ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΝΙΚΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΗΛΕΚΤΡΟΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

Ελαχιστοποίηση χαρακτηριστικών ταξινομητή για γονιδιακή σύνθεση

ΑΚΑΔΗΜΑΪΚΟ ΕΤΟΣ : 2008 – 2009

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ
ΤΟΥ ΦΟΙΤΗΤΗ

ΜΥΡΙΛΟΥ ΘΕΟΔΩΡΟΥ
ΑΜ:2000030057

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ

κ. Ζερβάκης Μιχάλης (επιβλέπων) Καθηγητής Πολυτεχνείου Κρήτης

κ. Μπάλας Κωνσταντίνος Αναπληρωτής Καθηγητής Πολυτεχνείου Κρήτης

κ. Καρυστινός Γεώργιος Επίκουρος Καθηγητής Πολυτεχνείου Κρήτης

ΠΡΟΛΟΓΟΣ

Η παρούσα διπλωματική εργασία με τίτλο «Ελαχιστοποίηση χαρακτηριστικών ταξινομητή για γονιδιακή σύνθεση», αφορά στην προσπάθεια εύρεσης του ελάχιστου αριθμού γονιδίων τα οποία δίνουν χρήσιμη πληροφορία για την παθολογία του καρκίνου του μαστού και κατ` επέκταση στην πιθανή εύρεση μιας μοναδικής γονιδιακής υπογραφής. Η διπλωματική εργασία ανατέθηκε στον τελειόφοιτο Μυρίλο Θεόδωρο από τον Καθηγητή του Τμήματος Ηλεκτρονικών Μηχανικών και Μηχανικών Υπολογιστών του Πολυτεχνείου Κρήτης, κύριο Ζερβάκη Μιχάλη.

Στο σημείο αυτό θα ήθελα να ευχαριστήσω θερμά τον καθηγητή μου, κύριο Ζερβάκη Μιχάλη για την συνολική του καθοδήγηση και στήριξη κατά τη διάρκεια εκπόνησης της διπλωματικής μου εργασίας.

Επίσης θα ήθελα να ευχαριστήσω το Διδάκτορα κύριο Μπλαζαντωνάκη Μιχάλη, που με ενδιαφέρον απάντησε σε όλες μου τις ερωτήσεις και με βοήθησε στις δυσκολίες που αντιμετώπισα κατά τη διάρκεια εκπόνησης της εργασίας.

Επιπλέον ευχαριστώ τους καθηγητές του τμήματος ΗΜΜΥ κύριους Καρυστινό Γεώργιο και Κωνσταντίνο Μπάλα για το χρόνο που αφιέρωσαν συμμετέχοντας στην επιτροπή αξιολόγησης της εργασίας.

Τέλος, επειδή με την εργασία αυτή ολοκληρώνονται οι προπτυχιακές μου σπουδές, θα ήθελα να ευχαριστήσω τους γονείς μου, τους φίλους μου και όσους άλλους μου συμπαραστάθηκαν και με υποστήριξαν σε όλες μου τις αποφάσεις όλα αυτά τα χρόνια.

Μυρίλος Θεόδωρος

Χανιά, Ιούνιος 2009

ΠΕΡΙΛΗΨΗ

Η εισαγωγή των μικροσυστοιχιών DNA στις επιστήμες της μοριακής βιολογίας και της ιατρικής έχει βοηθήσει τους επιστήμονες στην κατανόηση της έκφρασης των γονιδίων σε διάφορες περιπτώσεις καρκίνου, ενώ έχει ανοίξει νέους ορίζοντες στην εύρεση θεραπειών και φαρμάκων για την αντιμετώπισή του. Τα πρότυπα γονιδιακής έκφρασης που προκύπτουν για κάθε καρκίνο, δηλώνουν ποια γονίδια υπερεκφράζονται και ποια υποεκφράζονται, ξεχωρίζοντάς τον από άλλα είδη καρκίνου και δημιουργώντας την προοπτική καλύτερης διάγνωσης και πρόγνωσης.

Τα δεδομένα που προκύπτουν από τα πειράματα μικροσυστοιχιών DNA είναι υπεράριθμα σχετικά με το πλήθος των ασθενών που εξετάζονται και για το λόγο αυτό αναπτύσσονται συνεχώς μέθοδοι επιλογής γονιδίων, ούτως ώστε να καταλήξουμε σε κάποια γονιδιακή υπογραφή που να δίνει σημαντική πληροφορία για κάθε περίπτωση καρκίνου. Έτσι δημιουργούνται αποτελεσματικότερα μοντέλα ταξινόμησης.

Σκοπός αυτής της διπλωματικής είναι η εξαγωγή μιας γονιδιακής υπογραφής με ελάχιστο αριθμό γονιδίων τα οποία θα μας δίνουν χρήσιμη πληροφορία, για τον καρκίνο του μαστού . Βασίζεται στον αλγόριθμο RFE-LNW(Recursive Feature Elimination based on Linear Neuron Weights) και μελετά την προεπιλογή γονιδίων για την καλύτερη απόδοση του αλγορίθμου.

Τα γονίδια που εξάγει ο RFE-LNW ως τα πιο σημαντικά εμείς τα αφαιρούμε από τα train και test sets και ξανατρέχουμε τον αλγόριθμο. Αυτό το κάνουμε μέχρι να αφαιρεθούν όλα τα γονίδια ως σημαντικά. Κατόπιν δημιουργούμε νέα train και test sets από τα γονίδια που είχαμε αφαιρέσει κρατώντας μόνο αυτά που έδωσαν από κάποιο ποσοστό επιτυχίας και πάνω. Επαναλαμβάνουμε όλη αυτή τη διαδικασία και όταν φτάσουμε σε σημείο που κανένα γονίδιο δε δίνει κάτω από το ποσοστό επιτυχίας που έχουμε ορίσει, αυξάνουμε το ποσοστό επιτυχίας βάσει του οποίου κρατάμε γονίδια.

Το αποτέλεσμα αυτής της διπλωματικής εργασίας είναι η εύρεση ενός πολύ μικρού σετ γονιδίων βασιζόμενοι στα οποία μπορούμε να κρίνουμε με πολύ μεγάλο ποσοστό επιτυχίας αν κάποια ασθενής πρόκειται στα επόμενα 5 χρόνια να παρουσιάσει μετάσταση ή όχι.

ΠΕΡΙΕΧΟΜΕΝΑ

ΕΥΡΕΤΗΡΙΟ ΣΧΗΜΑΤΩΝ

ΕΥΡΕΤΗΡΙΟ ΠΙΝΑΚΩΝ

ΕΥΡΕΤΗΡΙΟ ΔΙΑΓΡΑΜΜΑΤΩΝ

Κεφάλαιο 1

Εισαγωγή

Μία από τις σημαντικότερες ασθένειες – αν όχι η σημαντικότερη – που πλήττουν την ανθρωπότητα στις μέρες μας είναι ο καρκίνος. Πρόκειται κυρίως για μια ασθένεια των πιο τελευταίων ετών και είναι μια από τις κυριότερες αιτίες θανάτου στις αναπτυγμένες χώρες. Το 13% όλων των θανάτων προκαλείται από καρκίνο [1], ενώ σύμφωνα με την Αμερικανική Αντικαρκινική Εταιρεία 7.6 εκατομμύρια άνθρωποι στον κόσμο πέθαναν από καρκίνο κατά τη διάρκεια του 2007 [2].

Έχει σημειωθεί σημαντική πρόοδος όσον αφορά την καταπολέμηση αυτής της μάστιγας – η ιατρική είναι σήμερα σε θέση να θεραπεύσει έναν καρκίνο στους δύο. Παρόλα αυτά, ο καρκίνος εξακολουθεί να αποτελεί πρόβλημα για τη δημόσια υγεία. Ορισμένοι καρκίνοι είναι ιάσιμοι, ή οι προοπτικές αποκατάστασης της υγείας αυξάνονται κατά πολύ, εάν υπάρξει έγκαιρη διάγνωση.

Σχεδόν όλοι οι καρκίνοι προκαλούνται από ανωμαλίες στο γενετικό υλικό των κυττάρων [3]. Επομένως, η πληροφορία που πηγάζει από τα γονίδια και συγκεκριμένα από την έκφρασή τους κατά την αναπαραγωγή των κυττάρων είναι πολύ σημαντική για την διάγνωση του καρκίνου σε κάποιον ασθενή, καθώς και για τον εντοπισμό διαφορών ανάμεσα σε διαφορετικά είδη καρκίνου. Επιπλέον, μπορεί να αποτελέσει χρήσιμο εφόδιο για την εύρεση νέων, πιο αποτελεσματικών μεθόδων θεραπείας.

Μέχρι το 1990, οι επιστήμονες μπορούσαν να μελετήσουν λίγα μόνο γονίδια κάθε φορά. Τώρα έχει αναπτυχθεί ένα νέο εργαλείο, που ονομάζεται μικροσυστοιχία DNA, η οποία είναι γνωστή ως DNA τσιπ και υπόσχεται να μεταφέρει την επιστήμη της κατανόησης γονιδίων σε νέο επίπεδο, με την ταυτόχρονη ανάλυση της έκφρασης χιλιάδων γονιδίων γρήγορα και αποτελεσματικά [4].

Στην παρούσα διπλωματική χρησιμοποιούνται δεδομένα που προκύπτουν από ανάλυση μικροσυστοιχιών DNA και που αφορούν τον καρκίνο του μαστού. Αποτελούνται από ένα πλήθος δειγμάτων – ασθενών με τις αντίστοιχες γονιδιακές τους εκφράσεις. Μας ενδιαφέρει η κατάταξη των ασθενών με βάση τις γονιδιακές τους εκφράσεις σε δύο κατηγορίες. Έτσι στην περίπτωση του καρκίνου του μαστού μας ενδιαφέρει η κατάταξή τους με βάση το γεγονός αν πρόκειται να εμφανιστεί μετάσταση μέσα στα επόμενα 5 χρόνια ή όχι. Για το σκοπό αυτό γίνεται χρήση μεθόδων αναγνώρισης προτύπων, όπου τα πρότυπα είναι οι ασθενείς και τα χαρακτηριστικά οι γονιδιακές τους εκφράσεις. Επομένως, τα δεδομένα μας χωρίζονται σε δύο υποσύνολα, ένα σύνολο δειγμάτων εκπαίδευσης (train set) για την εκπαίδευση του ταξινομητή και ένα ανεξάρτητο σύνολο δειγμάτων δοκιμής (independent test set) για τον έλεγχο της απόδοσής του.

Τα δεδομένα που προκύπτουν μέσω των πειραμάτων των μικροσυστοιχιών DNA είναι υπεράριθμα σχετικά με τον αριθμό των ασθενών που εξετάζονται. Έτσι για κάθε ασθενή που πάσχει από καρκίνο του μαστού σε κάθε ασθενή αντιστοιχούν 24188 γονίδια. Το γεγονός αυτό δυσκολεύει την ανάλυση δεδομένων τέτοιου τύπου από ήδη υπάρχουσες μεθόδους ταξινόμησης [5]. Επιπλέον, μεγάλο ποσοστό αυτών των δεδομένων που προκύπτουν από Microarray πειράματα είναι άσχετα και περιττά. Αυτά συνεισφέρουν στην ύπαρξη μη αξιόπιστων και χαμηλής απόδοσης αποτελεσμάτων.

Ανάμεσα στα χιλιάδες γονίδια υπάρχει περίπτωση να ανακαλύψουμε κάποια, των οποίων η έκφραση να είναι καθοριστική για την ταξινόμηση ενός ασθενή σε κάποια κατηγορία, δηλαδή να εντοπίσουμε πιο πληροφοριακά γονίδια, απορρίπτοντας ταυτόχρονα γονίδια που δεν περιέχουν κάποια σημαντική πληροφορία για το αποτέλεσμα. Η διαδικασία εύρεσης ενός βέλτιστου συνόλου γονιδίων είναι γνωστή ως επιλογή γονιδίων. Ο εντοπισμός ενός τέτοιου συνόλου είναι σημαντικός όσον αφορά το μοντέλο ταξινόμησης των ασθενών: με επιλογή γονιδίων καταλήγουμε σε αποδοτικότερα μοντέλα ταξινόμησης.

Μέχρι τώρα έχουν υλοποιηθεί και δοκιμαστεί πολλές μέθοδοι επιλογής γονιδίων, με σημαντικά αποτελέσματα. Σημαντική είναι η έρευνα που έχει γίνει γύρω από αυτό το θέμα και στόχος είναι η πρόβλεψη του κλινικού αποτελέσματος με όσο το δυνατόν λιγότερα γονίδια και με την μέγιστη δυνατή ακρίβεια. Πολλές έρευνες έχουν δείξει ότι μειώνοντας σημαντικά τον αριθμό των γονιδίων, η απόδοση της ταξινόμησης μπορεί να βελτιωθεί. Αυτό οδηγεί στη μείωση του κόστους του πειράματος, χωρίς να διακυβεύουμε την αξιοπιστία των αποτελεσμάτων.

Από τη σκοπιά της αναγνώρισης προτύπων θα μπορούσε κανείς να πει ότι το πρόβλημα της μελέτης των χιλιάδων γονιδίων είναι ένα κλασσικό παράδειγμα επιλογής χαρακτηριστικών. Όμως δεν είναι τόσο απλό, αφού σε τέτοιες περιπτώσεις έχουμε να αντιμετωπίσουμε τη διαφορά του πλήθους των γονιδίων σε σχέση με αυτό των δειγμάτων. Έχουμε στη διάθεσή μας έναν αριθμό γονιδίων της τάξης των χιλιάδων, ενώ έναν αριθμό δειγμάτων(ασθενών) της τάξης των μερικών δεκάδων μόνο. Αυτό το πρόβλημα είναι γνωστό ως Curse of Dimensionality. Οι χρησιμοποιούμενες μέθοδοι ταξινόμησης αποδίδουν σαφώς καλύτερα αν τα δείγματα(ασθενείς) έχουν πλήθος συγκρίσιμο με αυτό των χαρακτηριστικών(γονίδια), άρα κρίνεται επιτακτικό να καταφέρουμε να μειώσουμε τον αριθμό των γονιδίων χωρίς να χάσουμε χρήσιμη πληροφορία.

Σκοπός αυτής της εργασίας είναι επιλογή γονιδίων, τα οποία μας δίνουν χρήσιμη πληροφορία ώστε να καταλήξουμε σε ένα μικρότερο υποσύνολο των αρχικών γονιδίων. Παρόμοιες προσπάθειες έχουν γίνει και σε άλλες μελέτες. Για παράδειγμα ο Van't Veer και οι συνεργάτες του [6], στα δεδομένα του καρκίνου του μαστού που περιέχουν 24481 γονίδια, ως αρχικό βήμα επιλέγουν από αυτά τα 5000 που εμφανίζουν τα σημαντικότερα επίπεδα έκφρασης (υπερεκφράζονται ή υποεκφράζονται σημαντικά) σε σχέση με άλλα γονίδια. Οι Kim και Hamasaki [7] για τα δεδομένα της λευχαιμίας, του μελανώματος και του καρκίνου του πνεύμονα, επιλέγουν αρχικά εκείνα τα γονίδια των οποίων η έκφραση εμφανίζει την μεγαλύτερη διασπορά στο σύνολο των ασθενών, ισχυριζόμενοι ότι τα γονίδια που δεν έχουν μεγάλη διασπορά δεν είναι χρήσιμα για θέματα ταξινόμησης. Έτσι για τη λευχαιμία και

το μελάνωμα καταλήγουν σε 100 γονίδια (από 7000 και 3500 αντίστοιχα), ενώ για τον καρκίνο του πνεύμονα σε 200 γονίδια (από 12000).
Άλλο ένα παράδειγμα επιλογής δειγμάτων είναι αυτό του SVM-RFE δύο σταδίων, όπου στο πρώτο στάδιο γίνεται η προεπεξεργασία [8] (δεδομένα λευχαιμίας, καρκίνου παχέως εντέρου και λεμφώματος). Η SVM-RFE μέθοδος είναι μια επαναληπτική διαδικασία, σύμφωνα με την οποία τα γονίδια μειώνονται σταδιακά κατά κάποιο αριθμό *f*. Στο στάδιο της προεργασίας η SVM-RFE μέθοδος εκτελείται παραπάνω από μία φορά, κάθε φορά χρησιμοποιώντας διαφορετικό *f* και στο τέλος λαμβάνεται η ένωση των συνόλων που προκύπτουν για κάθε διαφορετική εκτέλεση του αλγορίθμου.

Στόχος μας είναι η εξαγωγή μιας γονιδιακής υπογραφής με όσο το δυνατόν λιγότερα γονίδια (αυτά που δίνουν την πιο χρήσιμη πληροφορία), βάσει της οποίας να μπορούμε με ψηλό ποσοστό επιτυχίας να κρίνουμε αν η ασθενής πρόκειται να εμφανίσει μετάσταση καρκίνου ή όχι.

Η προσπάθειά μας στηρίζεται πάνω στη μεθοδολογία που ακολουθήθηκε στο άρθρο <<Wrapper filtering criteria via linear neuron and kernel approaches>>. Ο ταξινομητής RFE-LNW, στον οποίο βασίστηκε το παραπάνω paper, εκπαιδεύεται βάσει ενός train set και στη συνέχεια ελέγχεται βάσει ενός test set. Βρίσκει ποια είναι τα πιο σημαντικά γονίδια, δηλαδή ποια δίνουν την πιο χρήσιμη πληροφορία. Στη συνέχεια αυτά εμείς τα αφαιρούμε από τα train και test set. Έτσι δημιουργούμε νέα train και test sets από τα παλιά σύνολα γονιδίων, τα οποία περιέχουν μόνο τα εναπομείναντα γονίδια, και επαναλαμβάνουμε τη διαδικασία με μειωμένο αριθμό γονιδίων έως ότου να έχουμε αφαιρέσει όλα τα γονίδια ως σημαντικά. Κατόπιν από τα γονίδια που είχαμε αφαιρέσει κρατάμε αυτά που μας έδιναν καλό ποσοστό επιτυχίας και δημιουργούμε πάλι νέα train και test sets και επαναλαμβάνουμε αυτή τη διαδικασία. Όταν φτάσουμε σε σημείο όπου κανένα γονίδιο δε μας δίνει κάτω από το ποσοστό επιτυχίας που είχαμε ορίσει, το αυξάνουμε και επαναλαμβάνουμε και επαναλαμβάνουμε την όλη διαδικασία.

Έτσι καταφέρνουμε από τον πολύ μεγάλο αριθμό γονιδίων που είχαμε αρχικά να καταλήξουμε σε αυτά που δίνουν την πιο χρήσιμη πληροφορία για τον

καρκίνο του παστού, και τα οποία είναι πολύ λίγα σε σχέση με τον αρχικό αριθμό.

Στο Κεφάλαιο 2 γίνεται μια εισαγωγή στην ασθένεια του καρκίνου, πως αυτή σχετίζεται με την γονιδιακή έκφραση, καθώς και μια περιεκτική περιγραφή του τρόπου με τον οποίο εξάγονται οι γονιδιακές εκφράσεις μέσω των πειραμάτων μικροσυστοιχιών DNA. Στο Κεφάλαιο 3 παρουσιάζεται το θεωρητικό υπόβαθρο πάνω στο οποίο βασιζόμαστε για την υλοποίησή μας. Γίνεται κάποια εισαγωγή στον τομέα της αναγνώρισης προτύπων και περιγράφεται ο ταξινομητής που χρησιμοποιείται. Στο Κεφάλαιο 4 αναλύεται η μεθοδολογία που ακολουθείται για την υλοποίηση του στόχου μας ενώ στο Κεφάλαιο 5 παρουσιάζονται τα αποτελέσματα που προκύπτουν. Τέλος στο Κεφάλαιο 6 επισημαίνονται τα συμπεράσματα στα οποία καταλήγουμε καθώς και η μελλοντική δουλειά η οποία δύναται να υλοποιηθεί στο μέλλον.

Κεφάλαιο 2

Μικροσυστοιχίες DNA – Συσχέτιση με την πάθηση του καρκίνου

Το αντικείμενο μελέτης αυτής της διπλωματικής εργασίας βασίζεται στην εξαγωγή συμπερασμάτων για τον καρκίνο, στηριζόμενη σε πληροφορία που μπορεί να αντληθεί από την έκφραση των γονιδίων. Ακολουθεί λοιπόν μια αναφορά στο τι είναι γονιδιακή έκφραση, πως αυτή σχετίζεται με την ασθένεια του καρκίνου ενώ τελικά περιγράφεται ο τρόπος με τον οποίο γίνεται ένα πείραμα μικροσυστοιχιών και παρουσιάζεται το πεδίο έρευνας της παρούσας διπλωματικής εργασίας.

