειροτερεύει αν οι D– βέλτιστοι σχεδιασμοί χρησιμοποιούνται για την εκτέλεση.

Όπως έχει ήδη περιγραφεί, στην μελέτη έχει προσαρμοστεί ένα γραμμικό μοντέλο και διεξάγεται ανάλυση της διακύμανσης. Οι παράγοντες με επίδραση είναι ο τύπος του σχεδιασμού, ο αριθμός των παρατηρήσεων ανά σχεδιασμό, η μέθοδος ταξινόμησης και το σύνολο δεδομένων. Το μοντέλο περιέχει έναν όρο διακοπής, κύριες επιδράσεις και αλληλεπιδράσεις δύο παραγόντων. Εξαιτίας του χαμένου σφάλματος της QDA, χάνεται και η ορθογωνιότητα. Συνεπώς, οι επιδράσεις των παραγόντων συγχέονται σε στην περίπτωση που χρησιμοποιούνται όλα τα διαθέσιμα αποτελέσματα. Προκειμένου να αποφευχθεί αυτό το πρόβλημα, παραλείπονται, επίσης, τα εναπομείνοντα αποτελέσματα της QDA και έτσι λαμβάνουμε και πάλι σαν αποτέλεσμα έναν ορθογώνιο σχεδιασμό.

Πίνακας 4.17: Ανάλυση του πίνακα διακύμανσης. Το R^2 είναι προσαρμοσμένο στην τιμή 0.984

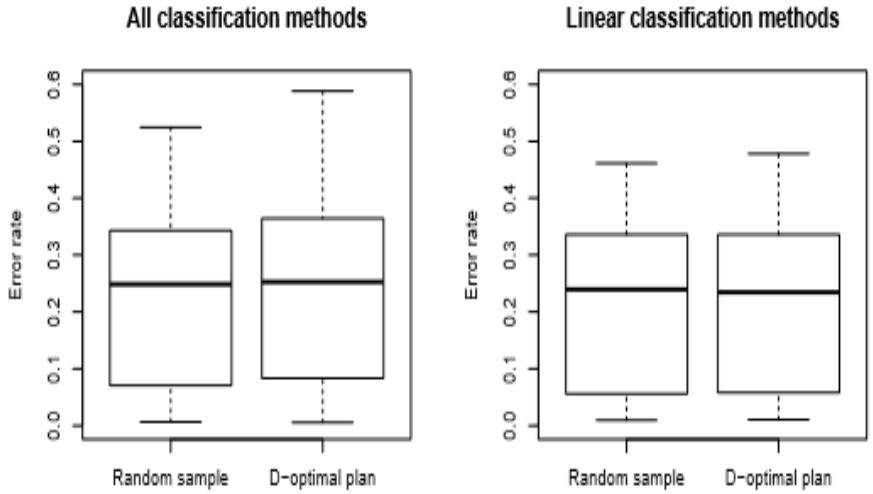
Επιδρών παράγοντας	Βαθμοί ελευθερίας	F -στατιστικό	p -τιμή
Σχεδιασμός	1	3.93	0.05
Αριθμός παρατηρήσεων	4	339.69	0.00
Μέθοδος ταξινόμησης	3	99.91	0.00
Σύνολο δεδομένων	6	177.43	0.00
Σχεδιασμός * Αριθμός παρατηρήσεων	4	3.68	0.01
Σχεδιασμός * Μέθοδος ταξινόμησης	3	7.48	0.00
Σχεδιασμός * Σύνολο δεδομένων	6	3.79	0.00
Αριθμός παρατηρήσεων * Μέθοδος ταξινόμησης	12	2.93	0.00
Αριθμός παρατηρήσεων * Σύνολο δεδομένων	24	10.23	0.00
Μέθοδος ταξινόμησης * Σύνολο δεδομένων	18	12.83	0.00
Υπόλοιπα	198		0.00