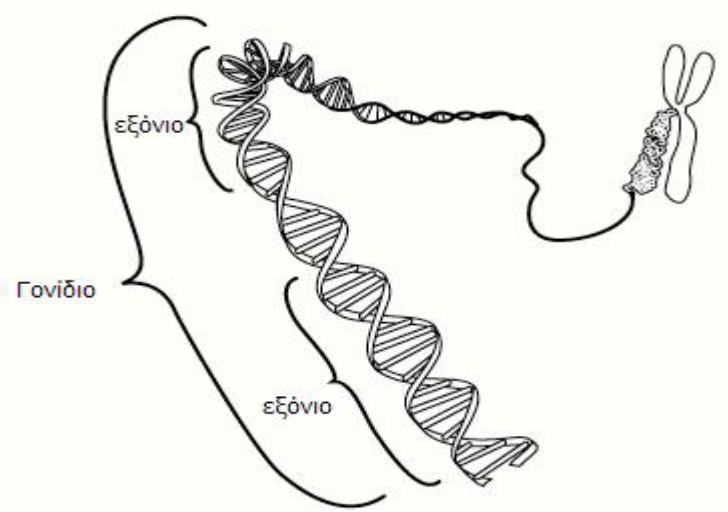
2.1 Γενετική έρευνα για τον καρκίνο

2.1.1 Γονιδιακή έκφραση

Το κύτταρο είναι κατά την βιολογία η βασική δομική και λειτουργική μονάδα που εκδηλώνει το φαινόμενο της ζωής. Ως κύτταρο νοείται το μικρότερο δομικό συστατικό της έμβιας ύλης το οποίο διαθέτει μορφολογική, φυσική και χημική οργάνωση καθώς και την ικανότητα της αφομοίωσης, της ανάπτυξης και της αναπαραγωγής [9].

Ο πυρήνας είναι συνήθως το μεγαλύτερο οργανίδιο ενός κυττάρου. Ελέγχει τις κυτταρικές δραστηριότητες και περιέχει τις γενετικές πληροφορίες που επιτρέπουν την αναπαραγωγή του κυττάρου. Μέσα στον πυρήνα του κάθε κυττάρου υπάρχουν 23 ζευγάρια δομών που είναι όμοια μεταξύ τους ανά ζεύγος και λέγονται χρωμοσώματα. Αυτά αποτελούνται από το DNA μέσα στο οποίο είναι γραμμένα με τάξη όλα τα χαρακτηριστικά ενός ανθρώπου [10]. Κάθε περιοχή του DNA που αναφέρεται σε ένα συγκεκριμένο χαρακτηριστικό ονομάζεται γονίδιο π.χ. γονίδιο που θα καθορίσει το χρώμα των μαλλιών, γονίδιο ύψους, γονίδια ευπάθειας σε ασθένειες κλπ. Υπάρχουν εκατομμύρια γονίδια σε κάθε κύτταρο. Το DNA βρίσκεται μέσα στον πυρήνα του κάθε

κυττάρου με τη μορφή διπλής αλυσίδας, είναι μπλεγμένο σαν κουβάρι και μορφοποιείται σε χρωμοσώματα μόνο στη φάση που το κύτταρο πολλαπλασιάζεται. Στο Σχήμα 1 φαίνεται η σχέση ενός γονιδίου με τη διπλή έλικα του DNA και το χρωμόσωμα.



Σχήμα 1. Απεικόνιση ενός γονιδίου σε σχέση με τη διπλή έλικα του DNA και ένα χρωμόσωμα (δεξιά).

Στη Γενετική και ειδικότερα στη Γονιδιολογία με τον όρο γονιδιακή έκφραση ή έκφραση γονιδίων (gene expression), χαρακτηρίζεται η διαδικασία εκείνη που προκαλεί τη μεταφορά κωδικοποιημένων πληροφοριών του γονιδίου στο λειτουργικό προϊόν του (γονιδιακό προϊόν), το οποίο μπορεί να είναι είτε RNA είτε πρωτεΐνη [11]. Πιο συγκεκριμένα, πρόκειται για την μεταγραφή της πληροφορίας που περιέχεται στο DNA, την αποθήκη της γενετικής πληροφορίας, σε μόρια αγγελιοφόρου RNA (mRNA) τα οποία στη συνέχεια μεταφράζονται σε πρωτεΐνες, που φέρουν εις πέρας τις σημαντικές λειτουργίες του κυττάρου.

Τα περισσότερα γονίδια περιέχουν κάποιες περιοχές που δεν κωδικοποιούν γονιδιακά προϊόντα, αλλά συχνά ρυθμίζουν τη γονιδιακή έκφραση. Οι περιοχές που κωδικοποιούν πραγματικά το προϊόν των γονιδίων είναι γνωστές ως εξόνια [12].

Εκτός από λίγες εξαιρέσεις, κάθε κύτταρο στο σώμα περιέχει ένα ολόκληρο σετ από χρωμοσώματα και παρόμοια γονίδια. Όμως, μόνο ένα ποσοστό αυτών των γονιδίων ενεργοποιείται και αυτό το ποσοστό που «εκφράζεται» δίνει μοναδικές ιδιότητες σε κάθε κύτταρο [13] . Τα γονίδια περιέχουν τις οδηγίες για τη δημιουργία mRNA, αλλά σε οποιαδήποτε στιγμή κάθε κύτταρο παράγει mRNA μόνο από το ποσοστό των γονιδίων που εκφράζονται. Εάν ένα γονίδιο χρησιμοποιείται για την παραγωγή mRNA, δηλαδή εκφράζεται, αυτό σημαίνει πως πρόκειται για ενεργό γονίδιο, αλλιώς θεωρείται ανενεργό.

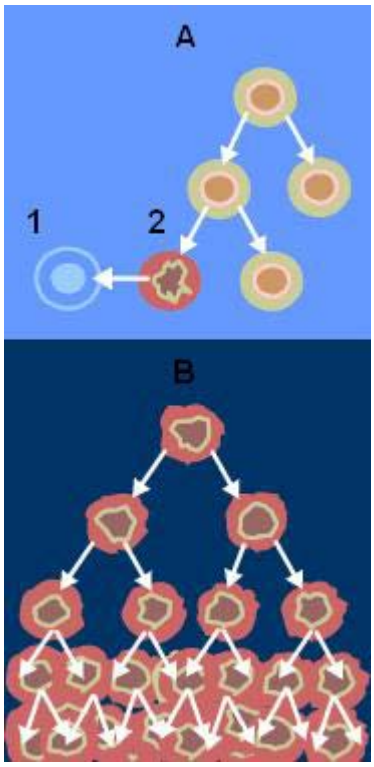
Πολλοί είναι εκείνοι οι παράγοντες που καθορίζουν εάν ένα γονίδιο θα είναι ενεργό ή ανενεργό, όπως η ώρα της ημέρας, το αν το κύτταρο διαιρείται, το περιβάλλον του καθώς και χημικά σήματα από άλλα κύτταρα [14]. Εκείνο το ποσοστό των γονιδίων που εκφράζεται δίνει και μοναδικές ιδιότητες σε κάθε κύτταρο. Σε διαφορετικού τύπου κύτταρα όπως είναι τα επιδερμικά, τα νευρικά ή τα ηπατικά, εκφράζονται διαφορετικά γονίδια και αυτό είναι κατά κύριο λόγο η αιτία διαφοροποίησής τους. Επομένως η γνώση του ποια γονίδια είναι ενεργά επιτρέπει στο να καθοριστεί ο τύπος ενός κυττάρου, η κατάστασή του, το περιβάλλον του και ούτω καθ’ εξής.

Κατά συνέπεια, το είδος και η ποσότητα των mRNAs που παράγονται από ένα κύτταρο αποτελεί αντικείμενο επιστημονικής μελέτης, ούτως ώστε να γίνει γνωστό ποια γονίδια εκφράζονται και κατ’ επέκταση πώς το κύτταρο αντιδρά στα διάφορα περιβαλλοντικά ερεθίσματα και στις δικές του μεταβαλλόμενες ανάγκες.

2.1.2 Καρκίνος και γονιδιακή έκφραση

Ο καρκίνος είναι μια κατηγορία ασθενειών ή διαταραχών στην οποία μια ομάδα κυττάρων επιδεικνύει ανεξέλεγκτη αύξηση (διαίρεση πέρα από τα κανονικά όρια), εισβολή (παρείσδυση και καταστροφή των παρακείμενων ιστών), και μερικές φορές μετάσταση (διάδοση σε άλλες θέσεις στο σώμα) [3]. Αυτές οι τρεις κακοήθεις ιδιότητες των καρκίνων τους διαφοροποιούν από τους καλοήθεις όγκους, οι οποίοι δεν εισβάλλουν ή δεν αναπαράγονται με μεταστάσεις.

Η ανεξέλεγκτη διαίρεση των κυττάρων προκαλείται από ζημιά στο DNA, με συνέπεια τις μεταλλάξεις στα γονίδια που ελέγχουν την κυτταροδιαίρεση. Όταν τα κανονικά κύτταρα βλάπτονται, πέρα από την επισκευή, αποβάλλονται μέσω της απόπτωσης. Τα κύτταρα καρκίνου αποφεύγουν την απόπτωση και συνεχίζουν να πολλαπλασιάζουν κατά τρόπο ανεξέλεγκτο (Σχήμα 2).
Ο καρκίνος μπορεί να προσβάλλει ανθρώπους όλων των ηλικιών, ακόμη και έμβρυα, αλλά ο κίνδυνος για τις περισσότερες περιπτώσεις αυξάνεται με την ηλικία. Εάν δεν θεραπευθεί, μπορεί τελικά να προκαλέσει το θάνατο.



Σχήμα 2. Κανονική διαίρεση κυττάρων (A) και καρκινική διαίρεση κυττάρων (B), με το 1 συμβολίζεται η απόπτωση και με το 2 το κατεστραμμένο κύτταρο.

Ο καρκίνος μπορεί να προκαλέσει πολλά διαφορετικά συμπτώματα, ανάλογα με την περιοχή και το χαρακτήρα της κακοήθειας και εάν υπάρχει μετάσταση. Η ύπαρξη καρκίνου μπορεί να υποψιαστεί για ποικίλους λόγους, αλλά η οριστική διάγνωση απαιτεί συνήθως την ιστολογική εξέταση του ιστού από έναν παθολόγο. Οι ιστοί είναι μεγάλες ομάδες ομοειδών κυττάρων, κατά σύσταση και ορισμένη φυσιολογική λειτουργία (π.χ. μυϊκός ιστός) και αποτελούν την μονάδα δεύτερης τάξης στον ανθρώπινο οργανισμό, μετά τα κύτταρα.

Ο καρκίνος γενικά ταξινομείται σε διαφορετικές κατηγορίες με δύο τρόπους: σύμφωνα με τον τύπο του ιστού από τον οποίο προέρχονται τα καρκινικά κύτταρα (ιστολογικός τύπος) και σύμφωνα με το μέρος του σώματος στο οποίο εμφανίζεται (ή εμφανίστηκε αρχικά, σε περίπτωση που έχουμε να κάνουμε με μετάσταση) [15]. Από ιστολογική σκοπιά υπάρχουν εκατοντάδες διαφορετικοί καρκίνοι οι οποίοι ομαδοποιούνται σε πέντε μεγάλες κατηγορίες, το καρκίνωμα, το σάρκωμα, το λέμφωμα, το μυέλωμα και την λευχαιμία, ενώ υπάρχουν και καρκίνοι μεικτών τύπων.

Η πρόβλεψη μιας θεραπείας για ένα άτομο που πάσχει από καρκίνο δεν είναι πάντα ξεκάθαρη υπόθεση. Μέσω της βιοψίας γίνεται αφαίρεση τμήματος ενός όγκου ή ενός ύποπτου ιστού του σώματος, προκειμένου να γίνει εξέταση των κυττάρων υπό το πρίσμα ενός μικροσκοπίου. Τα καρκινικά κύτταρα ποικίλουν μορφολογικά, από το να είναι όμοια στην εμφάνιση με τα φυσιολογικά κύτταρα μέχρι και να έχουν πλήρη έλλειψη δομής και οργάνωσης [16]. Οι παθολόγοι χρησιμοποιούν την ιστοπαθολογική εξέταση με σκοπό να καθορίσουν την πρόγνωση ή την προοπτική πλήρους ανάκαμψης των ασθενών από την πάθηση του καρκίνου. Παρόλα αυτά, τα φαινόμενα απατούν: μία θεραπεία η οποία μπορεί να έχει αποτέλεσμα σε κάποιο είδος καρκίνου, είναι πιθανό να αποτύχει σε κάποιο άλλο είδος το οποίο είναι μορφολογικά όμοιο.

Για το λόγο αυτό, οι επιστήμονες στρέφονται σε ένα νέο τρόπο ταξινόμησης του καρκίνου, που δε στηρίζεται πλέον στη μορφολογία των κυττάρων, αλλά σε μοριακά χαρακτηριστικά. Βασίζεται στο γεγονός ότι η γονιδιακή έκφραση των όγκων μπορεί να προσφέρει πολύ περισσότερη πληροφορία από τη μορφολογική και κατά συνέπεια να οδηγήσει σε πιο αξιόπιστα συστήματα ταξινόμησης των όγκων.

Η μελέτη της γονιδιακής έκφρασης ενέχει την παρατήρηση των ποσοτήτων mRNA ή πρωτεϊνών που παράγονται από ένα κύτταρο μια δεδομένη στιγμή. Η τεχνολογία μικροσυστοιχιών DNA είναι μια νέα, πολλά υποσχόμενη τεχνολογία, η οποία καθορίζει ποια γονίδια μεταξύ χιλιάδων που εξετάζονται

είναι ενεργά και επιτρέπει στους ερευνητές την προετοιμασία μοναδικών προτύπων γονιδιακής έκφρασης (gene expression profiles) για διαφορετικούς τύπους κυττάρων.

Αυτή η τεχνολογία, η οποία αναλύεται στην επόμενη ενότητα, φέρνει λοιπόν την επανάσταση στην ταξινόμηση και διάγνωση του καρκίνου, καθώς και στην λήψη αποφάσεων για την θεραπεία που πρόκειται να υιοθετηθεί σε κάθε περίπτωση καρκίνου.

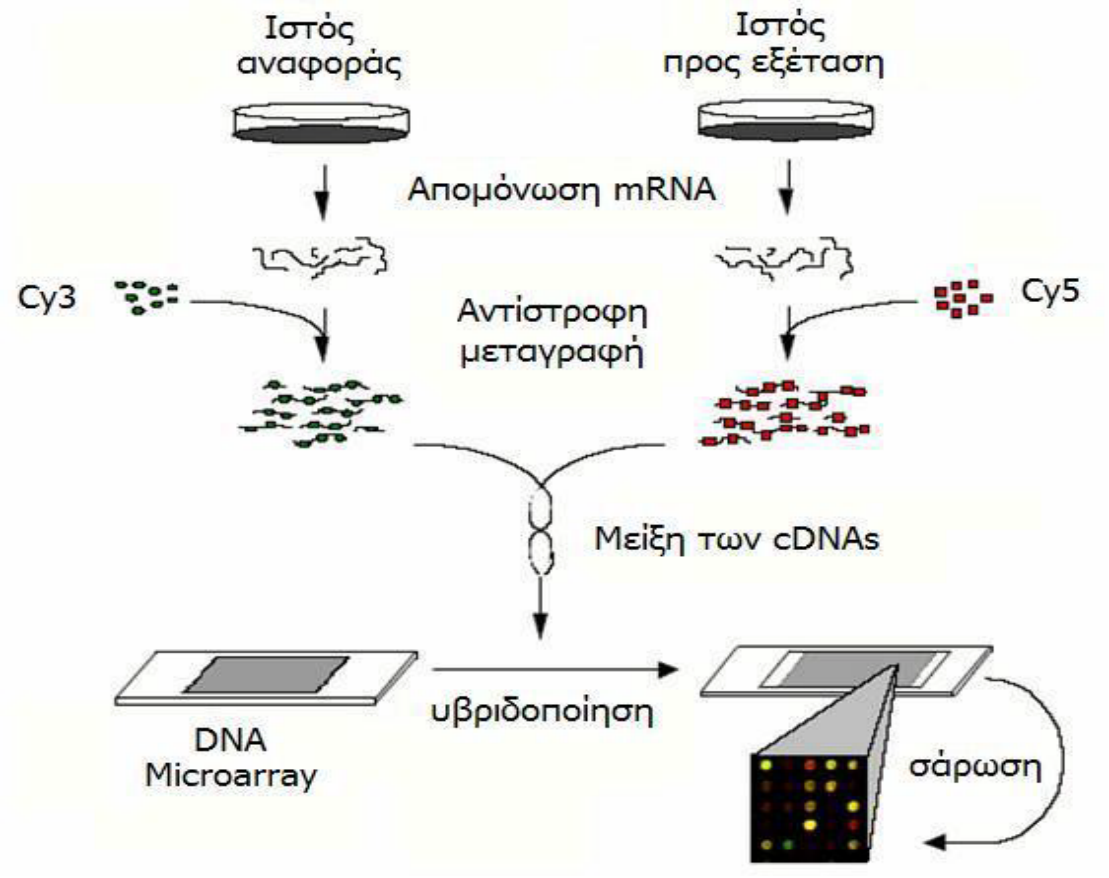
2.2 Τεχνολογία μικροσυστοιχιών DNA

Μία μικροσυστοιχία DNA μπορεί να θεωρηθεί ανάλογη ενός τσιπ υπολογιστή, όμως αντί να περιέχει ηλεκτρικά κυκλώματα, το τσιπ περιέχει χιλιάδες μικροσκοπικά κελιά [16]. Κατά τη δημιουργία μιας μικροσυστοιχίας, με τη βοήθεια της ρομποτικής τοποθετούνται μικροσκοπικές κηλίδες που περιέχουν τμήματα γονιδίων σε συγκεκριμένα σημεία (spots) πάνω σε μία γυάλινη αντικειμενοφόρο πλάκα, σαν αυτές που χρησιμοποιούνται στη μικροσκοπία. Τα τμήματα αυτά των γονιδίων αποτελούνται μόνο από τα εξόνια, τις περιοχές δηλαδή του γονιδίου που ουσιαστικά κωδικοποιούν την πρωτεΐνη και συντίθενται από συμπληρωματικό DNA (cDNA). Στην περίπτωση του καρκίνου, τα τμήματα γονιδίων προέρχονται από εκείνα τα γονίδια που έχουν προσδιοριστεί ως υπεύθυνα για την εμφάνιση του καρκίνου. Έτσι δημιουργείται ένα πλακίδιο μικροσυστοιχίας που περιέχει μια βιβλιοθήκη cDNA.

Γενικά στα πειράματα μικροσυστοιχιών DNA χρησιμοποιούνται δύο δείγματα mRNA, ένα δείγμα αναφοράς και ένα υπό εξέταση. Προκειμένου για παράδειγμα να διαγνωστεί ο καρκίνος, ο ύποπτος όγκος ή ιστός λαμβάνεται μέσω βιοψίας και εξάγεται το mRNA από τα κύτταρα, που αποτελεί το δείγμα υπό εξέταση. Το σκεπτικό αυτής της διαδικασίας είναι ότι, όπως έχουμε ήδη αναφέρει, κατά την έκφραση των γονιδίων γίνεται μεταγραφή του DNA σε mRNA. Το δείγμα αναφοράς θα προέρχεται από έναν αντίστοιχο φυσιολογικό ιστό.

Τα δύο δείγματα mRNA τότε μεταγράφονται αντίστροφα στις αντίστοιχες αλληλουχίες του πιο σταθερού, συμπληρωματικού DNA και μετά σημαίνονται χρησιμοποιώντας δύο διαφορετικές φθορίζουσες χρωστικές ουσίες (συνήθως μία χρωστική ουσία κόκκινου φθορισμού, Cy5 για το δείγμα υπό εξέταση και μια χρωστική ουσία πράσινου φθορισμού, Cy3 για το δείγμα αναφοράς) [17]. Στη συνέχεια υβριδοποιούνται ταυτόχρονα κάτω από αυστηρές συνθήκες σε ένα πλακίδιο μικροσυστοιχίας. Εάν το γονίδιο είναι ενεργό, τότε το φθορίζον cDNA προσαρτάται στο αντίστοιχο τμήμα γονιδίου και παράγει ένα φωτεινότερο χρώμα.

Στη συνέχεια γίνεται σάρωση του πλακιδίου μικροσυστοιχίας DNA με ομοεστιακό λέιζερ και η εικόνα που προκύπτει αναλύεται από υπολογιστή. Η σχετική έκφραση ενός γονιδίου στα δύο δείγματα αναλύεται με τη μέτρηση της αναλογίας των εντάσεων φθορισμού των δύο χρωστικών ουσιών σε μία συγκεκριμένη κηλίδα πάνω στο πλακίδιο μικροσυστοιχίας.



Σχήμα 3. Διαδικασία ενός πειράματος μικροσυστοιχίας DNA

Για κάθε γονίδιο, εάν στο δείγμα προς εξέταση η ποσότητα του mRNA που παράγεται από αυτό είναι μεγαλύτερη από εκείνη του δείγματος αναφοράς (το γονίδιο υπερεκφράζεται στο δείγμα προς εξέταση) τότε η κηλίδα θα φθορίζεται κόκκινη. Εάν το δείγμα αναφοράς παράγει περισσότερη ποσότητα mRNA από το δείγμα προς εξέταση (το γονίδιο υποεκφράζεται στο δείγμα προς εξέταση) τότε η κηλίδα θα φθορίζεται πράσινη. Τέλος εάν παράγεται ίση ποσότητα και στα δύο δείγματα η κηλίδα θα είναι κίτρινη. Έτσι δημιουργείται το πρότυπο γονιδιακής έκφρασης για το δείγμα της βιοψίας, το οποίο δίνει μια εικόνα του ποια γονίδια υπερεκφράζονται και ποια υποεκφράζονται στο δείγμα αυτό [18]. Στο Σχήμα 3 απεικονίζεται η συνολική διαδικασία ενός πειράματος μικροσυστοιχίας DNA που μόλις περιγράφηκε.

Όσον αφορά τον καρκίνο, οι μέχρι τώρα τεχνικές επιτρέπουν στους παθολόγους να κάνουν την διάγνωση, παρόλα αυτά πρόκειται για μία σύνθετη διαδικασία που απαιτεί την συμμετοχή πολλών εξειδικευμένων εργαστηρίων. Μέσω της τεχνολογίας microarray έχουμε μια μοναδική, πιο αποτελεσματική μέθοδο διάγνωσης. Επιπλέον η διάγνωση γίνεται βασισμένη στην πιο θεμελιώδη αιτία καρκίνου – την ελαττωματική γονιδιακή έκφραση. Επομένως πέρα από την βοήθεια που προσφέρει στον εντοπισμό κατάλληλης θεραπείας και φαρμάκων για κάθε καρκίνο, η τεχνολογία μικροσυστοιχιών DNA βοηθάει στην καλύτερη κατανόηση του μηχανισμού ανάπτυξης του καρκίνου.

2.3 Πεδίο έρευνας – Εξαγωγή δεδομένων μέσω μικροσυστοιχιών DNA

Στη συγκεκριμένη διπλωματική εργασία ασχολούμαστε με πειράματα που αφορούν τον καρκίνο του μαστού.

Ο καρκίνος του μαστού είναι ο συνηθέστερος καρκίνος στις γυναίκες και η συχνότητά εμφάνισής του αυξάνει σταθερά. Συμβαίνει όταν κάποια κύτταρα του μαστού χάνουν τον έλεγχο του πολλαπλασιασμού και διαιρούνται ανεξέλεγκτα [19]. Τα κύτταρα αυτά έχουν την δυνατότητα να διασπάσουν

κάποιους φυσιολογικούς ανατομικούς φραγμούς του μαστού και να διασπαρθούν στο υπόλοιπο σώμα.

Εξετάζονται ιστοί ασθενών με σκοπό να βρούμε τον ελάχιστο αριθμό γονιδίων τα οποία θα μας επιτρέψουν να κρίνουμε, με όσο μεγαλύτερη ακρίβεια μπορούμε μέσω του ταξινομητή που χρησιμοποιούμε αν πρόκειται μέσα στα επόμενα πέντε χρόνια να εμφανίσει η ασθενής μετάσταση ή όχι.

Κεφάλαιο 3

Αναγνώριση προτύπων – Βασικές έννοιες

Σε αυτό το κεφάλαιο παρουσιάζεται το θεωρητικό υπόβαθρο πάνω στο οποίο βασιζόμαστε για την υλοποίηση του προβλήματός μας. Δηλαδή θα γίνει ανάλυση κάποιων βασικών θεωρητικών στοιχείων που είναι απαραίτητα για την κατανόηση της υλοποίησης που θα ακολουθήσει σε επόμενο κεφάλαιο.

Η αναγνώριση προτύπων (pattern recognition) είναι ο επιστημονικός κλάδος του οποίου στόχος είναι η ταξινόμηση αντικειμένων σε διάφορες κατηγορίες (classes) [20]. Ανάλογα με την εφαρμογή, αυτά τα αντικείμενα μπορούν να είναι εικόνες ή κυματομορφές σημάτων ή οποιοσδήποτε τύπος μετρήσεων που πρέπει να ταξινομηθούν. Κάθε λοιπόν αντικείμενο προς ταξινόμηση αντιπροσωπεύεται από ένα διάνυσμα *N* συνιστωσών το οποίο περιέχει τα χαρακτηριστικά που περιγράφουν το συγκεκριμένο αντικείμενο. Σε αυτά τα αντικείμενα – διανύσματα αναφερόμαστε με τον γενικής χρήσης όρο «πρότυπα» (patterns). Η αναγνώριση προτύπων έχει μια μεγάλη ιστορία, αλλά πριν από την δεκαετία του ΄60 ήταν κυρίως το αποτέλεσμα θεωρητικής έρευνας στον τομέα της στατιστικής. Όπως με όλα τα άλλα, η εμφάνιση των υπολογιστών αύξησε τη ζήτηση για πρακτικές εφαρμογές της αναγνώρισης προτύπων, οι οποίες έθεσαν στη συνέχεια νέες απαιτήσεις για περαιτέρω θεωρητικές εξελίξεις. Η αναγνώριση προτύπων είναι πλέον ένα αναπόσπαστο τμήμα στα περισσότερα συστήματα τεχνητής νοημοσύνης που είναι υπεύθυνα για λήψη αποφάσεων ταξινόμησης.

Στην περίπτωσή μας τα αντικείμενα προς ταξινόμηση είναι οι ασθενείς. Μέσω ιστολογικής εξέτασης και περαιτέρω ανάλυσης με τεχνολογία μικροσυστοιχιών, λαμβάνονται τα χαρακτηριστικά τους, τα οποία είναι τα γονίδια με τις σχετικές τους εκφράσεις. Τελικά κάθε ασθενής αποτελεί ένα πρότυπο, που δεν είναι παρά ένα διάνυσμα που περιέχει τις γονιδιακές του εκφράσεις.

Ως χώρο χαρακτηριστικών (feature space) θεωρούμε έναν αφηρημένο χώρο, του οποίου η διάσταση καθορίζεται από τον αριθμό των χαρακτηριστικών που χρειάζονται για να περιγράψουν ένα πρότυπο. Μέσα σε αυτόν τον χώρο κάθε πρότυπο αναπαρίσταται ως σημείο (ή ως διάνυσμα) συναρτήσει των χαρακτηριστικών του. Έτσι εάν έχουμε ένα χαρακτηριστικό, το πρότυπο είναι ένα σημείο πάνω σε μία γραμμή, εάν έχουμε δύο χαρακτηριστικά είναι ένα σημείο στο επίπεδο, εάν έχουμε τρία χαρακτηριστικά το πρότυπο είναι ένα σημείο στον τρισδιάστατο χώρο και εάν έχουμε N χαρακτηριστικά είναι ένα σημείο στο N -διάστατο χώρο.

3.1 Ταξινόμηση προτύπων

Έχοντας κάνει κάποια εισαγωγή στην αναγνώριση προτύπων, θα προχωρήσουμε στην ανάλυση των θεμάτων ταξινόμησης των προτύπων.
Σε κάθε πρόβλημα ταξινόμησης έχουμε ένα δεδομένο αριθμό κλάσεων στις οποίες ταξινομούνται τα πρότυπα. Κάθε πρότυπο ταξινομείται ανάλογα με τα χαρακτηριστικά του σε μία συγκεκριμένη κλάση. Οπότε σε κάθε πρότυπο αντιστοιχεί μια ετικέτα (label) που δηλώνει την κλάση στην οποία αυτό ανήκει. Η δουλειά ενός ταξινομητή είναι η αντιστοίχιση ενός προτύπου σε μία ετικέτα. Επομένως εάν έχουμε *N* χαρακτηριστικά και *C* κλάσεις, ο ταξινομητής πραγματοποιεί την αντιστοίχιση από τον *N*-διάστατο χώρο χαρακτηριστικών σε ένα διακριτό σύνολο *C* ετικετών. Ο ρόλος δηλαδή ενός ταξινομητή είναι η διαίρεση του χώρου χαρακτηριστικών σε *C* διαφορετικά τμήματα, κάθε ένα από τα οποία αντιστοιχεί σε μια συγκεκριμένη κλάση. Για κάθε νέο πρότυπο προς ταξινόμηση, ο ταξινομητής ορίζει μία ετικέτα, το κατατάσσει δηλαδή σε έναν από τους *C* χώρους μέσα στο χώρο χαρακτηριστικών. Τα όρια τα οποία διαχωρίζουν τα *C* διαφορετικά τμήματα ονομάζονται όρια απόφασης (decision boundaries) [21]. Γενικά, σε περιοχές κοντά στα όρια απόφασης εμφανίζεται το υψηλότερο ποσοστό των λάθος ταξινομημένων αντικειμένων.

3.1.1 Εκπαίδευση ταξινομητών – Διαχωριστικές συναρτήσεις

Ένας ταξινομητής χρειάζεται εκπαίδευση η οποία μπορεί να είναι είτε με επίβλεψη (μάθηση με επίβλεψη, supervised learning) είτε χωρίς επίβλεψη

(μάθηση χωρίς επίβλεψη, unsupervised learning). Κατά την μάθηση με επίβλεψη ένα σύνολο από πρότυπα με προκαθορισμένες ετικέτες κλάσεων, το λεγόμενο σύνολο εκπαίδευσης (train set), χρησιμοποιείται για την εκπαίδευση του ταξινομητή. Σε αυτήν την περίπτωση εκμεταλλευόμαστε αυτήν την apriori πληροφορία για την δημιουργία του μοντέλου ταξινόμησης. Κατά την μάθηση χωρίς επίβλεψη έχουμε ένα σύνολο προτύπων χωρίς όμως προκαθορισμένες ετικέτες κλάσεων. Σε τέτοιες περιπτώσεις, ο στόχος μπορεί να είναι ο εντοπισμός ομοιοτήτων ανάμεσα στα χαρακτηριστικά ούτως ώστε να ομαδοποιήσουμε όμοια πρότυπα (clustering) ή ο καθορισμός της κατανομής των δεδομένων στον χώρο, γνωστό ως εκτίμηση πυκνότητας (density estimation) [22].

Ένα σύνολο εκπαίδευσης λοιπόν αποτελείται από ένα πλήθος *M* προτύπων x_i, *i=1,2,…,M,* σε κάθε ένα από τα οποία αντιστοιχεί μια ετικέτα κλάσης y_i. Συχνά όταν αναφερόμαστε στα πρότυπα θα χρησιμοποιούμε τις λέξεις δείγματα, αντικείμενα ή δεδομένα.

Μια διαχωριστική συνάρτηση (discriminant function) ή συνάρτηση απόφασης (decision function) *D(x)*, είναι μια συνάρτηση ενός προτύπου *x* και οδηγεί σε έναν κανόνα ταξινόμησης [21]. Έστω ότι έχουμε ταξινόμηση σε δύο κλάσεις ω_1 και ω_2, τότε μια διαχωριστική συνάρτηση *D(x)* είναι μια συνάρτηση για την οποία ισχύει:

$$D(x) > 0 \Rightarrow x \in \omega_1$$
$$D(x) < 0 \Rightarrow x \in \omega_2 \qquad (1)$$

Σε περίπτωση ισότητας (*D(x)* = 0) το πρότυπο *x* μπορεί να ταξινομηθεί αυθαίρετα σε οποιαδήποτε από τις δύο κλάσεις ω_1 και ω_2.

Σε περιπτώσεις προβλημάτων με περισσότερες από δύο κλάσεις δημιουργούνται περισσότερες συναρτήσεις απόφασης. Πιο συγκεκριμένα, για ένα σύνολο $\Omega=\{\omega_1,\omega_2,\ldots,\omega_C\}$ *C* κλάσεων, ορίζουμε *C* διαχωριστικές συναρτήσεις $g_i(x)$ τέτοιες ώστε

$$g_i(x) > g_j(x) \Rightarrow x \in \omega_1, \quad i,j = 1,2,\dots,C, i \neq j \tag{2}$$

Αυτό σημαίνει ότι το πρότυπο *x* ταξινομείται στην κλάση με την μεγαλύτερη διαχωριστική συνάρτηση. Φυσικά αν έχουμε δύο κλάσεις η διαχωριστική συνάρτηση της μορφής

$$D(x) = g_1(x) - g_2(x) \tag{3}$$

ισοδυναμεί με την περίπτωση δύο κλάσεων που δίνεται από την εξίσωση (1).

Η επιλογή της συνάρτησης απόφασης μπορεί να στηρίζεται σε προγενέστερη γνώση για τα πρότυπα που πρόκειται να ταξινομηθούν ή μπορεί να είναι μια συνάρτηση της οποίας οι παράμετροι να ρυθμίζονται από μία διαδικασία εκπαίδευσης. Υπάρχουν πολλές μορφές διαχωριστικών συναρτήσεων που διαφέρουν στην πολυπλοκότητα ξεκινώντας από την γραμμική διαχωριστική συνάρτηση (όπου η *D* είναι γραμμικός συνδυασμός των προτύπων εκπαίδευσης x_i) και φτάνοντας σε πολλών παραμέτρων μη γραμμικές συναρτήσεις.

Στην παρούσα διπλωματική εργασία για την εκπαίδευση των ταξινομητών χρησιμοποιείται μάθηση με επίβλεψη. Επίσης περιοριζόμαστε σε ταξινόμηση σε δύο κλάσεις. Ως είσοδο λοιπόν έχουμε ένα σύνολο εκπαίδευσης, σύμφωνα με το οποίο κατασκευάζεται η διαχωριστική συνάρτηση βάση της οποίας θα ληφθεί η απόφαση ταξινόμησης ενός προτύπου *x*.

3.1.2 Γραμμικά διαχωριστικές συναρτήσεις – Όρια απόφασης

Θεωρούμε και πάλι την περίπτωση ταξινόμησης σε δύο κλάσεις ω_1 και ω_2. Συναρτήσεις απόφασης που είναι γραμμικός συνδυασμός των *N* χαρακτηριστικών ενός προτύπου *x* είναι γνωστές ως γραμμικά διαχωριστικές συναρτήσεις (linear discriminant functions) [21]. Μια τέτοια συνάρτηση έχει τη μορφή:

$$D(x) = w^T x + w_0 = \sum_{i=1}^{N} w_i x_i + w_0 \quad (4)$$

Η παραπάνω σχέση είναι μια γραμμική διαχωριστική συνάρτηση της οποίας ο ακριβής προσδιορισμός γίνεται με καθορισμό του διανύσματος βαρών *w* και του κατωφλίου w_0. Η εξίσωση *D(x) = 0* είναι η εξίσωση ενός υπερεπιπέδου με μοναδιαίο κάθετο διάνυσμα στην κατεύθυνση του *w* και μια κάθετη απόσταση $|w_0|/|w|$ από την αρχή των συντεταγμένων. Επομένως το *w* δηλώνει την κατεύθυνση του υπερεπιπέδου και το w_0την θέση του στο *N*-διάστατο χώρο. Στην περίπτωση όπου w_0= 0, το υπερεπίπεδο περνάει από την αρχή των συντεταγμένων. Το υπερεπίπεδο λοιπόν που ορίζεται από τη σχέση *D(x) = 0,* αποτελεί και το όριο απόφασης. Το όριο απόφασης διαχωρίζει τα πρότυπα που ταξινομούνται στην κλάση ω_1 από αυτά που ταξινομούνται στην κλάση ω_2. Η τιμή της διαχωριστικής συνάρτησης για ένα πρότυπο *x* είναι και η κάθετη απόστασή του από το υπερεπίπεδο.

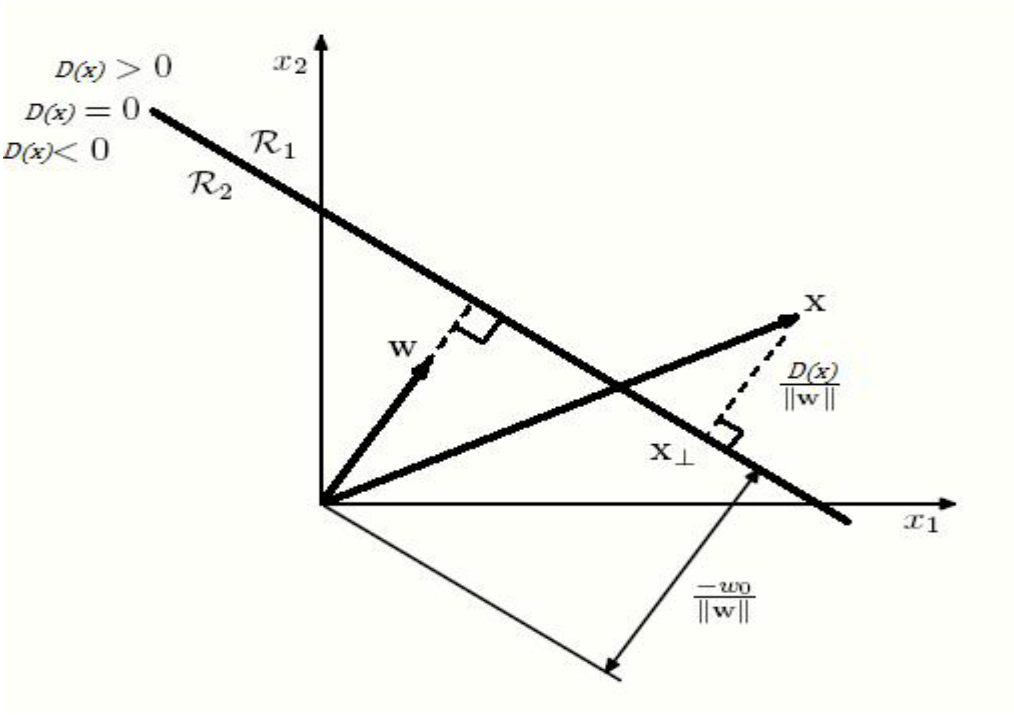
Εδώ χρειάζεται να επισημάνουμε ότι υπάρχουν περισσότερα του ενός υπερεπίπεδα τα οποία είναι ικανά να διαχωρίσουν τις κλάσεις μεταξύ τους. Μέσω της διαδικασίας της εκπαίδευσης εντοπίζεται και το βέλτιστο υπερεπίπεδο, σύμφωνα πάντα και με τον τρόπο λειτουργίας του κάθε ταξινομητή.

Για τις γραμμικές συναρτήσεις απόφασης, σε περίπτωση όπου έχουμε δύο χαρακτηριστικά, το όριο απόφασης είναι μια γραμμή, σε περίπτωση τριών χαρακτηριστικών είναι ένα επίπεδο και εάν έχουμε περισσότερα των τριών χαρακτηριστικά είναι ένα υπερεπίπεδο στον πολυδιάστατο χώρο.

Σε περίπτωση που έχουμε δύο πρότυπα x_1 και x_2 τα οποία βρίσκονται και τα δύο πάνω στο υπερεπίπεδο τότε

$$w^t x_1 + w_0 = w^t x_2 + w_0 = 0$$

$$\Rightarrow w^t (x_2 - x_1) = 0 \quad (5)$$



Σχήμα 4. Αναπαράσταση μιας γραμμικής διαχωριστικής συνάρτησης στην περίπτωση δύο κλάσεων

Είναι προφανές ότι το διάνυσμα $x_1 - x_2$ ανήκει και αυτό στο υπερεπίπεδο, οπότε από την εξίσωση 5 συμπεραίνουμε ότι το διάνυσμα βαρών w είναι ορθογώνιο σε κάθε διάνυσμα που ανήκει στο υπερεπίπεδο και δηλώνει την κατεύθυνσή του.

Όσα περιγράψαμε μόλις φαίνονται στο Σχήμα 4 για την περίπτωση δύο κλάσεων και δύο χαρακτηριστικών, όπου η διαχωριστική συνάρτηση *D(x)* χωρίζει τον χώρο των χαρακτηριστικών σε δύο υποπεριοχές R_1 και R_2.
Ένα σύνολο αντικειμένων διαφορετικών κλάσεων που διαχωρίζονται πλήρως από μία γραμμική διαχωριστική συνάρτηση λέμε ότι είναι γραμμικώς διαχωριζόμενο. Ομοίως αν δύο κλάσεις διαχωρίζονται από μία γραμμική διαχωριστική συνάρτηση λέμε ότι είναι γραμμικά διαχωριζόμενες [23].

3.1.3 Εκτίμηση απόδοσης ταξινομητή, γενίκευση και overffiting

Αποτέλεσμα της εκπαίδευσης είναι ένας ταξινομητής ικανός να ταξινομήσει πρότυπα άγνωστα σε αυτόν, δηλαδή πρότυπα που δεν ανήκουν στο σύνολο

εκπαίδευσης. Αυτό δε σημαίνει φυσικά ότι κάθε νέο πρότυπο θα ταξινομείται σωστά. Για να ελέγξουμε το κατά πόσο ο ταξινομητής μπορεί να ταξινομήσει σωστά νέα πρότυπα, χρησιμοποιούμε ως είσοδο ένα νέο σύνολο προτύπων προς ταξινόμηση, το λεγόμενο independent test set. Για αυτό το σύνολο είναι γνωστές οι ετικέτες κλάσεων που αντιστοιχούν στα πρότυπα, δεν τις εισάγουμε όμως στον ταξινομητή. Μετά την ταξινόμηση γίνεται εκτίμηση του κατά πόσο σωστά έγινε η κατάταξη των προτύπων στις διαφορετικές κλάσεις, συγκρίνοντας τις ετικέτες κλάσεων που προκύπτουν με εκείνες που ήδη γνωρίζουμε. Υπολογίζεται λοιπόν το σφάλμα ταξινόμησης, το οποίο επιθυμούμε να είναι όσο το δυνατόν μικρότερο.