Ο Πίνακας 4.17 παρουσιάζει τα αποτελέσματα ανάλυσης της διακύμανσης. Παρά το γεγονός ότι οι υπόλοιποι παράγοντες έχουν μεγαλύτερη επιρροή στο προκύπτον σφάλμα, ο τύπος του σχεδιασμού φαίνεται να έχει κι αυτός κάποιο αντίκτυπο. Η εκτιμώμενη επίδραση του τύπου του σχεδιασμού είναι 0.040. Αυτό σημαίνει πως η εκτέλεση στη βάση ενός D–βέλτιστου σχεδιασμού αυξάνει ελαφρά την τάξη σφάλματος.

Αν διεξαχθεί ανάλυση της διακύμανσης ειδικά για τις γραμμικές μεθόδους ταξινόμησης (LDA και SVMDOT), προκύπτει αντίθετο αποτέλεσμα. Ο τύπος του σχεδιασμού φαίνεται να έχει επίδραση στην τάξη του σφάλματος αλλά η εκτιμώμενη επίδραση του παράγονται είναι -0.054. Το τελευταίο σημαίνει ότι η D–βελτιστοποίηση μάλλον βοηθά στη μείωση της τάξης σφάλματος στην περίπτωση γραμμικών μεθόδων ταξινόμησης.

Πίνακας 4.18: Ανάλυση του πίνακα διακύμανσης για γραμμικές μεθόδους ταξινόμησης. Το R^2 είναι προσαρμοσμένο στην τιμή 0.994.

Επιδρών παράγοντας	**Βαθμοί ελευθερίας**	**F -στατιστικό**	**p -τιμή**
Σχεδιασμός	1	5.59	0.02
Αριθμός παρατηρήσεων	4	352.54	0.00
Μέθοδος ταξινόμησης	1	29.34	0.00
Σύνολο δεδομένων	6	213.16	0.00
Σχεδιασμός * Αριθμός παρατηρήσεων	4	3.10	0.02
Σχεδιασμός * Μέθοδος ταξινόμησης	1	3.00	0.09
Σχεδιασμός * Σύνολο δεδομένων	6	3.62	0.00
Αριθμός παρατηρήσεων * Μέθοδος ταξινόμησης	4	8.27	0.00
Αριθμός παρατηρήσεων * Σύνολο δεδομένων	24	10.30	0.00
Μέθοδος ταξινόμησης * Σύνολο δεδομένων	6	1.18	0.00
Υπόλοιπα	82		0.32



Σχήμα 4.12: Σφάλμα που προκύπτει από την εκτέλεση σε τυχαία δείγματα και σε D–βέλτιστους σχεδιασμούς

Ο Πίνακας 4.19 παρουσιάζει τα αποτελέσματα ανάλυσης της διακύμανσης για ξεχωριστά σύνολα δεδομένων. Για όλα τα σύνολα δεδομένων εκτός της ισορροπίας, του κρασιού και της ιατρικής, ο τύπος του σχεδιασμού φαίνεται να έχει επίδραση πάνω στο σφάλμα. Για το σύνολο δεδομένων του συκωτιού, οι D –βέλτιστοι σχεδιασμοί είναι οφέλιμοι, ενώ για τα σύνολα δεδομένων του στήθους, του διαβήτη και της επιχείρησης, τα τυχαία δείγματα καταλήγουν σε μικρότερα σφάλματα.

Πίνακας 4.19: Εκτιμώμενες επιδράσεις παραγόντων και F –testγια ξεχωριστά σύνολα δεδομένων.