Εάν ο ταξινομητής είναι πολύ σύνθετος (η διαχωριστική συνάρτηση έχει πολλές παραμέτρους, π.χ. πολυώνυμο μεγάλου βαθμού), μπορεί να προσαρμοστεί ακριβώς στα δεδομένα του συνόλου εκπαίδευσης [21]. Αυτό σημαίνει ότι μπορεί να ταξινομεί με επιτυχία τα πρότυπα εκπαίδευσης αλλά να αδυνατεί να ταξινομήσει νέα πρότυπα. Αυτό είναι το φαινόμενο του overfitting. Είναι προτιμότερη λοιπόν η επιλογή ενός πιο απλού ταξινομητή που να ταξινομεί μεν καλά τα πρότυπα εκπαίδευσης – ίσως όχι με 100% επιτυχία – αλλά να μπορεί να ταξινομήσει με μεγάλο ποσοστό επιτυχίας και νέα πρότυπα. Αυτό είναι και το ζήτημα της γενίκευσης (generalization) το οποίο καλείται να ικανοποιεί κάθε ταξινομητής προκειμένου να θεωρηθεί καλή η απόδοσή του.

3.2 Μείωση της διάστασης του χώρου χαρακτηριστικών

Το πλήθος των προτύπων σε ένα train set που απαιτείται για την κατασκευή της συνάρτησης απόφασης, αυξάνει εκθετικά με την διάσταση του χώρου χαρακτηριστικών. Εάν ένα διάστημα στον μονοδιάστατο χώρο χρειάζεται για να γεμίσει πυκνά M ισαπέχοντα σημεία, το αντίστοιχο τετράγωνο στον δυσδιάστατο χώρο χρειάζεται M^2, ο αντίστοιχος κύβος στον τρισδιάστατο χώρο M^3 και ούτω καθεξής. Αυτό το φαινόμενο είναι γνωστό ως “curse of dimensionality” [20]. Σε περιπτώσεις όπου το πλήθος των χαρακτηριστικών είναι πολύ μεγάλο σχετικά με το πλήθος των προτύπων είναι πιθανό να εμφανιστεί το φαινόμενο του overfitting. Για ένα δεδομένο μοντέλο

ταξινόμησης, το πρόβλημα του overfitting γίνεται λιγότερο σοβαρό όσο αυξάνουμε το πλήθος των δειγμάτων εκπαίδευσης. Παρόλα αυτά σε ορισμένες περιπτώσεις αυτό είναι αδύνατο, οπότε ακολουθείται μια διαδικασία μείωσης της διάστασης του χώρου χαρακτηριστικών (dimensionality reduction) [23].

Παράλληλα η ύπαρξη πολλών χαρακτηριστικών απαιτεί πολλή μνήμη και υπολογιστική ισχύ. Όσο περισσότερα είναι τα χαρακτηριστικά τόσο πιο αργή είναι η διαδικασία της εκπαίδευσης. Επίσης, ανάμεσα σε πολλά χαρακτηριστικά, είναι πιθανόν να υπάρχουν κάποια τα οποία δεν είναι χρήσιμα για τον διαχωρισμό των κλάσεων, δηλαδή υπάρχει πλεονασμός (redundancy) – περιττή, άχρηστη πληροφορία – οπότε δεν ισχύει ότι όσο περισσότερα χαρακτηριστικά έχουμε στην διάθεσή μας, τόσο περισσότερη πληροφορία αντλούμε για την απόφαση ταξινόμησης.

Η ύπαρξη λοιπόν πολλών χαρακτηριστικών μπορεί να προκαλέσει σοβαρά προβλήματα σε πολλούς αλγορίθμους μάθησης, υποβιβάζοντας την απόδοσή τους. Προκειμένου λοιπόν να βελτιωθεί η ικανότητα γενίκευσης ενός ταξινομητή, και να επιταχυνθεί η διαδικασία εκπαίδευσης, μειώνονται τα χαρακτηριστικά.

Οι γονιδιακές εκφράσεις που προκύπτουν από τα πειράματα μικροσυστοιχιών DNA είναι πολυάριθμες σε σχέση με το πλήθος των ασθενών που συμμετέχουν στο πείραμα. Επομένως η μείωση της διάστασης των χαρακτηριστικών είναι και σε αυτή *τη*ν περίπτωση κρίσιμη, προκειμένου να εντοπιστούν πιο πληροφοριακά γονίδια και να προκύψουν πιο αποδοτικά μοντέλα ταξινόμησης.

Υπάρχουν δύο διαφορετικές προσεγγίσεις για την μείωση της διάστασης των χαρακτηριστικών (dimensionality reduction). Αυτές είναι η εξαγωγή χαρακτηριστικών (feature extraction) και η επιλογή χαρακτηριστικών (feature selection) [24].

3.2.1 Εξαγωγή χαρακτηριστικών

Κατά την εξαγωγή χαρακτηριστικών, νέα χαρακτηριστικά , λιγότερα από τα αρχικά, προκύπτουν με συνδυασμό – γραμμικό ή μη γραμμικό – των αρχικών [21]. Αυτό που γίνεται στην ουσία είναι μία αντιστοίχηση από τον *N*-διάστατο χώρο των χαρακτηριστικών σε έναν άλλο χώρο λιγότερων διαστάσεων. Ένα μειονέκτημα αυτής της μεθόδου είναι ότι κανένα από τα αρχικά χαρακτηριστικά δεν παραλείπεται στον υπολογισμό των νέων, οπότε ακόμα και εκείνα τα χαρακτηριστικά που δεν παρέχουν πληροφορία για το σωστό διαχωρισμό των κλάσεων λαμβάνονται υπόψη.

3.2.2 Επιλογή χαρακτηριστικών – Filter και Wrapper μέθοδοι

Η δεύτερη μέθοδος για την μείωση της διάστασης του χώρου χαρακτηριστικών είναι η επιλογή χαρακτηριστικών. Γίνεται προσπάθεια να βρεθεί ένα υποσύνολο των αρχικών χαρακτηριστικών το οποίο θα είναι αρκετό για να περιγράψει τα πρότυπα και να εκπαιδεύσει τον ταξινομητή [25].
Η βέλτιστη επιλογή χαρακτηριστικών για προβλήματα μάθησης με επίβλεψη απαιτεί εξαντλητική έρευνα όλων των δυνατών υποσυνόλων των χαρακτηριστικών. Εάν έχουμε στην διάθεσή μας μεγάλο αριθμό χαρακτηριστικών, αυτό είναι μη πρακτικό. Επομένως συνήθως γίνεται έρευνα για ένα ικανοποιητικό σύνολο χαρακτηριστικών αντί για το βέλτιστο δυνατό.

Στη στατιστική η πιο δημοφιλής μορφή επιλογής χαρακτηριστικών είναι η σταδιακή μείωση χαρακτηριστικών (stepwise regression). Είναι ένας αλγόριθμος σε κάθε επανάληψη του οποίου γίνεται ταξινόμηση των χαρακτηριστικών βάση κάποιου κριτηρίου και αφαιρείται το χειρότερο χαρακτηριστικό (ή επιλέγεται το καλύτερο). Σε περιπτώσεις όπου το πλήθος των χαρακτηριστικών είναι μεγάλο, δύναται η αφαίρεση (ή η επιλογή) περισσότερων του ενός χαρακτηριστικών τη φορά, με το κόστος να έχουμε υποβίβαση της απόδοσης ταξινόμησης [23]. Η εκτίμηση του κάθε υποσυνόλου μπορεί για παράδειγμα να γίνει με εκπαίδευση του ταξινομητή και εκτίμηση της απόδοσής του χρησιμοποιώντας μόνο τα συγκεκριμένα

χαρακτηριστικά από τα train και test sets, όπως γίνεται και στην παρούσα διπλωματική.

Σε αντίθεση με την εξαγωγή χαρακτηριστικών, χαρακτηριστικά τα οποία δεν δίνουν χρήσιμη πληροφορία μπορούν να απορριφθούν και έτσι καταλήγουμε σε πιο πληροφοριακά σύνολα χαρακτηριστικών με τα οποία επιτυγχάνεται καλύτερη απόδοση του ταξινομητή. Οι αλγόριθμοι επιλογής χαρακτηριστικών χωρίζονται σε δύο ευρείες κατηγορίες, τις filter μεθόδους και τις wrapper.

3.2.2.1 Filter μέθοδοι

Οι Filter μέθοδοι επικεντρώνονται σε γενικά χαρακτηριστικά των δεδομένων χρησιμοποιώντας διάφορες μεθόδους, όπως Fisher`s ratio,T-statistics,x^2-statistics [25-28] και πολλές άλλες. Τα γονίδια κατατάσσονται ανάλογα με το πόσο καλά τα πηγαίνουν κατά την εφαρμογή της εκάστοτε μεθόδου και στη συνέχεια αυτά τα οποία έχουν καταταχθεί ψηλά σε σχέση με τα υπόλοιπα, και τα οποία παρέχουν την υψηλότερη δυνατότητα διάκρισης και την υψηλότερη ακρίβεια ταξινόμησης, επιλέγονται ως marker genes. Οι Filter μέθοδοι δηλαδή βασίζονται σε γενικά χαρακτηριστικά των δεδομένων εκπαίδευσης ώστε να γίνει επιλογή χαρακτηριστικών χωρίς να περιλαμβάνεται οποιαδήποτε διαδικασία μάθησης ταξινομητή [29].

3.2.2.2 Wrapper μέθοδοι

Αντίθετα, οι Wrapper μέθοδοι απαιτούν την εκπαίδευση ενός ταξινομητή και κάνουν χρήση της απόδοσής του ώστε να εκτιμηθεί και να καθοριστεί ποια χαρακτηριστικά θα επιλεχθούν. Τείνουν να βρίσκουν χαρακτηριστικά με μεγάλο ρίσκο να εμφανίσουν το φαινόμενο του overfitting, όπως επίσης τείνουν να είναι υπολογιστικά πιο δαπανηρές από τις filter μεθόδους.

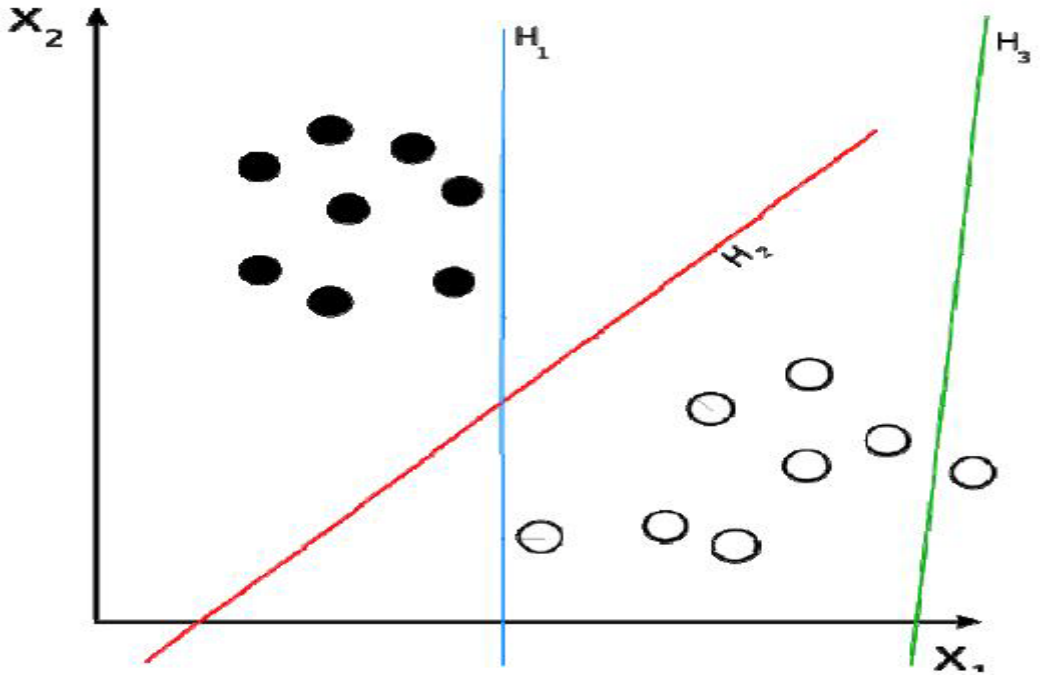
Λειτουργούν με έναν επαναληπτικό τρόπο κατά τον οποίο ο ταξινομητής χρησιμοποιείται για να αποδώσει ένα βάρος συσχέτισης σε κάθε χαρακτηριστικό και στη συνέχεια τα χαρακτηριστικά με τα χαμηλότερα βάρη εξαλείφονται σε κάθε επανάληψη. Σε κάθε επόμενη επανάληψη τα βάρη

επαναπροσδιορίζονται και προσαρμόζονται δυναμικά, ενώ η όλη διαδικασία συνεχίζεται επαναληπτικά. Στο τέλος το μικρότερο set χαρακτηριστικών, το οποίο επιτυγχάνει την ψηλότερη ακρίβεια ταξινόμησης, επιλέγεται ως marker genes set.

3.3 Support Vector Machines (SVMs)

Όπως έχουμε αναφέρει και παραπάνω υπάρχουν διάφορα υπερεπίπεδα που μπορούν να διαχωρίσουν τις κλάσεις μεταξύ τους. Τα γραμμικά Support Vector Machines (SVMs) είναι ένα σύνολο από μεθόδους μάθησης με επίβλεψη, οι οποίες κατασκευάζουν διαχωριστικά υπερεπίπεδα στο N-διάστατο χώρο των χαρακτηριστικών, τέτοια ώστε να μεγιστοποιείται το περιθώριο ανάμεσα στις κλάσεις. Η απόσταση ενός τέτοιου υπερεπιπέδου από το αντικείμενο κάθε κλάσης είναι η μέγιστη δυνατή [30].

Ας υποθέσουμε ότι έχουμε διαχωρισμό σε δύο κλάσεις, καθώς επίσης και ότι τα δεδομένα των δύο αυτών κλάσεων είναι γραμμικά διαχωριζόμενα. Για να υπολογίσουμε το περιθώριο ανάμεσα στις κλάσεις κατασκευάζονται δύο παράλληλα υπερεπίπεδα, ένα σε κάθε πλευρά του διαχωριστικού επιπέδου, καθένα από τα οποία περιέχει ένα τουλάχιστον αντικείμενο από την κάθε κλάση. Τα δύο αυτά υπερεπίπεδα ισαπέχουν από το βασικό υπερεπίπεδο, και προσδιορίζουν το περιθώριο ανάμεσα στις δύο κλάσεις. Ένας καλός διαχωρισμός επιτυγχάνεται από εκείνο το υπερεπίπεδο που έχει τη μεγαλύτερη απόσταση από τα κοντινότερα αντικείμενα των δύο κλάσεων, δηλαδή το περιθώριο είναι το μέγιστο, οπότε έχουμε μικρότερο σφάλμα γενίκευσης. Στο Σχήμα 5 φαίνεται η διαφορά ανάμεσα σε ένα οποιοδήποτε υπερεπίπεδο που διαχωρίζει δύο κλάσεις και σε αυτό που επιτυγχάνει το μέγιστο περιθώριο ανάμεσα στις κλάσεις.



Σχήμα 5. Δύο γραμμικώς διαχωριζόμενα σύνολα δεδομένων. Το υπερεπίπεδο H_3 δε διαχωρίζει τις δύο κλάσεις. Τα υπερεπίπεδα H_1 και H_2 διαχωρίζουν τις δύο κλάσεις όμως το H_1 πετυχαίνει μικρό περιθώριο ανάμεσα στις κλάσεις, ενώ το H_2 μέγιστο.

Εκείνα τα σημεία που βρίσκονται στα όρια του περιθωρίου, δηλαδή βρίσκονται πάνω στα υπερεπίπεδα εκατέρωθεν του διαχωριστικού υπερεπιπέδου, ονομάζονται support vectors [20]. Μια ιδιαιτερότητα των SVMs είναι ότι τα βάρη της συνάρτησης απόφασης που προκύπτει υπολογίζονται μόνο βάση των support vectors.

Έστω λοιπόν ότι $X=\{x_1, x_2, \ldots, x_M\}$ είναι το σύνολο των M αντικειμένων που χρησιμοποιούνται για την εκπαίδευση του ταξινομητή, όπου x_i είναι ένα διάνυσμα χαρακτηριστικών (πρότυπο) στον N-διάστατο χώρο χαρακτηριστικών. Σε κάθε πρότυπο αντιστοιχεί μια ετικέτα κλάσης y_i , η οποία θα είναι είτε +1 εάν το αντικείμενο x_i ανήκει στην κλάση ω_1, είτε -1 εάν αντιστοιχεί στην κλάση ω_2. Επειδή αναφερόμαστε σε γραμμικά SVMs η διαχωριστική συνάρτηση $D(x)$ θα είναι της μορφής:

$$D(x) = w^T x + w_0 \qquad (6)$$

Για *D(x)* > 0 το πρότυπο *x* ταξινομείται στην κλάση ω_1 (*y* = +1) διαφορετικά ταξινομείται στην κλάση ω_2 (*y* = −1) . Επομένως όλα τα πρότυπα εκπαίδευσης ταξινομούνται σωστά εάν ισχύει

$$y_i(w^T x_i + w_0) \geq 0 \quad \text{για i=1,2,3,…,M} \tag{7}$$

Έστω *H* το διαχωριστικό υπερεπίπεδο για το οποίο ισχύει ότι *D(x)* = 0. Ορίζουμε δύο ισαπέχοντα υπερεπίπεδα H_1 και H_2 εκατέρωθεν του *H* για τα οποία ισχύει:

$$H_1: w^T x + w_0 = +1$$

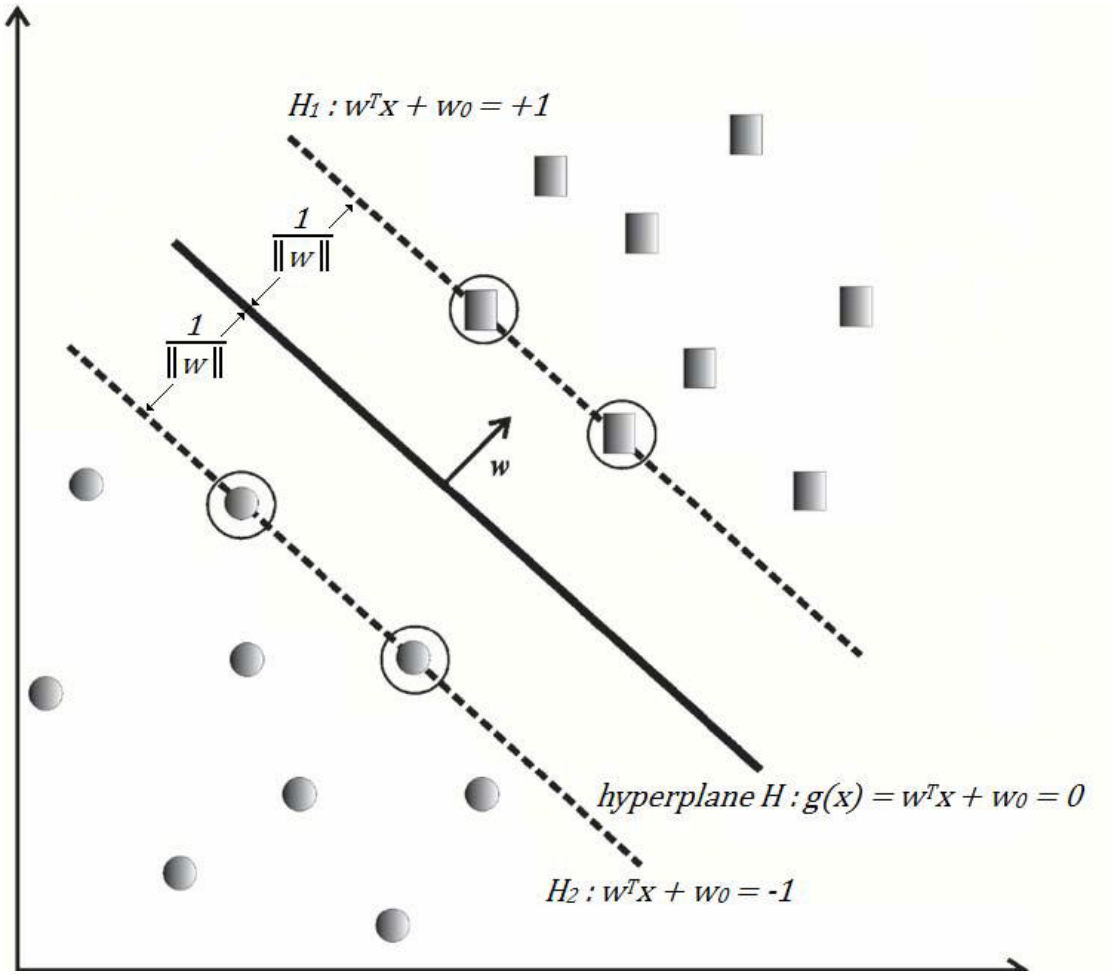
$$H_2: w^T x + w_0 = -1 \tag{8}$$

οπότε για κάθε πρότυπο εκπαίδευσης x_i με ετικέτα κλάσης y_i έχουμε τις εξής σχέσεις:

$$w^T x_i + w_0 \geq +1 \quad \text{για } y_i = +1 \tag{9}$$

$$w^T x_i + w_0 \leq -1 \quad \text{για } y_i = -1 \tag{10}$$

Η απόσταση ανάμεσα σε κάθε ένα από αυτά τα υπερεπίπεδα και το διαχωριστικό υπερεπίπεδο *H* είναι 1/||*w*||, οπότε το περιθώριο έχει πλάτος 2/||*w*||. Το Σχήμα 6 δείχνει το διαχωριστικό υπερεπίπεδο και τα δύο υπερεπίπεδα H_1 και H_2 για δύο γραμμικά διαχωριζόμενα σύνολα δεδομένων, όπου τα support vectors είναι τα κυκλωμένα σημεία.



Σχήμα 6. Υπερεπίπεδα και support vectors για δύο γραμμικά διαχωριζόμενα σύνολα δεδομένων

Για να έχουμε ελάχιστο σφάλμα γενίκευσης, πρέπει να μεγιστοποιηθεί το περιθώριο, δηλαδή η ποσότητα 2/||*w*||. Αυτό σημαίνει ότι πρέπει να βρεθεί μια λύση που να ελαχιστοποιεί το ||*w*||. Οπότε η συνάρτηση κόστους θα είναι:

$$J(w) = \frac{1}{2}\|w\|^2 \tag{11}$$

η οποία πρέπει να ελαχιστοποιηθεί υπό τον όρο ότι:

$$y_i(w^T x_i + w_0) \geq 1 \quad \text{για i=1,2,.....M} \tag{12}$$

Αυτός είναι και ο στόχος ενός hard margin SVM , η ελαχιστοποίηση δηλαδή του σφάλματος γενίκευσης μεγιστοποιώντας το περιθώριο ανάμεσα στα δύο υπερεπίπεδα.

Σε πολλά προβλήματα όμως είναι πολύ πιθανό να μην υπάρχει κάποιο γραμμικό όριο απόφασης που να διαχωρίζει πλήρως τις κλάσεις οπότε το να ψάχνει κανείς να βρει ένα βέλτιστο υπερεπίπεδο είναι άνευ σημασίας [20]. Αυτό σημαίνει ότι το περιθώριο που ορίζεται από τα δύο υπερεπίπεδα H_1 και H_2 δεν είναι άδειο αλλά υπάρχουν πρότυπα του train set.

Σε αυτές τις περιπτώσεις τα δείγματα εκπαίδευσης μπορεί να ανήκουν σε μία από τις παρακάτω τρεις κατηγορίες:

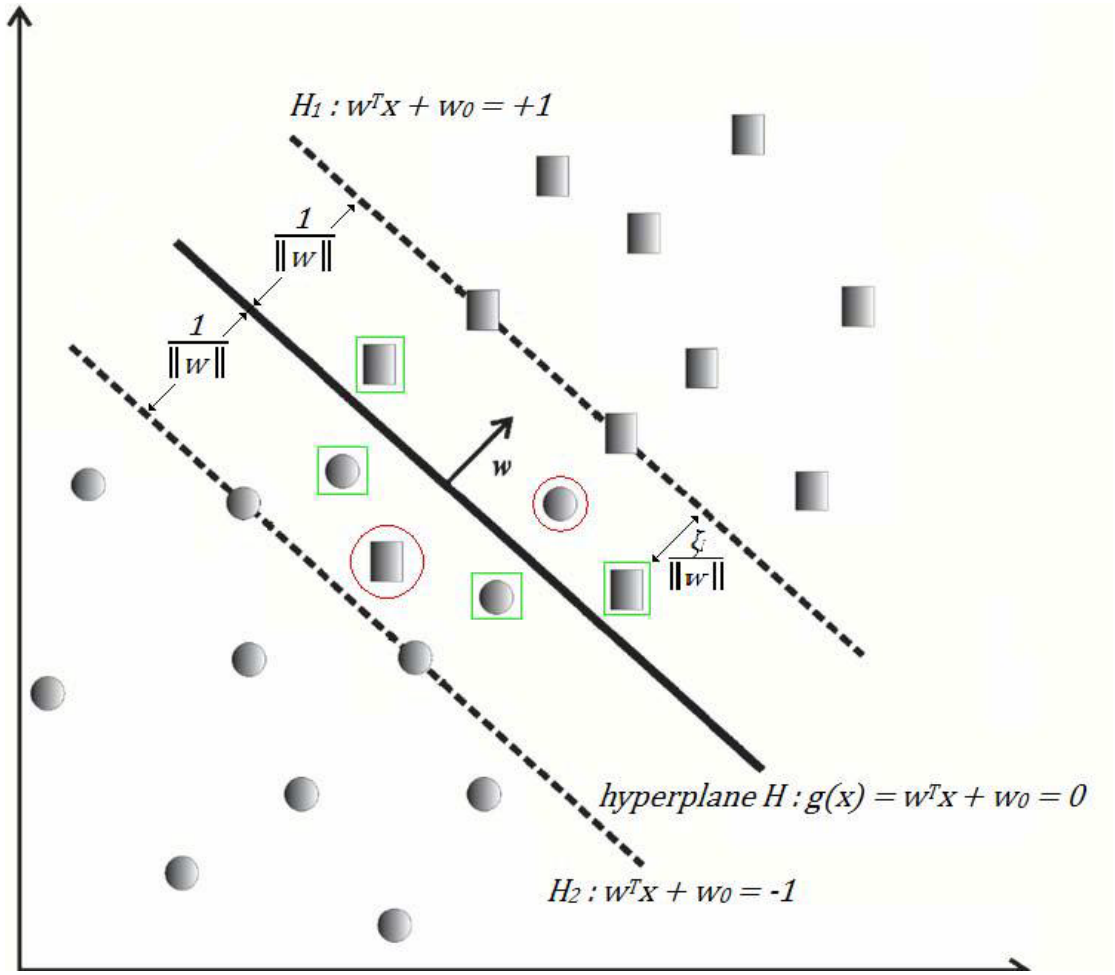
- Να βρίσκονται εκτός περιθωρίου και να είναι σωστά ταξινομημένα.

- Να βρίσκονται εντός περιθωρίου και να είναι σωστά ταξινομημένα (στο Σχήμα 7 τα σημεία με το τετράγωνο περίγραμμα). Αυτό σημαίνει ότι ικανοποιούν την ανισότητα:

$$0 \leq y_i(w^T x_i + w_0) < 1$$

- Να βρίσκονται εντός ή εκτός περιθωρίου και να μην είναι σωστά ταξινομημένα (στο Σχήμα 7 τα σημεία με το κυκλικό περίγραμμα). Αυτό σημαίνει ότι ικανοποιούν την ανισότητα:

$$y_i(w^T x_i + w_0) < 0$$



Σχήμα 7. Μη διαχωριζόμενα γραμμικά σύνολα δεδομένων. Με τετράγωνο περίγραμμα είναι τα δεδομένα που ταξινομούνται σωστά και με κυκλικό αυτά που ταξινομούνται λάθος.

Οι soft margin SVMs επιδιώκουν την βελτίωση στην απόδοση χαλαρώνοντας τους περιορισμούς των hard margin SVMs. Για αυτό το λόγο εισάγονται κάποιες μεταβλητές ξ_i, $i=1,2,\ldots M$ μέσα στον περιορισμό της σχέσης 17 οπότε και έχουμε:

$$y_i(w^T x_i + w_0) \geq 1 - \xi_i \tag{13}$$

Αυτή η σχέση ικανοποιεί και τις τρεις παραπάνω κατηγορίες δεδομένων που περιγράφηκαν. Πιο συγκεκριμένα, η πρώτη κατηγορία αντιστοιχεί σε $\xi_i = 0$, η δεύτερη σε $0< \xi_i \leq 1$ και η τρίτη σε $\xi_i > 1$. Οι μεταβλητές ξ_i είναι γνωστές ως slack variables. Η απόσταση ενός σημείου x_i που ανήκει στην δεύτερη

κατηγορία από το υπερεπίπεδο που αντιστοιχεί στην κλάση στην οποία ανήκει είναι ξ_i /||*w*||, όπως φαίνεται στο παραπάνω σχήμα (Σχήμα 7). Ο στόχος μας λοιπόν είναι να έχουμε ένα όσο το δυνατόν μεγαλύτερο περιθώριο, διατηρώντας όμως ταυτόχρονα τον αριθμό των δειγμάτων με $\xi_i > 0$ όσο γίνεται πιο χαμηλό. Με μαθηματικούς όρους αυτό ισοδυναμεί με την ελαχιστοποίηση της ακόλουθης συνάρτησης κόστους:

$$J(w, w_0, \xi) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{M}\xi_i \qquad (14)$$

υπό τον περιορισμό της σχέσης 18 και για

$$\xi_i \geq 0 \text{ για } i = 1,2,\dots,M \qquad (15)$$

Η *C* είναι μια θετική σταθερά η οποία ελέγχει την σχετική επιρροή των παραμέτρων *w* και ξ_i. Καθορίζει την ισορροπία ανάμεσα στο να έχουμε ένα μεγάλο περιθώριο, με το κόστος περισσότερων λάθος ταξινομημένων δειγμάτων και στο να έχουμε ένα μικρό περιθώριο, με λιγότερα όμως λάθος ταξινομημένα δείγματα. Όσο μεγαλύτερη είναι η παράμετρος *C*, τόσο περισσότερο πλησιάζουμε στην συμπεριφορά ενός hard margin SVM, ενώ για πολύ μικρές τιμές της *C* προκύπτει μεγάλο σφάλμα γενίκευσης [31]. Δεν υπάρχει κάποια βέλτιστη τιμή για την παράμετρο *C*. Η απόφαση για το ποια θα είναι η τιμή της μπορεί να ληφθεί είτε με δοκιμές είτε με εφαρμογή της LOOCV μεθόδου για βελτιστοποίησή της.

Έχουμε να κάνουμε λοιπόν με ένα μη γραμμικό πρόβλημα βελτιστοποίησης το οποίο υπόκειται σε ένα σύνολο γραμμικών περιορισμών ανισότητας [20]. Η ελαχιστοποίηση της 19 υπό τον περιορισμό των 18 και 20 απαιτεί την ικανοποίηση των λεγόμενων Karush-Kuhn-Tucker (KKT) συνθηκών που είναι οι εξής:

$$\frac{\partial L}{\partial w} = 0 \Longrightarrow w = \sum_{i=1}^{M}\lambda_i x_i y_i \qquad (16)$$

$$\frac{\partial L}{\partial w_0} = 0 \Longrightarrow w = \sum_{i=1}^{M} \lambda_i y_i \quad (17)$$

$$\frac{\partial L}{\partial \xi_i} = 0 \Longrightarrow C - \mu_i - \lambda_i = 0, i = 1,2,\ldots.M \quad (18)$$

$$\lambda_i[y_i(w^T x_i + w_0) - 1 + \xi_i] = 0,\ i = 1,2,\ldots.,M \quad (19)$$

$$\mu_i \xi_i = 0, i = 1,2,\ldots M \quad (20)$$

$$\mu_i \geq 0, \lambda_i \geq 0, i = 1,2,\ldots M \quad (21)$$

όπου λ είναι το διάνυσμα των πολλαπλασιαστών Lagrange λ_i, $i=1,2,\ldots,M$ και L είναι η συνάρτηση Lagrange που ορίζεται ως:

$$L(w, w_0, \xi, \lambda, \mu) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{M}\xi_i \sum_{i=1}^{M}\mu_i\xi_i \sum_{i=1}^{M}\lambda_i[y_i(w^T x_i + w_0) - 1 + \xi_i] \quad (22)$$

Η δυαδική αναπαράσταση της συνάρτησης Lagrange είναι η εξής:

μεγιστοποίηση του $L(w, w_0, \lambda, \xi, \mu)$ (23)

υπό τον όρο ότι $w = \sum_{i=1}^{M} \lambda_i y_i x_i$ (24)

$$\sum_{i=1}^{M} \lambda_i y_i = 0 \quad (25)$$

$$C - \mu_i - \lambda_i = 0 \quad (26)$$

και $\lambda_i \geq 0 \mu_i \geq 0,\ i = 1,2,\ldots M$ (27)

Αντικαθιστώντας τους παραπάνω περιορισμούς στην συνάρτηση Lagrange προκύπτει:

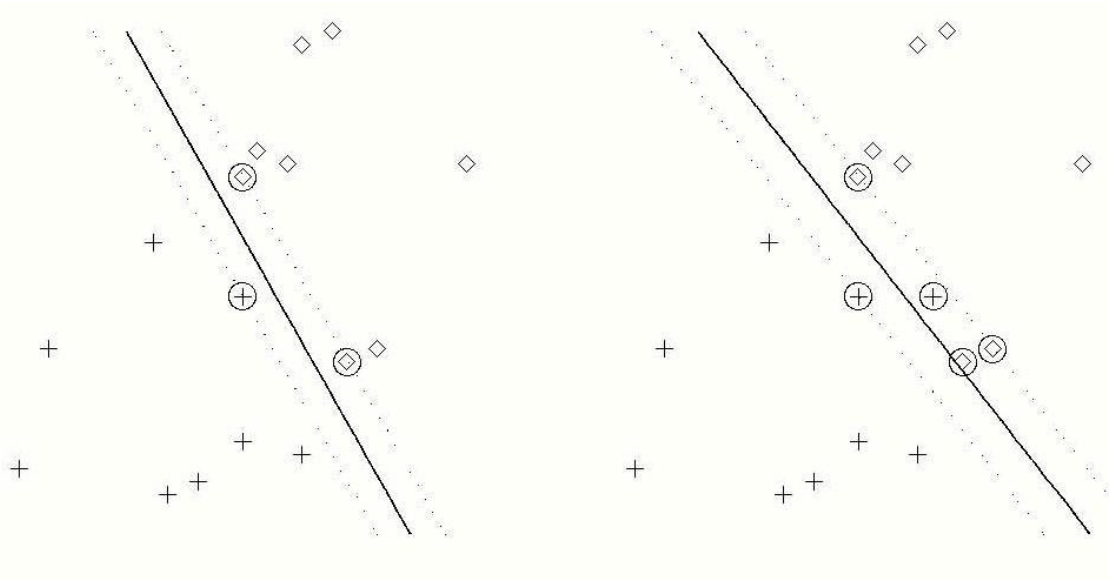
$$max_{\lambda}(\sum_{i=1}^{M}\lambda_i - \frac{1}{2}\sum_{i=1}^{M}\ \sum_{i=1}^{M}\lambda_i\lambda_j y_i y_j x_i^T x_j) \quad (28)$$

υπό τον όρο ότι $0 \le \lambda_i \le C, i = 1,2,\dots M$ (29)

και $\sum_{i=1}^{M}\lambda_i y_i = 0$ (30)

Η σπουδαιότητα της δυαδικής μορφής είναι ότι εκφράζει το κριτήριο βελτιστοποίησης ως εσωτερικό γινόμενο των προτύπων x_i, κάτι το οποίο είναι ιδιαίτερα σημαντικό. Η συνάρτηση κόστους δεν εξαρτάται ρητά από την διάσταση του χώρου χαρακτηριστικών, κάτι το οποίο έχει ιδιαίτερες συνέπειες για τα μη γραμμικά SVMs τα οποία δε θα μας απασχολήσουν σε αυτή τη διπλωματική.

Εκείνα τα πρότυπα x_i για τα οποία ο πολλαπλασιαστής Lagrange λ_i είναι διάφορος του 0, είναι τα support vectors και αποτελούν τα πιο πληροφοριακά πρότυπα του train set. Όπως φαίνεται και από την σχέση 29, το διάνυσμα βαρών *w* υπολογίζεται μόνο από τα πρότυπα x_i για τα οποία $\lambda_i > 0$. Όλα τα υπόλοιπα δείγματα του train set είναι περιττά και δεν λαμβάνονται υπόψη στο σχηματισμό της συνάρτησης απόφασης. Εάν $\xi_i = 0$ τότε τα support vectors βρίσκονται πάνω στα υπερεπίπεδα εκατέρωθεν του διαχωριστικού υπερεπιπέδου *H* (τα λεγόμενα margin vectors). Εάν $\xi_i > 1$ τότε ταξινομούνται λάθος, ενώ αν $\xi_i < 1$ ταξινομούνται σωστά. Στο Σχήμα 8 φαίνεται η περίπτωση γραμμικά διαχωριζόμενων κλάσεων σε σχέση με την περίπτωση των μη διαχωριζόμενων γραμμικά κλάσεων. Τα support vectors είναι περιγεγραμμένα από ένα κύκλο. Στην δεξιά εικόνα ένα support vector ταξινομείται στη λάθος κλάση.



Σχήμα 8. Διαφορά μεταξύ γραμμικά διαχωριζόμενων κλάσεων (αριστερά) και γραμμικά μη διαχωριζόμενων κλάσεων (δεξιά)

Επίσης για $\xi_i \neq 0$ από τη σχέση 25 προκύπτει ότι $\mu_i = 0$ οπότε από την σχέση 31 προκύπτει ότι $\lambda_i = C$. Οπότε για όλα τα support vectors που δεν βρίσκονται πάνω σε κάποιο από τα υπερεπίπεδα H_1 και H_2, η τιμή του πολλαπλασιαστή Lagrange λ_i είναι ίση με την τιμή της σταθεράς C.

Με βάση τα παραπάνω, η διαχωριστική συνάρτηση μπορεί να γραφτεί ως:

$$D(x) = \sum_{i=0}^{M} y_i \lambda_i (x^T x_i) + w_0 \quad (31)$$

όπου για $D(x) = 0$ προκύπτει το διαχωριστικό υπερεπίπεδο. Το w_0 μπορεί να υπολογιστεί αντικαθιστώντας ένα από τα support vectors στην παρακάτω σχέση:

$$y_j \left((w x_j) + w_0 \right) = 1 \quad (32)$$

Για μία πιο πλήρη ενημέρωση όσον αφορά τα SVMs μπορεί κάποιος να ανατρέξει στο [15].

3.4 Ο προτεινόμενος ταξινομητής

Η επιλογή γονιδίων στα πλαίσια της ταξινόμησης προτύπων, μπορεί να λυθεί ως πρόβλημα επιλογής χαρακτηριστικών γνωρισμάτων. Γενικά, ένας αλγόριθμος επιλογής χαρακτηριστικών γνωρισμάτων αποτελείται κυρίως από δύο βασικά συστατικά: τη διαδικασία αναζήτησης και το κριτήριο αξιολόγησης (Dash and Liu, 1997). Η διαδικασία αναζήτησης παράγει τα υποσύνολα χαρακτηριστικών γνωρισμάτων που είναι υποψήφια για την αξιολόγηση. Στη διαδικασία αναζήτησης, τα υποψήφια υποσύνολα χαρακτηριστικών γνωρισμάτων μπορούν να παραχθούν είτε ακολουθιακά είτε τυχαία. Η sequential forward selection (SFS) αρχίζει από ένα κενό σύνολο και προσθέτει επαναληπτικά τα χαρακτηριστικά γνωρίσματα, ενώ η sequential backward elimination (SBE) αρχίζει από το πλήρες σύνολο χαρακτηριστικών γνωρισμάτων και διαγράφει επαναληπτικά τα χαρακτηριστικά γνωρίσματα. Το κριτήριο αξιολόγησης μετρά την ποιότητα των υποψηφίων υποσυνόλων χαρακτηριστικών γνωρισμάτων που παράγονται από τη διαδικασία αναζήτησης. Σε κάθε βήμα της επαναληπτικής διαδικασίας, το χαρακτηριστικό γνώρισμα που οδηγεί στη μέγιστη βελτίωση μετά από την προσθήκη ή τη λιγότερη υποβάθμιση μετά από τη διαγραφή επιλέγεται. Γενικά, η επιλογή χαρακτηριστικών γνωρισμάτων μπορεί να εκτελεσθεί με δύο τρόπους: τις filter και τις wrapper μεθόδους.

Αν και ένα πλήθος προσεγγίσεων των Wrapper μεθόδων, βασιζόμενων κυρίως σε SVMs (Support vector Machines), έχουν προταθεί τα τελευταία χρόνια[25,32,33], δεν έχει γίνει ακόμα καμία προσπάθεια να ενσωματωθούν οι μέθοδοι Filter και Wrapper. Η πιο σχετική προσπάθεια για αυτό το σκοπό είναι το μοντέλο επιλογής γονιδιακής έκφρασης GEMS (Genes Expression Model Selection) [34,35]. Χρησιμοποιεί το filter criterion για να κατατάξει τα γονίδια, ενώ ταυτόχρονα η διαδικασία επιλογής προτιμά το γονίδιο που μεγιστοποιεί την απόδοση ταξινόμησης. Συγκεκριμένα χρησιμοποιεί έναν νέο κριτήριο αξιολόγησης που λειτουργεί σα φίλτρο, αποκαλούμενο LS Bound measure, για την επιλογή γονιδίων. Το νέο κριτήριο έχει τα πλεονεκτήματα και των μεθόδων filter και των μεθόδων wrapper. Κατ' αρχάς, το κριτήριο προέρχεται από την leave-one-out cross validation (LOOCV) διαδικασία των

least squares support vector machines (LS-SVM) και είναι στενά συνδεδεμένο με το ανώτερο όριο των αποτελεσμάτων της μεθόδου LOOCV. Επομένως, το κριτήριο βρίσκει τα γονίδια που οδηγούν στην ακριβή ταξινόμηση. Δεύτερον, η εκτίμηση του ανώτερου ορίου περιλαμβάνει την εκπαίδευση του ταξινομητή μόνο μια φορά, χωρίς επαναλαμβανόμενη χρήση της cross validation διαδικασίας. Κατά συνέπεια, η υπολογιστική πολυπλοκότητα μειώνεται σημαντικά έναντι της κλασσικής μεθόδου wrapper.