Σύνολο δεδομένων	**Επίδραση τύπουσχεδιασμού**	**F - στατιστικό**	**p -τιμή**	R^2
Ισορροπία	-0.023	1.57	0.23	0.996
Στήθος	0.198	5.59	0.04	0.957
Διαβήτης	0.075	26.40	0.00	0.999
Συκώτι	-0.110	9.39	0.01	0.989
Κρασί	0.025	0.34	0.57	0.988
Επιχείρηση	0.128	9.53	0.01	0.985
Ιατρική	-0.016	1.71	0.22	1.000

4.11 D – βέλτιστοι σχεδιασμοί σαν βάση για επιλογή μεταβλητών και εκτέλεση

Σε αυτή την ενότητα συγκρίνονται τα σφάλματα που προκύπτουν από την επιλογή μεταβλητών και την εκτέλεση στη βάση D–βέλτιστων σχεδιασμών, με χρήση της vs est doptimal και τυχαίων δειγμάτων και της vs random obs.

Ο Πίνακας 4.20 παρουσιάζει τη μέση σχετική απόκλιση MRD(vs est random obs, vs est doptimal) για διαφορετικούς αριθμούς παρατηρήσεων και μεταβλητών, μεθόδων ταξινόμησης και κριτηρίων επιλογής μεταβλητών.

Πίνακας 4.20: Μέση σχετική απόκλιση MRD (vs random obs, est dopitmal) (αποκλειστικά για το σύνολο δεδομένων της ίριδας)του σφάλματος που προκύπτει από την vs est doptimal και την vs est random obs.

Συσχέτιση	**v**	**LDA**	**QDA**	**CART**	**SVMDOT**	**SVMRBF**	**Σύνολο**
k = p+1	0.25p	0.14	-0.19	-0.06	-0.18	-0.22	-0.10
	0.5p	-0.01	-	-0.07	-0.01	-0.26	-0.09
	0.75p	0.28	-	-0.07	-0.09	-0.23	-0.03
k = 0.1n	0.25p	0.09	-0.10	0.04	-0.04	-0.06	-0.01
	0.5p	0.12	0.17	0.03	0.08	-0.16	0.04
	0.75p	0.07	0.16	0.05	-0.03	-0.11	0.02
k = 0.5n	0.25p	0.07	0.02	0.03	0.02	0.04	0.04
	0.5p	0.02	-0.05	0.04	0.08	0.06	0.03
	0.75p	0.12	0.00	0.00	0.06	-0.02	0.03
Δέντρο	v	LDA	QDA	CART	SVMDOT	SVMRBF	Σύνολο
k = p+1	0.25p	-0.12	-0.12	-0.04	-0.09	-0.17	-0.11
	0.5p	-0.03	-0.53	0.03	-0.11	-0.22	-0.11
	0.75p	0.34	-	0.03	0.11	-0.24	0.06
k = 0.1n	0.25p	-0.10	-0.12	0.04	-0.05	-0.09	-0.06
	0.5p	0.12	0.33	0.02	0.01	-0.12	0.07
	0.75p	0.11	0.08	0.01	-0.03	-0.09	0.01
k = 0.5n	0.25p	-0.03	0.02	-0.01	-0.03	-0.06	-0.02
	0.5p	0.13	0.10	-0.05	-0.09	-0.08	0.00
	0.75p	0.10	-0.15	0.07	0.02	0.05	-0.01

Στον Πίνακα 4.21φαίνονται οι σχετικές συχνότητες με τις οποίες το σφάλμα που προκύπτει από την vs est doptimal είναι μικρότερο από το αντίστοιχο της vs est random obs.

Πίνακας 4.21: Σχετική συχνότητα με την οποία το σφάλμα που προκύπτει από την vs est doptimal είναι χαμηλότερο από το αντίστοιχο που προκύπτει μέσω της vs est randomobs.

Συσχέτιση	**v**	**LDA**	**QDA**	**CART**	**SVMDOT**	**SVMRBF**	**Σύνολο**
k = p+1	0.25p	0.57	0.14	0.14	0.29	0.14	0.26
	0.5p	0.57	-	0.29	0.43	0.29	0.39
	0.75p	0.71	-	0.29	0.57	0.43	0.50
k = 0.1n	0.25p	0.57	0.43	0.71	0.57	0.57	0.57
	0.5p	0.86	0.80	0.86	0.86	0.14	0.70
	0.75p	0.71	0.75	0.57	0.57	0.57	0.62
k = 0.5n	0.25p	0.57	0.71	0.43	0.29	0.71	0.54
	0.5p	0.43	0.57	0.57	0.71	0.57	0.57
	0.75p	0.57	0.71	0.43	0.43	0.29	0.49
Δέντρο	v	LDA	QDA	CART	SVMDOT	SVMRBF	Σύνολο
k = p+1	0.25p	0.29	0.50	0.29	0.57	0.43	0.41
	0.5p	0.57	0.00	0.43	0.29	0.29	0.37
	0.75p	1.00	-	0.57	0.71	0.00	0.57
k = 0.1n	0.25p	0.57	0.14	0.71	0.43	0.29	0.43
	0.5p	0.71	0.43	0.43	0.57	0.43	0.51
	0.75p	0.71	0.80	0.57	0.43	0.43	0.58
k = 0.5n	0.25p	0.29	0.43	0.57	0.43	0.29	0.40
	0.5p	0.71	0.71	0.57	0.29	0.43	0.54
	0.75p	0.57	0.29	0.14	0.43	0.57	0.40

Για την QDA εξαιτίας του μικρού αριθμού των παρατηρήσεων που εκτελούνται για k = p+1, δεν είναι διαθέσιμα όλα τα σφάλματα. Στον πίνακα 20 οι τιμές ποικίλλουν γύρω από το μηδέν. Για k = p+1 οι περισσότερες τιμές είναι αρνητικές και για την SVMDOT, επίσης το MRD (vs est random obs, vs est dopitmal) είναι κυρίως μικρότερο του μηδενός. Αυτό σημαίνει ότι οι D – βέλτιστοι σχεδιασμοί μάλλον αυξάνουν την τάξη σφάλματος σε αυτές τις περιπτώσεις. Στον Πίνακα 4.21 φαίνεται ότι κυρίως για k μεσσαίου μεγέθους και μεγάλο αριθμό μεταβλητώνν, η επιλογή μεταβλητών και η εκτέλεση σε D–βέλτιστους σχεδιασμούς μπορεί να είναι οφέλιμη. Συχνά, οι μεγαλύτερες σχετικές συχνότητες παρατηρούνται για την LDA.

Όπως και στις προηγούμενες ενότητες, εφαρμόζεται ένα γραμμικό μοντέλο στα αποτελέσματα. Οι επιδρώντες παράγοντες είναι ο τύπος του σχεδιασμού (D–βέλτιστος ή τυχαίος), ο αριθμός των παρατηρήσεων ανά σχεδιασμό, ο αριθμός των επιλεγμένων μεταβλητών, το κριτήριο επιλογής, η μέθοδος ταξινόμησης και το σύνολο δεδομένων. Για άλλη μια φορά, τα αποτελέσματα από την QDA παραλείπονται. Ο Πίνακας 4.22 παρουσιάζει τα αποτελέσματα ανάλυσης της διακύμανσης.

Ο τύπος του σχεδιασμού επηρεάζει το σφάλμα. Εφόσον η εκτιμώμενη επίδραση του τύπου σχεδιασμού είναι 0.036, οι D – βέλτιστοι σχεδιασμοί μάλλον αυξάνουν την τάξη σφάλματος.

Πίνακας 4.22: Ανάλυση του πίνακα διακύμανσης κύριων επιδράσεων. Το R^2 είναι προσαρμοσμένο στην τιμή 0.971.