Ο αλγόριθμος που χρησιμοποιούμε σε αυτή την εργασία είναι ο RFE-LNW[36]. Αυτός πηγαίνει ένα βήμα πιο πέρα το σκεπτικό της ενσωμάτωσης των Filter και Wrapper μεθόδων, ενσωματώνοντας επίσης γενικά χαρακτηριστικά σε κλασικά εργαλεία ταξινόμησης προτύπων, δηλαδή εμφυτεύοντας filtering criteria σε λειτουργίες wrapper, καταφέρνοντας έτσι να βελτιώσει την απόδοση σε σχέση με τις ήδη υπάρχουσες Wrapper μεθόδους. Ο αλγόριθμος RFE-LNW (Recursive Feature Elimination based on Linear Neuron Weights) [36] ήταν μια προσπάθεια γεφύρωσης του χάσματος ανάμεσα στις δύο φιλοσοφίες. Βασίζεται σε ένα linear neuron (LN) εμπλουτισμένο με μια παραλλαγή του Fisher`s metric

3.4.1 Ο συντελεστής Fisher

Η επιλογή διαφορετικά εκπεφρασμένων γονιδίων είναι επιθυμητός στόχος των μεθόδων επιλογής γονιδίων [38] και αυτό έχει επισημανθεί σε πολλές μελέτες. Σε όλες αυτές χρησιμοποιούνται παραλλαγές του συντελεστή του Fisher ο οποίος δίνεται από την παρακάτω εξίσωση:

$$f_1(g_i) = \frac{(\mu_+(g_i) - \mu_-(g_i))^2}{\sigma_+(g_i)^2 + \sigma_-(g_i)^2} \quad (33)$$

Μια παραλλαγή του συντελεστή του Fisher μπορεί να εκφραστεί ως:

$$f_2(g_i) = \frac{\sum_{j=1}^{M} |g_{ij} - c(g_i)|}{\sigma_+(g_i) + \sigma_-(g_i)} \quad (34)$$

Όπου $\mu_+(g_i)$, $\mu_-(g_i)$, $\sigma_+(g_i)$ και $\sigma_-(g_i)$ είναι οι μέσες και οι τυπικές αποκλίσεις των εκφράσεων του γονιδίου g_i για τη θετική και την αρνητική κλάση αντίστοιχα. M είναι ο αριθμός των δειγμάτων και

$$c(g_i) = \frac{(\mu_+(g_i)+\mu_-(g_i))}{2} \tag{35}$$

Μια άλλη παραλλαγή χαμηλού υπολογιστικού κόστους είναι:

$$f_3 = \frac{|\mu_+(g_i)+\mu_-(g_i)|}{\sigma_+(g_i)+\sigma_-(g_i)} \tag{36}$$

Μπορεί εύκολα να δειχτεί ότι οι εξισώσεις (33), (34) και (36) εκφράζουν το ίδιο πράγμα. Χρησιμοποιώντας αυτές τις εξισώσεις για να ορίσουμε τα βάρη σε ένα σετ δοθέντων γονιδίων, είναι προφανές ότι τα γονίδια των οποίων η έκφραση διαφοροποιείται περισσότερο στις δύο καταστάσεις (π.χ. -1 στην παθολογική κατάσταση και +1 στην κανονική) λαμβάνουν υψηλότερα βάρη από αυτά που διαφοροποιούνται λιγότερο ανάμεσα στις δύο κλάσεις. Γονίδια τα οποία εκφράζονται με ακριβώς τον ίδιο τρόπο στις δύο καταστάσεις ορίζονται να έχουν την ελάχιστη τιμή βάρους, η οποία είναι η μηδενική.

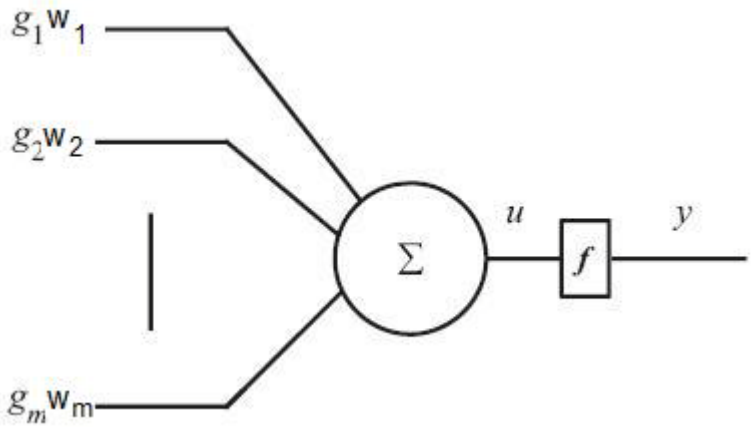
3.4.2 Ο ταξινομητής RFE-LNW

Οι περισσότερες προσεγγίσεις επιλογής δεικτών που έχουν εφαρμοστεί στο πεδίο των DNA microarrays λόγω της υψηλής διαστασιμότητας (μεγάλος αριθμός γονιδίων) των δεδομένων χρησιμοποιούν γραμμικά εργαλεία. Στον RFE-SVM χρησιμοποιείται ένας γραμμικός πυρήνας για να εκτιμήσει το διάνυσμα του βάρους του διαχωριστικού υπερεπιπέδου, η απόλυτη τιμή του οποίου χρησιμοποιείται στη συνέχεια ως κριτήριο ταξινόμησης των γονιδίων.

Τα LNs (Linear Neurons) χάρη στο σχεδιασμό τους (γραμμικός συνδυασμός εισόδων) μπορούν επίσης να προσεγγίσουν οποιαδήποτε γραμμική συνάρτηση.

Για αυτό το λόγο ο RFE-LNW χρησιμοποιεί ένα LN για να προσεγγίσει το διαχωριστικό υπερεπίπεδο ανάμεσα στις αρνητικές και στις θετικές κλάσεις. Εκμεταλλευόμενος μια τόσο ανοιχτή αρχιτεκτονική μπορεί να επιλέξει ανάμεσα σε μια πληθώρα σχημάτων εκμάθησης, ή εύκολα να ενσωματώσει μια νέα διαδικασία εκμάθησης κατάλληλα προσαρμοσμένη στο υπό εξέταση πρόβλημα, ενώ επίσης μπορεί να επεκταθεί σε πολυστρωματική υλοποίηση.

Στη δική μας δουλειά έχει εφαρμοστεί ένα γραμμικό νευρωνικό δίκτυο M εισόδων, όπου M ο αριθμός των γονιδίων, το οποίο μπορεί να έχει δύο πιθανές εξόδους (0 για την αρνητική κλάση και 1 για τη θετική), ώστε να προσεγγίσουμε το διαχωριστικό υπερεπίπεδο που ξεχωρίζει τις δύο κλάσεις.



Σχήμα 9. Ένας νευρώνας προσαρμοσμένος στο πρόβλημα

Πιο συγκεκριμένα χρησιμοποιώντας τη σιγμοειδή συνάρτηση $f(u)$ παίρνουμε:

$$y = \frac{1}{1+e^{u}} = f(u) \tag{37}$$

$$u = \sum_{i=1}^{M} w_i g_i \tag{38}$$

$$\dot{f}(u) = y(1-y) \tag{39}$$

Σημείωση: $\dot{f}(u) \geq 0, \alpha\varphi\acute{o}\ 0 \leq y \leq 1$

3.4.3 Εκπαίδευση του RFE-LNW

Σε αυτή την ενότητα θα παρατεθεί το μαθηματικό υπόβαθρο της διαδικασίας εκπαίδευσης, την οποία χρησιμοποιήσαμε για να εξάγουμε τα βάρη του LN. Η συνάρτηση σφάλματος, την οποία θέλουμε να ελαχιστοποιήσουμε δίνεται από τον τύπο:

$$E = \frac{1}{2}\sum_{j=1}^{M}(d_j - y_j)^2 \tag{40}$$

Όπου το M αντιστοιχεί στον αριθμό των δειγμάτων (ασθενών), το d_j αναπαριστά την επιθυμητή συσχετισμένη με το δείγμα j έξοδο του νευρώνα και το y_j είναι η πραγματική έξοδος που παράγεται από το νευρώνα για το δοθέν δείγμα. Μέσω της μεθόδου gradient descent για την ελαχιστοποίηση της τελευταίας εξίσωσης αναπροσαρμόζουμε το βάρος w_i το οποίο είναι συσχετισμένο με το γονίδιο g_i ως εξής:

$$w_i(t+1) = w_i(t) - \left(\mu\frac{\partial E}{\partial w_i}\right) = w_i(t) - \mu\sum_{j=1}^{M}\left(\frac{\partial E}{\partial y_j}\frac{\partial y_j}{\partial u}\frac{\partial u}{\partial w_i}\right)$$

Επηρεάζουμε την αναπροσαρμογή του βάρους πολλαπλασιάζοντας με το Fisher`s metric:

$$w_i(t+1) = w_i(t) - \frac{\mu}{2}\sum_{j=1}^{M}\left[\left(\frac{\partial E}{\partial y_j}\frac{\partial y_j}{\partial u_j}\frac{\partial u_j}{\partial w_i}\right)\frac{|g_{ij}-c(g_i)|}{\sigma_+(g_i)+\sigma_-(g_i)}\right] =$$

$$= w_i(t) - \frac{\mu}{2}\sum_{j=1}^{M}\left(-2\left(d_j - y_j\right)y_j\left(1-y_j\right)g_{ij}\right)f_2(g_i) =$$

$$= w_i(t) + \mu\sum_{j=1}^{M}\left(d_j - y_j\right)y_j\left(1-y_j\right)g_{ij}f_2(g_i) =$$

$$= w_i(t) + \mu\sum_{j=1}^{M}\left(d_j - y_j\right)\dot{f}(u_j)g_{ij}f_2(g_i)$$

Τελικά:

$$w_i(t+1) = w_i(t) + \mu \sum_{j=1}^{M} e_j \dot{f}(u_j) g_{ij} \frac{|g_{ij} - c(g_i)|}{\sigma_+(g_i) + \sigma_-(g_i)} \quad (41)$$

όπου το t αναπαριστά την εκάστοτε επανάληψη, μ είναι ο ρυθμός εκμάθησης και

$$e_j = (d_j - y_j) \quad (42)$$

Δουλεύοντας με πρόσημα, ιδέα η οποία εισήχθη στην resilient back-propagation εκμάθηση [37], η (41) μπορεί να εκφραστεί ως εξής:

- Στην περίπτωση όπου $f_2(g_i) = 1$, το οποίο είναι ίδιο με τη στάνταρ back-propagation διαδικασία, έχουμε:

$$w_i(t+1) = w_i(t) + \mu \sum_{j=1}^{M} e_j \dot{f}(u_j) g_{ij} \quad (43)$$

$$w_i(t+1) = w_i(t) + \mu \sum_{j=1}^{M} sign\left(e_j \dot{f}(u_j)\right) sign(g_{ij}) \quad (44)$$

- Ή γενικά:

$$w_i(t+1) = w_i(t) + \mu \sum_{j=1}^{M} sign\left(e_j \dot{f}(u_j)\right) sign(g_{ij}) f_2(g_i)$$

(45)

$$w_i(t+1) = w_i(t) + \sum_{j=1}^{M} |d_j - y_i| sign\left(e_j \dot{f}(u_j)\right) sign(g_{ij}) f_2(g_i)$$

(46)

3.4.3.1 Σύγκλιση του αλγορίθμου του RFE-LNW

Η εξίσωση (43) είναι ο βασικός αλγόριθμος εκμάθησης gradient descent, ο οποίος έχει αποδειχθεί ότι συγκλίνει και ότι επιτυγχάνει ελαχιστοποίηση της

συνάρτησης σφάλματος. Η εξίσωση (44) συγκλίνει εφόσον κρατώντας το πρόσημο της gradient κατευθυνόμαστε προς το ελάχιστο, το οποίο εν τέλει θα επιτευχθεί χρησιμοποιώντας το κατάλληλο ρυθμό εκμάθησης. Υπάρχει η πιθανότητα να παγιδευτούμε σε ένα τοπικό ελάχιστο, αλλά αυτό είναι ένα γενικότερο πρόβλημα των νευρωνικών δικτύων. Στην πραγματικότητα αναμένουμε ότι η (44) θα συγκλίνει πιο γρήγορα από την (43) επειδή τα e_j και $\dot{f}(u_j)$ στην (43) μπορούν να πάρουν πολύ μικρές τιμές καταλήγοντας σε πολύ μικρές τιμές του βάρους, το οποίο οδηγεί με τη σειρά του σε μικρές τιμές της συνάρτησης σφάλματος, πράγμα το οποίο καθυστερεί τη σύγκλιση [37]. Παίρνοντας μόνο το πρόσημο της gradient στη (44) οδηγούμαστε προς το ελάχιστο με μεγαλύτερα βήματα, επιταχύνοντας τη σύγκλιση, τουλάχιστον όταν η διαδικασία βρίσκεται ακόμα μακριά από το ελάχιστο. Αυτό αναγκάζει τον αλγόριθμο να συγκλίνει πολύ γρηγορότερα ειδικά στα πρώτα βήματα της διαδικασίας, όπου ο αριθμός των γονιδίων είναι ακόμα υπερβολικά μεγάλος.

Η Εξίσωση (44) όμως παίρνοντας μόνο το πρόσημο, αποδίδει ίδιες τιμές βάρους σε γονίδια που εκφράζονται πολύ διαφορετικά και σε γονίδια που εκφράζονται λιγότερο διαφορετικά. Είναι σωστό τα πρώτα να πάρουν μεγαλύτερο βάρος από τα δεύτερα, πράγμα το οποίο επιτυγχάνεται με την εξίσωση (45) στην οποία εισάγεται ο όρος $f_2(g_i)$.

Η εξίσωση (41), από την οποία προέρχεται η (45) είναι στην ουσία η μέθοδος gradient descent πολλαπλασιασμένη με την παραλλαγή του Fisher`s metric της εξίσωσης (34). Το κλάσμα $\frac{|g_{ij}-c(g_i)|}{\sigma_+(g_i)+\sigma_-(g_i)}$ που βρίσκεται μέσα στο άθροισμα από j=1 έως M της εξίσωσης (41) περιέχει όρους στον παρανομαστή ανεξάρτητους από το j, πράγμα το οποίο σημαίνει ότι ο παρονομαστής δεν επηρεάζει την τιμή του αθροίσματος και ότι μπορεί να βγει έξω από το άθροισμα. Ο αριθμητής τώρα μπορεί να φραχθεί σε μια μέγιστη τιμή υπό τον όρο ότι τα δεδομένα του πίνακα g_{ij} είναι κανονικοποιημένα σε ένα διάστημα [-α,α]. Σε αυτή την περίπτωση η μέγιστη πιθανή τιμή του αριθμητή είναι 2α. Μπορούμε λοιπόν να θεωρήσουμε ότι η (41) μπορεί να μετασχηματιστεί στην:

$$w_i(t+1) = w_i(t) + \dot{\mu} \sum_{j=1}^{M} e_j \dot{f}(u_j) g_{ij}$$

όπου $$\dot{\mu} = \mu \frac{2\alpha\, M}{\sigma_+(g_i)+\sigma_-(g_i)}$$

και $\dot{\mu}$ είναι ο μέγιστος ρυθμός εκμάθησης που μπορεί να συναντήσουμε.

Όσον αφορά τα δικά μας δεδομένα, είναι κανονικοποιημένα στο διάστημα

[-3,3] και η μέγιστη τιμή του αριθμητή είναι 4,012 για τα 78 δείγματά μας. Όπως προειπόθηκε παραπάνω τα e_j και $\dot{f}(u_j)$ μπορεί να πάρουν πολύ μικρές τιμές καταλήγοντας σε πολύ μικρές τιμές του βάρους, το οποίο οδηγεί με τη σειρά του σε μικρές τιμές της συνάρτησης σφάλματος, πράγμα το οποίο καθυστερεί τη σύγκλιση. Για αυτό το λόγο παίρνουμε στην εξίσωση (45) το πρόσημο της gradient descent ώστε να επιταχύνουμε τη σύγκλιση πηγαίνοντας προς το ελάχιστο με μεγαλύτερα βήματα, τουλάχιστον όταν η διαδικασία βρίσκεται ακόμα μακριά από το ελάχιστο. Η μικρή διαστασιμότητα μπορεί να επιβραδύνει την σύγκλση. Καθώς η διαδικασία προχωράει και το πρόβλημα της διαστασιμότητας μειώνεται σημαντικά ο λόγος των δειγμάτων (ασθενείς) προς τα χαρακτηριστικά (γονίδια) αυξάνεται και το πρόβλημα της εκτίμησης του διαχωριστικού υπερεπιπέδου γίνεται δυσκολότερο, επιβραδύνοντας τη σύγκλιση. Αυτό κάνει αναγκαία την αύξηση είτε των εποχών, είτε του ρυθμού εκμάθησης. Σε αυτά τα τελευταία βήματα η εξίσωση (46) μπορεί να χρησιμοποιηθεί για να επιταχυνθεί η σύγκλιση χρησιμοποιώντας ένα μεταβλητό ρυθμό εκμάθησης $|d_j - y_j|$. Πολύ απλά κατανοεί κανείς ότι όσο είμαστε μακριά από το στόχο το $|d_j - y_j|$ παίρνει μεγάλη τιμή επιταχύνοντας τη σύγκλιση, ενώ όσο πλησιάζουμε το στόχο, το $|d_j - y_j|$ αρχίζει να παίρνει χαμηλότερες τιμές επιβραδύνοντας έτσι τη σύγκλιση. Με άλλα λόγια στα τελευταία βήματα της διαδικασίας επιλογής χαρακτηριστικών (γονιδίων) κατευθυνόμαστε προς το στόχο γρήγορα όσο

ακόμα βρισκόμαστε μακριά από αυτόν, ενώ επιβραδύνουμε όσο τον πλησιάζουμε για καλύτερο συντονισμό του διαχωριστικού υπερεπιπέδου.

Οι εξισώσεις (45) και (46) είναι οι τελικοί κανόνες που χρησιμοποιούνται για την αναπροσαρμογή του βάρους, ενώ οι εξισώσεις (43) και (44) χρησιμοποιήθηκαν για επεξήγηση της τελικής επιλογής.