Επιδρών παράγοντας	Βαθμοί ελευθερίας	F - στατιστικό	p - τιμή
Σχεδιασμός	1	8.44	0.00
Αριθμός παρατηρήσεων	4	639.94	0.00
Αριθμός μεταβλητών	2	126.02	0.00
Κριτήριο	1	58.27	0.00
Μέθοδος ταξινόμησης	3	179.32	0.00
Σύνολο δεδομένων	6	1093.77	0.00
Υπόλοιπα	1551		

Τα αποτελέσματα της ανάλυσης διακύμανσης για γραμμικές μεθόδους ταξινόμησης δίνονται στον Πίνακα 4.23. Όπως σε προηγούμενες ενότητες, η εκτιμώμενη επίδραση παράγοντα είναι -0.031, δηλαδή αρνητική για γραμμικές μεθόδους ταξινόμησης. Έτσι, για γραμμικές μεθόδους ταξινόμησης η επιλογή μεταβλητών και η εκτέλεση σε D–βέλτιστους σχεδιασμούς φαίνεται να είναι οφέλιμη.

Πίνακας 4.23: Ανάλυση του πίνακα διακύμανσης των κύριων επιδράσεων για γραμμικές μεθόδους ταξινόμησης. Το R^2είναι προσαρμοσμένο στην τιμή 0.980.

Επιδρών παράγοντας	Βαθμοί ελευθερίας	F - στατιστικό	p - τιμή
Σχεδιασμός	1	2.99	0.08
Αριθμός παρατηρήσεων	4	330.33	0.00
Αριθμός μεταβλητών	2	114.71	0.00
Κριτήριο	1	72.43	0.00
Μέθοδος ταξινόμησης	1	0.86	0.35
Σύνολο δεδομένων	6	394.70	0.00
Υπόλοιπα	741		

Πίνακας 4.24: Εκτιμώμενες επιδράσεις παραγόντων και F- tests για μεμονωμένα σύνολα δεδομένων.

Σύνολο δεδομένων	Επίδραση τύπου σχεδιασμού	F - στατιστικό	p -τιμή	R^2
Ισορροπία	-0.032	4.42	0.04	0.988
Στήθος	0.237	83.32	0.00	0.974
Διαβήτης	0.084	12.84	0.00	0.987
Συκώτι	-0.275	35.59	0.00	0.874
Κρασί	0.090	15.40	0.00	0.982
Επιχείρηση	0.138	27.82	0.00	0.965
Ιατρική	0.009	0.28	0.60	0.997

Ο Πίνακας 4.24 παρουσιάζει τις επιδράσεις των παραγόντων του τύπου σχεδιασμού και τα αποτελέσματα του F–test για μεμονωμένα σύνολα δεδομένων. Για όλα τα σύνολα δεδομένων, εκτός από αυτό της ιατρικής, ο τύπος του σχεδιασμού φαίνεται να έχει επίδραση στην τάξη σφάλματος. Για την πλειονότητα των συνόλων δεδομένων, οι D–βέλτιστοι σχεδιασμοί οδηγούν σε ελαφρά χειρότερα σφάλματα σε σχέση με τα τυχαία δείγματα. Μόνο για τα σύνολα δεδομένων της ισορροπίας και της επιχείρησης η επιλογή μεταβλητών και η εκτέλεση σε D- βέλτιστους σχεδιασμούς είναι μάλλον οφέλιμη.

4.12 Σύνοψη

Σε αυτό το κεφάλαιο μελετήθηκαν η καταλληλότητα των D– βέλτιστων σχεδιασμών για την επιλογή μεταβλητών και την εκτέλεση των μεθόδων ταξινόμησης. Στην μελέτη προσομοίωσης αποδείχθηκε ότι για τα περισσότερα σύνολα δεδομένων η εφαρμογή μιας μεθόδου επιλογής μεταβλητών είναι οφέλιμη και ότι τα κριτήρια που χρησιμοποιήθηκαν (συσχέτισης και δέντρα), είναι κατάλληλα για την ανίχνευση σημαντικών μεταβλητών για την ταξινόμηση. Θα πρέπει να σημειωθεί πως το κριτήριο της συσχέτισης αποδείχθηκε ελαφρώς καλύτερο.

Οι D – βέλτιστοι σχεδιασμοί δεν αποδείχθηκαν χρήσιμοι για την επιλογή μεταβλητών, εφόσον ο τύπος του σχεδιασμού που χρησιμοποιήθηκε σαν βάση για την επιλογή δεν έχει σημαντική επίδραση πάνω στην τάξη σφάλματος. Παρόλο που στο παράδειγμα των δύο συνόλων δεδομένων φάνηκε ότι βάσει D–βέλτιστων σχεδιασμών προκύπτουν ελαφρώς διαφορετικές μεταβλητές σε σχέση με τα τυχαία δείγματα, αυτές οι διαφορές δεν φαίνεται να έχουν επιρροή πάνω στο σφάλμα.

Αντίθετα, οι D –βέλτιστοι σχεδιασμοί σαν βάση για εκτέλεση έχουν, μάλλον, μικρή επίδραση πάνων στο σφάλμα. Δυστυχώς, μέσω των D– βέλτιστων σχεδιασμών η τάξη σφάλματος αυξάνεται. Παρόλα αυτά, για γραμμικές μεθόδους ταξινόμησης, οι D – βέλτιστοι σχεδιασμοί σαν βάση για εκτέλεση των μεθόδων ταξινόμησης φαίνονται οφέλιμοι. Θα πρέπει, ωστόσο, να σημειωθεί πως οι υπόλοιποι επιδρώντες παράγοντες, όπως το σύνολο δεδομένων ή η μέθοδος ταξινόμησης έχουν μεγαλύτερη επιρροή στο αποτέλεσμα.

Το ίδιο εφαρμόζεται για D –βέλτιστους σχεδιασμούς σαν βάση για επιλογή μεταβλητών και εκτέλεση. Το σφάλμα που προκύπτει αυξάνεται ελαφρά για D– βέλτιστους σχεδιασμούς αν ληφθούν υπόψη οι μέθοδοι ταξινόμησης. Αλλά, για γραμμικές μεθόδους ταξινόμησης το προκύπτον σφάλμα μάλλον βελτιώνεται.

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. A new variable selection method for classification, Silvia Casado, Joaquin PacheroBonrostro, Laura Nuñez Letamendia

2. A SELECTIVE OVERVIEW OF VARIABLE SELECTION IN HIGH DIMENSIONAL FEATURE SPACE, Jianqing Fan and JinchiLv, Princeton University and University of Southern California

3. An Introduction to Variable and Feature Selection, Isabelle Guyon, André Elisseeff

4. Akaike, (1969), H. Fitting autoregressive models for prediction. Annals of the Institute of Statistical Mathematics, 243

5. Akaike, H. (1973), Information theory and an extension of the maximum likelihood principle. In second International Symposium on Information Theory, Eds B.N. Petrov and F. Csaki

6. Big Data Opportunities and Challenges: Discussions from Data Analytics Perspectives, Zhi – Hua Zhou, NiteshV.Chawla, YaochuJin, and Graham J. Williams

7. Breiman, L., 2001 Random forests. Machine Learning 45, 5-32

8. Brieman, 1984 L., Friedman, J.H., Olshen, R.A., Stone, J.Classification and regression trees. Wadsworth, Belmont

9. Bledsoe, 1961 W.W The use of biological concepts in the analytical study of systems. Paper presented at ORSA-TIMS National Meeting, San Fransisco

10. Bliss L. 2002 Particle Swarm Optimization.Pace University DPS program

11. C.Pumplün, S. Rüping and C. Weihs. D-optimal plans in observational studies. Technical Report 44/2005, SFB 475, Complexity reduction in multivariate data structures

12. C. Pumplün C. Weihs and A. Presser. Experimental design for variable selection in databases. In C. Weihs and Gaul, editors Classification – The Ubiquitous Challenge. Springe