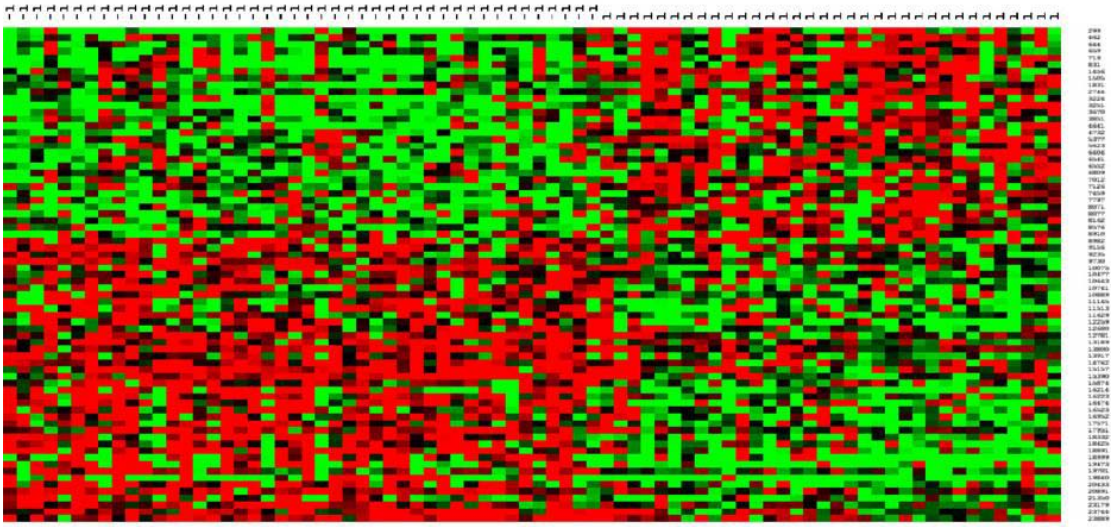
Σαν συμπερασματική παρατήρηση της ενότητας αυτής σημειώνεται ότι εκπαιδεύοντας ένα νευρώνα με μια κατάλληλη διαδικασία εκμάθησης μπορούμε τελικά να εφαρμόσουμε ένα κριτήριο φιλτραρίσματος, όπως το Fisher`s ratio, σε μια μέθοδο wrapper. Αυτό γίνεται αλλάζοντας τον αλγόριθμο gradient descent, ο οποίος έχει αποδειχτεί ότι συγκλίνει, πολλαπλασιάζοντας με το fisher`s metric, παίρνοντας δηλαδή από την εξίσωση

$$w_i(t+1) = w_i(t) - \left(\mu \frac{\partial E}{\partial w_i}\right)$$

την $$w_i(t+1) = w_i(t) - \left(\mu \frac{\partial E}{\partial w_i}\right) \frac{\sum_{j=1}^{M} |g_{ij} - c(g_i)|}{\sigma_+(g_i) + \sigma_-(g_i)}$$

3.4.4 Εφαρμογή στην επιλογή διαφορετικά εκπεφρασμένων γονιδίων

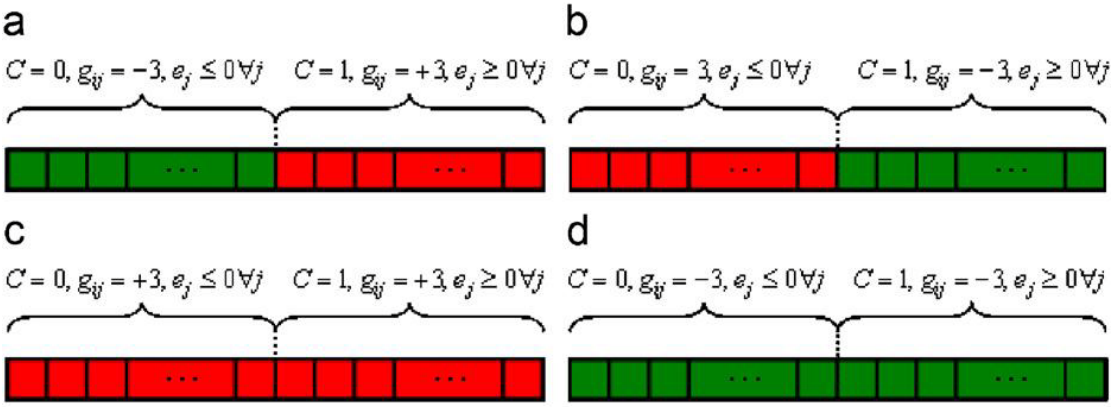
Θα μπορούσαμε να απεικονίσουμε την έκφραση των γονιδίων για κάθε κατάσταση σε μια έγχρωμη παλέτα στην οποία κάθε σειρά αναπαριστά ένα γονίδιο και κάθε στήλη ένα δείγμα (ασθενή), όπως ακριβώς παρακάτω, στο Σχήμα 10.



Σχήμα 10. Χρωματική απεικόνιση της έκφρασης των γονιδίων που μελετάμε για κάθε μια από τις δύο καταστάσεις.

Εδώ έχει ακολουθηθεί απεικόνιση σε πράσινο-κόκκινο-μαύρο. Ένα κελί χρώματος κόκκινου υποδεικνύει ότι το συγκεκριμένο γονίδιο εκφράζεται περισσότερο στην παθολογική κατάσταση, απ` ότι στην κανονική. Ένα πράσινο κελί υποδεικνύει ακριβώς το αντίθετο, ενώ ένα μαύρο κελί υποδεικνύει ότι το συγκεκριμένο γονίδιο εκφράζεται ακριβώς με τον ίδιο τρόπο και στις 2 καταστάσεις. Τα χρώματα μπορούν να μετασχηματιστούν και λογαριθμικά σε ένα κλειστό διάστημα, για παράδειγμα στο [-1,1], όπου τα -1, 0 και +1 εκφράζουν τα χρώματα πράσινο, μαύρο, κόκκινο αντίστοιχα.

Ο ταξινομητής RFE-LNW μπορεί να επιλέξει τα γονίδια που εκφράζονται διαφορετικά στις δύο καταστάσεις που μας ενδιαφέρουν ανάμεσα από τον μεγάλο αριθμό αρχικών γονιδίων. Το Σχήμα 11 δείχνει την έκφραση ενός υποτιθέμενου γονιδίου g_i στην αρνητική (C=0) και στη θετική (C=1) κλάση αντίστοιχα.



Σχήμα 11. Γονίδια που εκφράζονται με τον ίδιο τρόπο και γονίδια που εκφράζονται διαφορετικά στις δύο κλάσεις.

Στις περιπτώσεις (a) και (b) το υποτιθέμενο γονίδιο εκφράζεται διαφορετικά στις δύο περιπτώσεις που μας ενδιαφέρουν και απεικονίζεται με πράσινο (αρνητικές τιμές) στην αρνητική κλάση και κόκκινο (θετικές τιμές) στη θετική κλάση αντίστοιχα. Στις άλλες δύο περιπτώσεις (c) και (d) δεν παρατηρείται διαφοροποίηση στην έκφραση του γονιδίου στις δύο κλάσεις. Συνδυάζοντας την περίπτωση (a) και την εξίσωση (44) και επικεντρώνοντας την προσοχή μας στην αρνητική κλάση (πράσινο χρώμα), παρατηρούμε ότι ο όρος $sign\left(e_j \dot{f}(u_j)\right) sign(g_{ij}) \geq 0$ ισχύει. Πράγματι $e_j \leq 0$ (αφού $d_j = 0$) (42) και $y_i \epsilon [0 \dots 1]$. Το $\dot{f}(u_j)$ από την εξίσωση (39) είναι θετικό και το g_{ij}= -1. Εστιάζοντας στη θετική κλάση (κόκκινο τμήμα) του Σχήματος 12 (a) χρησιμοποιώντας την ίδια αιτιολόγηση παρατηρούμε πάλι ότι $\left(e_j \dot{f}(u_j)\right) sign(g_{ij}) \geq 0$. Εφόσον ο όρος e_j είναι συνήθως μη μηδενικός, ο όρος άθροισης στην εξίσωση (44) παράγει θετικό αποτέλεσμα. Ακολουθώντας πάλι την ίδια λογική μπορεί εύκολα να αποδειχτεί ότι στην περίπτωση (b) του σχήματος η εξίσωση (44) παράγει αρνητικό αποτέλεσμα, και ότι ο όρος άθροισης της ίδιας εξίσωσης στις περιπτώσεις (c) και (d) παράγει αποτελέσματα κοντά στο μηδέν αφού οι όροι αλληλοεξουδετερώνονται. Βλέπουμε επίσης ότι τα διαφορετικά εκπεφρασμένα γονίδια (περιπτώσεις(a) και (b)) παίρνουν μεγαλύτερες απόλυτες τιμές βάρους απ` ότι τα γονίδια που εκφράζονται με τον ίδιο τρόπο (περιπτώσεις (c) κ αι (d)). Από την εξίσωση (43) δεν μπορούμε να το

πάρουμε αυτό το αποτέλεσμα, αφού εξαρτάται από την τιμή και όχι το πρόσημο του όρου e_j, ο οποίος υπάρχει η πιθανότητα να είναι πολύ μικρός και να αλλοιώνει το αποτέλεσμα.

Η Εξίσωση (44) πάλι παίρνοντας μόνο το πρόσημο, αποδίδει ίδιες τιμές βάρους σε γονίδια που εκφράζονται πολύ διαφορετικά και σε γονίδια που εκφράζονται λιγότερο διαφορετικά. Είναι σωστό τα πρώτα να πάρουν μεγαλύτερο βάρος από τα δεύτερα, πράγμα το οποίο επιτυγχάνεται με την εξίσωση (45) στην οποία εισάγεται ο όρος $f_2(g_i)$.

3.4.5 Incremental learning

Οι εξισώσεις (43) - (46) αναπροσαρμόζουν τον όρο του βάρους $w(t+1)$ αφού πρώτα έχουν εισαχθεί όλα τα παραδείγματα στο δίκτυο, δηλαδή αφού έχουν εκτιμηθεί οι όροι άθροισης. Αυτό στη θεωρία νευρωνικών δικτύων αναφέρεται ως batch learning. Εναλλακτικά τα βάρη μπορούν να αναπροσαρμοστούν αυξητικά (incrementally) λαμβάνοντας υπ` όψιν ένα δείγμα τη φορά. Σε αυτή την περίπτωση οι όροι άθροισης μπορούν να εξαλειφθούν. Έτσι οι εξισώσεις (43) – (46) μπορούν να μετασχηματιστούν στις:

$$w_i(t+1) = w(t) + \mu e \dot{f}(u) g_i \quad (47)$$

$$w_i(t+1) = w(t) + \mu \, sign\left(e \, \dot{f}(u)\right) sign(g_i) \quad (48)$$

$$w_i(t+1) = w(t) + \mu \, sign\left(e \, \dot{f}(u)\right) sign(g_i) f_2(g_i) \quad (49)$$

$$w_i(t+1) = w(t) + |d - y| sign(e \, \dot{f}(u)) sign(g_i) f_2(g_i) \quad (50)$$

Έχει αποδειχτεί ότι για το συγκεκριμένο τομέα επιτυγχάνονται καλύτερα αποτελέσματα με τη χρήση incremental learning. Να σημειωθεί ότι η εξίσωση

(50) χρησιμοποιήθηκε από το σημείο που είχαν απομείνει 100 γονίδια μέχρι το πέρας της διαδικασίας. Η σύγκλιση αποδεικνύεται στην παράγραφο 3.4.3.1. Η μόνη διαφορά αυτών των εξισώσεων είναι ότι τώρα λαμβάνεται υπ` όψιν ένα δείγμα τη φορά.

3.4.6 Ο αλγόριθμος του ταξινομητή RFE-LNW

Ο Πίνακας 1 συνοψίζει τον αλγόριθμο του ταξινομητή RFE-LNW

Πίνακας 1.
Ο αλγόριθμος του ταξινομητή RFE-LNW

1. M είναι ο αρχικός αριθμός γονιδίων
2. Όσο (M≥0)
3. Αναπροσαρμογή του διανύσματος των βαρών w χρησιμοποιώντας τις εξισώσεις (49) και (50). (Σε αυτή την εργασία η εξίσωση (49) χρησιμοποιήθηκε όσο ο αριθμός των εναπομείναντων γονιδίων ήταν μεγαλύτερος από 100, ενώ από εκείνο το σημείο και μετά χρησιμοποιήθηκε η εξίσωση (50))
4. Κατάταξη των γονιδίων σύμφωνα με τις απόλυτες τιμές τους στο διάνυσμα των βαρών w.
5. Απομάκρυνση των γονιδίων με τη μικρότερη απόλυτη τιμή βάρους(M←M - 1). Σε κάθε επανάληψη μπορεί να αφαιρεθούν περισσότερα του ενός γονίδια.
6. Εκτίμηση της ακρίβειας ταξινόμησης των M εναπομείναντων γονιδίων με τη χρήση ενός γραμμικού SVM ταξινομητή.
7. Τέλος βρόχου.
8. Ανάδειξη ως γονίδια δείκτες το σετ των εναπομείναντων γονιδίων τα οποία επιτυγχάνουν την καλύτερη ακρίβεια ταξινόμησης.

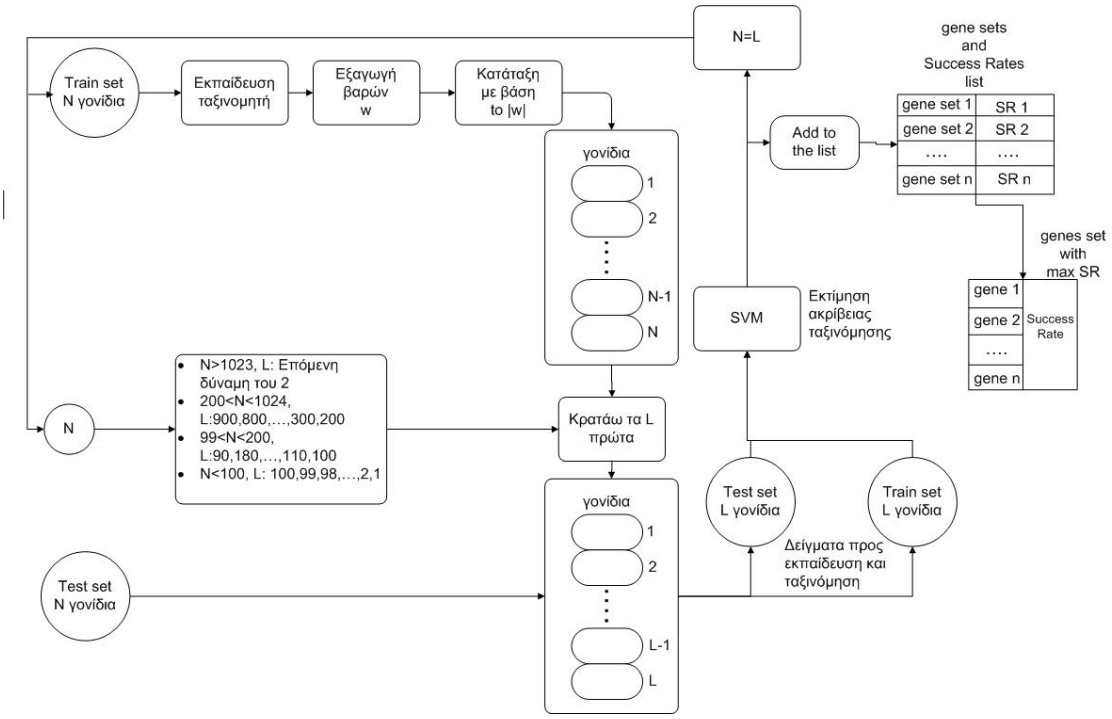
Πρέπει να επισημανθεί ότι εκτός από το πλεονέκτημα της χρήσης ενός νευρώνα ως τη μόνη μονάδα εκμάθησης, καταφέρνει να εφαρμόσει filter

criteria με πραγματικά wrapper τρόπο, όπου τα βάρη επανεκτιμούνται και πιθανώς προσαρμόζονται από επανάληψη σε επανάληψη. Εξαλείφοντας γονίδια μειώνεται τη διαστασιμότητα του προβλήματος και έτσι το νέο υπερεπίπεδο επανεκτιμάται σε ένα νέο χώρο, ο οποίος είναι μειωμένων διαστάσεων από τον προηγούμενο, με ένα νέο direction vector w. Τα βάρη των γονιδίων αλλάζουν τιμή από επανάληψη σε επανάληψη, αφού αρχικά τα πολλά γονίδια που εισάγουμε συνεπάγονται την ύπαρξη κάποιων, και μάλιστα πολλών, τα οποία δίνουν μη χρήσιμη πληροφορία και επισκιάζουν την σημασία και την επίδραση αυτών που δίνουν χρήσιμη πληροφορία. Τα γονίδια που μας δίνουν χρήσιμη πληροφορία φαίνονται όσο προχωράει η διαδικασία και το πρόβλημα της διαστασιμότητας μειώνεται. Αξίζει να παρατηρηθεί η διαφορά με τη filter μέθοδο όπου το fisher metric των εναπομείναντων γονιδίων παραμένει σταθερό κατά τη διάρκεια της διαδικασίας επιλογής γνωρισμάτων (γονιδίων), όπως ακριβώς και το διαχωριστικό όριο των κλάσεων.

3.4.7 Τυπική διαδικασία επιλογής γονιδίων

Η αφετηρία μας είναι το αρχικό σύνολο γονιδίων. Χρησιμοποιείται το train set για την αξιολόγηση του ποια γονίδια θα αφαιρεθούν. Η διαδικασία επιλογής γονιδίων είναι μια επαναληπτική διαδικασία, σε κάθε βήμα της οποίας αφαιρείται ένα πλήθος γονιδίων, μέχρι να μειώσουμε στο 1 γονίδιο.

Σε κάθε στάδιο αυτής της επαναληπτικής διαδικασίας, είναι αναγκαίο να εκτιμηθεί η απόδοση του ταξινομητή μας (RFE-LNW) χρησιμοποιώντας μόνο τα εναπομείναντα γονίδια. Έτσι θα μπορούν να προκύψουν συμπεράσματα για την πληροφορία που κρύβεται σε αυτά τα γονίδια, σχετικά με το πρόβλημα της ταξινόμησης. Αυτό γίνεται με αξιολόγηση μέσω ενός ανεξάρτητου σετ γονιδίων (independent test set evaluation).Τελικά σε κάθε επανάληψη της διαδικασίας, υπολογίζονται τα νέα train και test sets, αφαιρώντας από κάθε δείγμα τους τις εκφράσεις των γονιδίων που απορρίπτονται. Με αυτά τα νέα σύνολα εκπαίδευσης και δοκιμής, γίνεται εκπαίδευση του ταξινομητή και εκτίμηση της απόδοσής του. Ο αλγόριθμος του ταξινομητή που χρησιμοποιούμε αναπαρίσταται παρακάτω στο Σχήμα 12.



Σχήμα 12. Σχηματική αναπαράσταση ταξινομητή RFE-LNW

Πιο συγκεκριμένα, με το train set εκπαιδεύεται ο ταξινομητής και στη συνέχεια ελέγχεται η απόδοσή του, με έναν SVM τόσο με το train όσο και με το test set. Για την ακρίβεια υπολογίζεται το ποσοστό επιτυχίας και όχι το ποσοστό σφάλματος. Έτσι ελέγχεται κάθε φορά εάν ο εκάστοτε ταξινομητής ανταποκρίνεται ικανοποιητικά τόσο σε δεδομένα γνωστά σε αυτόν (έλεγχος απόδοσης με το train set) όσο και σε άγνωστα σε αυτόν δεδομένα (έλεγχος απόδοσης με το independent Test set), προκειμένου να διαπιστωθεί η ικανότητα γενίκευσης.

Για κάθε ένα από τα γονίδια εκτιμάται το κατά πόσο αυτό ατομικά συνεισφέρει στον διαχωρισμό των κλάσεων, το πόσο δηλαδή σημαντικό αυτό είναι. Υπολογίζεται λοιπόν κάποιο βάρος για κάθε γονίδιο. Όσο μεγαλύτερη είναι η απόλυτη τιμή αυτού, τόσο πιο σημαντικό θεωρείται το γονίδιο για την απόφαση ταξινόμησης. Κάθε βάρος υπολογίζεται με πληροφορία για το κάθε γονίδιο και δε λαμβάνεται υπόψη η αμοιβαία πληροφορία που πιθανόν να υπάρχει ανάμεσα στα γονίδια. Ο υπολογισμός του βάρους γίνεται με τις εξισώσεις (49) και (50). Οπότε εξάγεται ένα διάνυσμα βαρών w διάστασης N – όσα είναι και τα γονίδια. Άρα σε κάθε γονίδιο g_i αντιστοιχεί ένα βάρος w_i.

Χρησιμοποιούμε όπως είπαμε ως κριτήριο ταξινόμησης την απόλυτη τιμή του βάρους και τελικά τα γονίδια ταξινομούνται ξεκινώντας από αυτό που έχει το μεγαλύτερο $|w_i|$ και καταλήγοντας σε αυτό με το μικρότερο $|w_i|$.

Έχοντας κατατάξει λοιπόν τα γονίδια από το πιο σημαντικό στο λιγότερο σημαντικό, το επόμενο βήμα είναι η αφαίρεση ενός πλήθους λιγότερο σημαντικών γονιδίων. Κάθε φορά που μειώνονται τα γονίδια πρέπει να υπολογίζονται νέα βάρη. Σε κάθε επανάληψη της μεθόδου υπολογίζονται τα νέα train και test sets, τα οποία αποτελούνται από τα γονίδια του προηγούμενου train και test set έχοντας αφαιρέσει τα γονίδια που ήταν στις χαμηλότερες θέσεις ταξινόμησης αντίστοιχα. Το νέο train set που προκύπτει χρησιμοποιείται για εκπαίδευση του ταξινομητή και για εξαγωγή των νέων βαρών για τα εναπομείναντα γονίδια. Η διαδικασία αυτή είναι επαναληπτική και συνεχίζεται μέχρις ότου φτάσουμε στο 1 γονίδιο. Στο τέλος έχουμε στην διάθεσή μας ένα πλήθος συνόλων γονιδίων, καθένα από τα οποία είναι υποσύνολο των μεγαλυτέρων του και γενικά υποσύνολο των αρχικών γονιδίων.

Στη συγκεκριμένη διπλωματική ακολουθούμε την εξής διαδικασία για την αφαίρεση γονιδίων. Αρχικά (ενώ ακόμα έχουμε πολλά γονίδια) αφαιρούμε τόσα ώστε να φτάσουμε στη αμέσως χαμηλότερη δύναμη του δύο. Αυτό το κάνουμε έως ότου έχουν απομείνει 1024 γονίδια. Από εκείνο το σημείο και μετά αφαιρούμε κάθε φορά 100 γονίδια μέχρι να φτάσουμε στα 200. Κατόπιν συνεχίζουμε αφαιρώντας 10 γονίδια τη φορά μέχρι να μας έχουν απομείνει 100. Από εκείνο το σημείο και μέχρι το πέρας της διαδικασίας, 1 εναπομείναν γονίδιο, συνεχίζουμε αφαιρώντας 1 γονίδιο τη φορά. Ο Πίνακας 2 περιγράφει με πόσα γονίδια στο train και στο test set θα ξεκινάει κάθε επανάληψη.