13. D – optimal plans for variable selection in data bases, Schiffer, Julia; Weihs, Claus

14 .Design of Experiments: The D – Optimal Approach and Its Implementation As a Computer Algorithm Doksum, K., Tang, S., Tsui, K-W.,2008. Nonparametric variable selection: The earth algorithm. Journal of the American Statistical Association 103: (484), 1609-1620

15. Doksum, K.A., Samarov, A., 1995.Nonparametric estimation of globlal functional and a measure of explanatory power of covariates in regression. Annals of Statistics 23, 1443-1473

16.Experimental Design for Variable Selection in Data Bases, ConstanzePumplün, Claus Weihs, and Andrea Preuser, University of Dortmund- Department of Statistics- 44221 Dortmund, Germany

17. Efron, B.(1983) Estimating the error rate of a predictive rule: Improvement over cross- validation. J. Amer. Statist.Assos., 78, 316-331

18. Efroymson, M. A. Multiple regression analysis, in Mathematical methods for digital computers. Wiley, New York, 1960

19. Friedman, J., 1991.Multivariate adaptive regression splines. The annals of statistics

20. Gibbs Variable Selection Usind BUGS, IoannisNtzoufras- Department of Business Administration- University of the Aegean- Chios- Greece

21. Goldberg, D. E. Genetic Algorithms in Search, Optimization, and Machine Learning. Reading. Kluwer Academic Publishers, Boston, MA, 1989

22. Hastie, T., Tibshirani, R., Friedman, J. The elements of statistical learning, seond edition, Springer 2009

23. Holland, J.H. Adaption in natural and artificial systems. MIT Press, 1992

24. HIGH DIMENSIONAL GRAPHS AND VARIABLE SELECTION WITH THE LASSO, By NICOLAI MEINSHAUSEN AND PETER BÜHLMANN, ETH Zürich

25. HIGH–DIMENSIONAL VARIABLE SELECTION BY LARRY WASSERMAN AND KATHRYN ROEDER, Carnegie Mellon University

26. Kim, K.J., Cho S. B.A comprehensive overview of the applications of artificial life. Artificial Lifw, 12 (2006), 153-182

27. Miller, A. Subset selection in regression, second edition. Chapman & Hall/CRC, 2002

28. Nonparametric Variable Selection and Classification: The CATCHAlforithm, Shijie Tang, Lisha Chen, Kam – WahTsui, KjellDoksum

29 .Plackett R. L. and Burman, J.P (1946): The design of optimum multifactorial experiments. Biometrica, 33, 305-325

30. Reliability Meets Big Data: Opportunities and Challenges, William Q. Meeker- Department of Statistica- Center for Nondestructive Evaluation-Iowa State

University-Ames IA 50011, Yili Hong- Department of Statistics- Virginia Tech-Blacksburg VA 24061

31. Shen, Q., Jiang, J. H., Jiao, C. X. Shen, G. L., Yu, R.Q . Modified particle swarm optimization algorithm for variable selection in MLR and PLS modeling: QSAR studies of antagonism of angiotensin II antagonists. European Journal of Pharmaceutical Sciences, 22 (2004), 145-152

32. Statistical inference in massive data sets, Runze Li, Dennis K. J. Lin and Bing Li

33. The Variable Selection Problem, Edward I. George, University of Texas at Austin, September 2000

34. Tibshirani, R. Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society B, 58 (1996), 267-28

35. Todeschini, R., Consonni, V. Molecular descriptors for chemoinformatics, second edition. Wiley –VCH, 2009

36. Variable Selection for Classification and Regression in Large p, Small n Problems, Wei – Yin Loh

37. Variable selection methods: an introduction, Matteo Cassotti and Francesca Grisoni, Milano Chemometrics and QSAR Research Group – Dept. of Enviromental Sciences, University of Milano – Bicocca, P.za dellaScienza 1 – 20126 Milano (Italy)

38. Zou and Hastie, T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society B, 67 (2005), 301-320