Πίνακας 2.

Αριθμός γονιδίων των train και test set στην αρχή κάθε επανάληψης

1	2	…	6	7	8		14	15	16		23	24	25		122
24188	2^{14}	…	2^{10}	900	800	…	200	190	180	…	100	99	98	…	1

Ο ταξινομητής RFE-LNW μετά το τέλος των επαναλήψεων έχει 122 σύνολα γονιδίων για τα οποία έχει υπολογίσει μέσω του test set το ποσοστό επιτυχίας ταξινόμησης (Success rate). Από αυτά λοιπόν τα σύνολα γονιδίων επιλέγει αυτό με την υψηλότερη ακρίβεια ταξινόμησης.

Κεφάλαιο 4

Μεθοδολογία υλοποίησης

Στο προηγούμενο κεφάλαιο παρουσιάστηκε η εισαγωγή στην αναγνώριση προτύπων και η θεωρητική ανάλυση των όσων χρησιμοποιούνται σε αυτή τη διπλωματική εργασία. Στη συγκεκριμένη περίπτωση τα πρότυπα που πρόκειται να ταξινομηθούν είναι οι ασθενείς που έχουν στο παρελθόν παρουσιάσει καρκίνο του μαστού, τα χαρακτηριστικά τους είναι οι σχετικές εκφράσεις των γονιδίων και πρόκειται να γίνει ταξινόμηση σε δύο κλάσεις, ανάλογα με τη σχετική έκφραση των γονιδίων. Οι δύο αυτές κλάσεις είναι αφενός οι ασθενείς που πρόκειται να παρουσιάσουν μετάσταση καρκίνου και αφετέρου οι ασθενείς που δεν πρόκειται να παρουσιάσουν μετάσταση μέσα σε πέντε χρόνια.

Σκοπός μας είναι να ελαχιστοποιήσουμε τον αριθμό των γονιδίων, βάσει των οποίων θα μπορούμε με ψηλό ποσοστό επιτυχίας να κρίνουμε αν η ασθενής πρόκειται να εμφανίσει μετάσταση καρκίνου ή όχι. Χρησιμοποιούμε τον ταξινομητή RFE-LNW. Αυτός εκπαιδεύεται βάσει ενός train set και στη συνέχεια ελέγχεται βάσει ενός test set. Βρίσκεται ποια είναι τα πιο σημαντικά γονίδια, δηλαδή ποια δίνουν την πιο χρήσιμη πληροφορία και στη συνέχεια αυτά αφαιρούνται από τα train και test set. Έτσι δημιουργούμε νέα train και test sets από τα παλιά σύνολα γονιδίων, τα οποία περιέχουν μόνο τα εναπομείναντα γονίδια, και επαναλαμβάνουμε τη διαδικασία με μειωμένο αριθμό γονιδίων με σκοπό να αφαιρέσουμε όλα τα γονίδια ως σημαντικά. Κατόπιν από τα γονίδια που είχαμε αφαιρέσει κρατάμε αυτά που μας έδιναν καλό ποσοστό επιτυχίας και δημιουργούμε πάλι νέα train και test sets και επαναλαμβάνουμε αυτή τη διαδικασία. Έτσι ενώ είχαμε ξεκινήσει με 24188 γονίδια, καταλήγουμε σε 97, τα οποία μας δίνουν τη δυνατότητα να κρίνουμε με μεγάλη επιτυχία αν η ασθενής πρόκειται στα επόμενα πέντε χρόνια να παρουσιάσει μετάσταση.

Εδώ πρέπει να σημειωθεί ότι δε γίνεται cross validation. Ξεκινάμε με δοθέντα train και test sets και από αυτά αφού ελέγξουμε και βρούμε τα ποσοστά επιτυχίας, αφαιρούμε γονίδια με σκοπό να καταλήξουμε σε μια γονιδιακή υπογραφή με τον ελάχιστο αριθμό γονιδίων

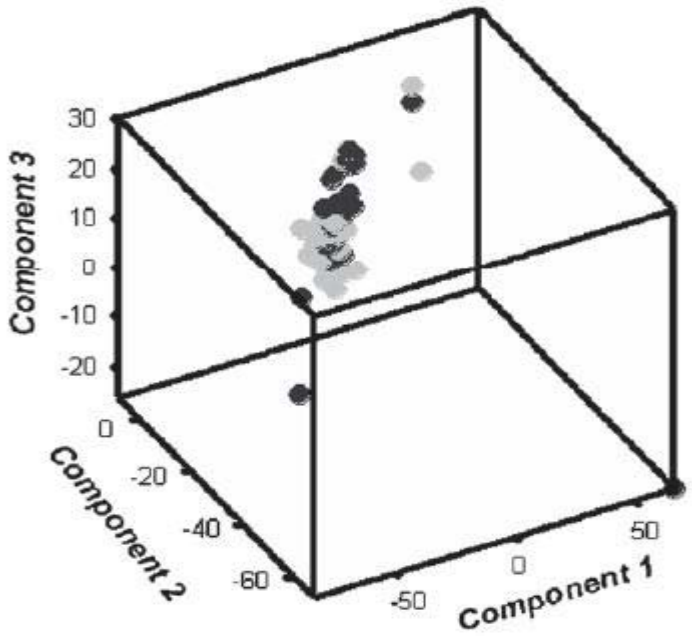
Αρχικά γίνεται μια περιγραφή των δεδομένων του καρκίνου του μαστού και στη συνέχεια περιγράφεται η μεθοδολογία που ακολουθείται σε αυτή τη διπλωματική εργασία.

4.1 Περιγραφή των δεδομένων

Τα δεδομένα του καρκίνου του μαστού τα οποία εξετάζονται στην παρούσα διπλωματική εργασία, προέρχονται από τη δημοσιεύσεις του Van't Veer [6]. Είναι στη διάθεσή μας ένα σύνολο εκπαίδευσης (train set) και ένα ανεξάρτητο σύνολο δοκιμής (independent test set).

Για τα δεδομένα του καρκίνου του μαστού έχουμε συνολικά 24188 γονίδια και το train set αποτελείται από 78 δείγματα, 44 εκ των οποίων χαρακτηρίζονται αρνητικά και αντιστοιχούν στους ασθενείς που παραμένουν υγιείς για διάστημα τουλάχιστον πέντε ετών, ενώ τα υπόλοιπα 34 δείγματα χαρακτηρίζονται θετικά και αντιστοιχούν στους ασθενείς που εμφανίζουν μετάσταση μέσα σε διάστημα πέντε ετών. Το test set αποτελείται από 19 δείγματα, 7 αρνητικά και 12 θετικά.

Προκειμένου να μπορέσει να γίνει μια πρώτη σύγκριση των δεδομένων μας, εφαρμόζεται σε αυτά μείωση στις τρεις διαστάσεις μέσω PCA μεθοδολογίας για να βγει κάποιο συμπέρασμα για την κατανομή των δεδομένων στο χώρο. Το αποτέλεσμα φαίνεται στο Σχήμα 13.



Σχήμα 13. Απεικόνιση των τριών principal components των του καρκίνου του μαστού.

Παρατηρούμε υπάρχει μεγάλη επικάλυψη ανάμεσα στις δύο κλάσεις και μάλιστα η διασπορά τους είναι μικρή με αποτέλεσμα να είναι δύσκολο να διαχωριστούν μεταξύ τους.

Συνεπώς, το πρόβλημα της επιλογής γονιδίων είναι πολύ ενδιαφέρον και κρίσιμο ούτως ώστε να βρεθεί ένα καλύτερο υποσύνολο γονιδίων, βάση του οποίου να γίνεται καλός διαχωρισμός των δύο κλάσεων.

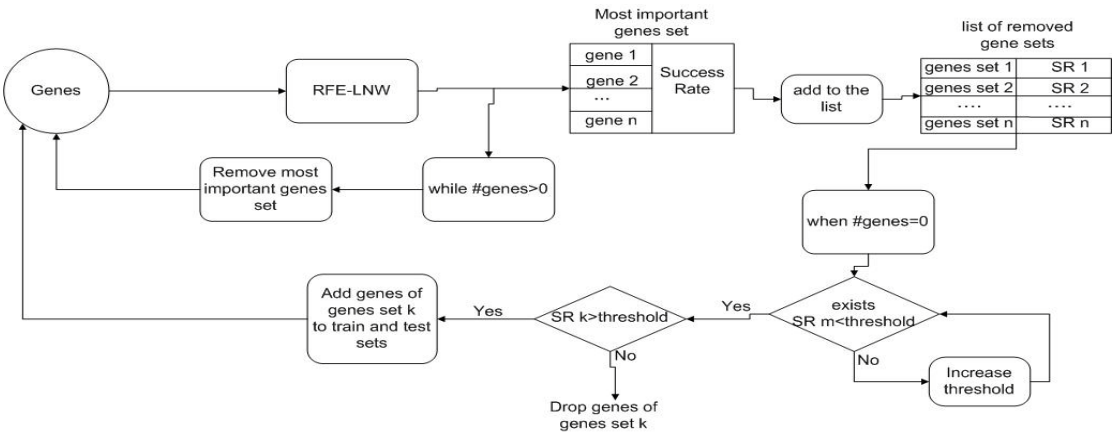
Κάθε πρότυπο – ασθενής είναι ένα διάνυσμα που αποτελείται από *N* χαρακτηριστικά – εκφράσεις γονιδίων. Έχουμε ταξινόμηση σε δύο κλάσεις οπότε σε κάθε πρότυπο x_i αντιστοιχεί μια ετικέτα κλάσης $y_i \epsilon$ [-1,+1]. Το κάθε γονίδιο θα συμβολίζεται με g_i, i*=1,2,…,N*. Οπότε η έκφραση του γονιδίου g_i, για τον ασθενή x_j, j*=1,2,…,M* θα είναι το *i*-οστό στοιχείο του διανύσματος x_j, δηλαδή το $(x_j)_i$.

4.2 Μεθοδολογία επιλογής γονιδίων

Το πλήθος των γονιδιακών εκφράσεων που προκύπτουν με τα πειράματα Microarrays είναι πολύ μεγάλο και όπως έχουμε ήδη εξηγήσει, σε τέτοιες πολυδιάστατες περιπτώσεις είναι αναγκαία η μείωση των διαστάσεων ούτως ώστε να καταλήξουμε σε αποδοτικότερα μοντέλα ταξινόμησης. Επιπλέον, στην κατασκευή διαγνωστικών τεστ είναι σημαντική από πρακτική σκοπιά η ικανότητα επιλογής μικρών υποσυνόλων γονιδίων [25]. Η μελέτη λίγων γονιδίων μπορεί να βοηθήσει τους βιολόγους να αποκτήσουν σημαντική γνώση των μηχανισμών που είναι υπεύθυνοι για μία συγκεκριμένη ασθένεια, κάτι το οποίο μπορεί να οδηγήσει σε ανακάλυψη της κατάλληλης θεραπείας και σε έγκαιρη διάγνωση [39].

Σκοπός αυτής της διπλωματικής είναι χρησιμοποιώντας τον ταξινομητή RFE-LNW να καταλήξουμε σε μια γονιδιακή υπογραφή με τον ελάχιστο αριθμό γονιδίων τα οποία μας δίνουν χρήσιμη πληροφορία για το αν πρόκειται η ασθενής να παρουσιάσει μετάσταση μέσα στα επόμενα 5 χρόνια ή όχι.

Το επόμενο σχήμα παρουσιάζει απλουστευμένα τη διαδικασία την οποία ακολουθήσαμε. Στη συνέχεια αναλύουμε λεπτομερώς τη μεθοδολογία που ακολουθήσαμε για την επιλογή των γονιδίων και τελικά παραθέτουμε το Σχήμα 15, το οποίο παρουσιάζει λεπτομερώς τη διαδικασία επιλογής των γονιδίων.



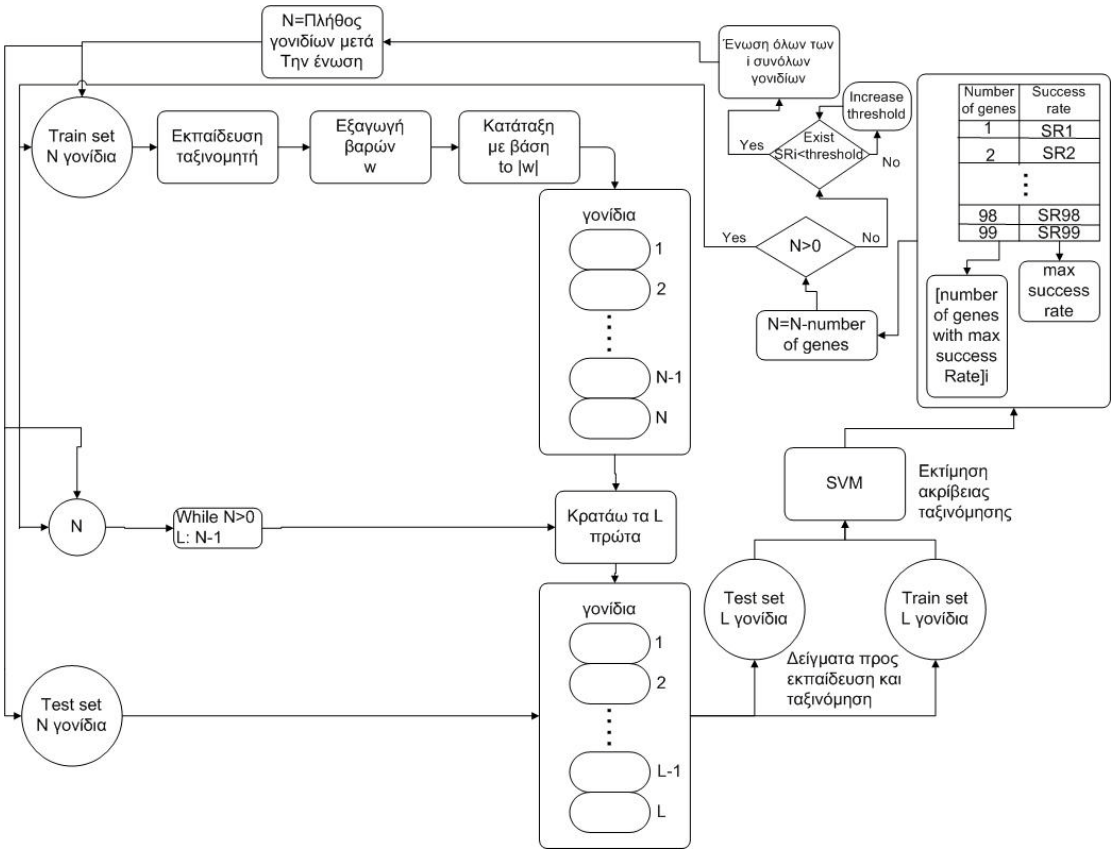
Σχήμα 14. Απλουστευμένη παρουσίαση της διαδικασίας που ακολουθήσαμε για την επιλογή γονιδίων

Σε αυτή την παράγραφο περιγράφουμε τη μεθοδολογία που ακολουθήθηκε για να καταλήξουμε στο ζητούμενο σετ γονιδίων. Αφετηρία μας είναι το αρχικό σετ γονιδίων του Van`t Veer το οποίο έχει περιγραφεί πιο πάνω. Ακολουθούμε την τυπική διαδικασία επιλογής γονιδίων η οποία έχει περιγραφεί στην παράγραφο 3.4.6. Ο RFE-LNW όπως είπαμε στο τέλος βρίσκει ποιο σύνολο γονιδίων έχει το μέγιστο ποσοστό επιτυχίας (Success Rate) και αυτό το σύνολο επιλέγει κάθε φορά ως marker genes. Αυτά τα γονίδια εμείς τα αφαιρούμε από τα αρχικά train και test sets. Δημιουργούμε νέα train και test sets τα οποία είναι υποσύνολα των αρχικών, αφού δεν περιέχουν τα γονίδια που αφαιρέσαμε. Αφαιρούμε δηλαδή τα γονίδια που εκτιμά ο ταξινομητής μας ως πιο σημαντικά. Αυτό το σετ γονιδίων που αφαιρέσαμε τα κρατάμε κάπου μαζί με το ποσοστό επιτυχίας του γιατί θα το χρειαστούμε σε επόμενη φάση. Συνεχίζουμε λοιπόν με τα νέα – μειωμένα train και test sets και κάνουμε την ίδια ακριβώς διαδικασία. Έτσι μειώνουμε κι άλλο τα train και test sets αφαιρώντας και άλλα γονίδια ως σημαντικά. Σε κάθε τρέξιμο (Run) δηλαδή του κώδικα του ταξινομητή αφαιρούνται κάποια γονίδια (αυτά με το μεγαλύτερο ποσοστό επιτυχίας). Αυτό το κάνουμε έως ότου αφαιρεθούν όλα τα γονίδια από τα train και test sets ως σημαντικά. Στην διαδικασία από την έναρξη με ένα train και test set μέχρι και το σημείο που αυτά αδειάζουν αναφερόμαστε με τον όρο φάση. Καταλαβαίνουμε ότι κάθε φάση έχει έναν αριθμό τρεξιμάτων (Runs). Όπως προειπόθηκε κρατάμε κάπου τα γονίδια που αφαιρούνται μαζί με το ποσοστό επιτυχίας τους. Όταν τα train και test sets αδειάσουν στο τέλος μιας φάσης, είναι το σημείο που θα χρειαστούμε τα γονίδια που είχαν αφαιρεθεί, καθώς και τα ποσοστά επιτυχίας τους. Σε αυτό το σημείο δημιουργούμε πάλι νέα train και test sets, τα οποία αυτή τη φορά αποτελούνται από γονίδια που είχαν αφαιρεθεί στην προηγούμενη φάση. Θα χρησιμοποιήσουμε όμως μόνο γονίδια τα οποία μας δίνουν ποσοστό επιτυχίας από ένα κατώφλι και πάνω. Για να προχωρήσουμε παραδείγματος χάριν στη δεύτερη φάση αποφασίσαμε να κρατήσουμε μόνο τα γονίδια που έδωσαν ποσοστό επιτυχίας από 78 και πάνω. Έτσι ενώσαμε τα εν λόγω γονίδια που είχαν αφαιρεθεί κατά την πρώτη φάση και δημιουργήσαμε νέα train και test sets με τα οποία ξεκίνησε η δεύτερη. Αυτή η διαδικασία επαναλαμβάνεται κάθε φορά για να ξεκινήσουμε την επόμενη φάση. Στην πορεία καταλήξαμε σε κάποια σημεία που δεν υπήρχαν γονίδια

που να μας δίνουν ποσοστό επιτυχίας κάτω από το κατώφλι που είχαμε επιλέξει. Τότε ανεβάζαμε αυτό το κατώτατο όριο βάσει του οποίου κρατούσαμε τα γονίδια που θα αποτελούσαν τα καινούρια μας train και test sets. Η διαδικασία επαναλαμβάνεται μέχρι να καταλήξουμε σε μια γονιδιακή υπογραφή με τον ελάχιστο δυνατό αριθμό γονιδίων. Σκοπός μας είναι να καταλήξουμε σε μια γονιδιακή υπογραφή η οποία θα περιέχει κάτω από 100 γονίδια. Στο Σχήμα 15 αναπαρίσταται λεπτομερώς η διαδικασία που μόλις περιγράφηκε.

Πρέπει να αναφέρουμε ότι τροποποιήσαμε τον ταξινομητή ούτως ώστε να επιλέγει σε κάθε run, ως marker genes, μόνο μεταξύ των συνόλων γονιδίων τα οποία περιέχουν λιγότερα από 100 γονίδια. Δηλαδή από τα 99 τελευταία iterations τα οποία δίνουν σύνολα γονιδίων με πλήθος κάτω από 100 γονίδια το καθένα επιλέγουμε αυτό με το μεγαλύτερο ποσοστό επιτυχίας. Αν το μέγιστο ποσοστό επιτυχίας επετεύχθη από δύο ή περισσότερα από αυτά τα σύνολα γονιδίων, ο ταξινομητής μας επέλεγε ως marker genes αυτό με τα λιγότερα γονίδια. Αυτό το κάνουμε γιατί όπως είπαμε η γονιδιακή υπογραφή στην οποία θέλουμε να καταλήξουμε θέλουμε να περιέχει τον ελάχιστο αριθμό γονιδίων.

Αξίζει επίσης να αναφέρουμε ότι στον ταξινομητή για αυτή τη διπλωματική χρησιμοποιήθηκαν οι τιμές 100 για το ρυθμό εκμάθησης και 3000 εποχές μέχρι το σημείο που απομένουν 100 γονίδια, ενώ χρησιμοποιήθηκε μεταβλητός ρυθμός εκμάθησης, αυτός της εξίσωσης (50), και 200 εποχές από εκείνο το σημείο και μέχρι το πέρας της διαδικασίας.



Σχήμα 15. Σχηματική αναπαράσταση της διαδικασίας που ακολουθήσαμε.

Κεφάλαιο 5

Παρουσίαση αποτελεσμάτων

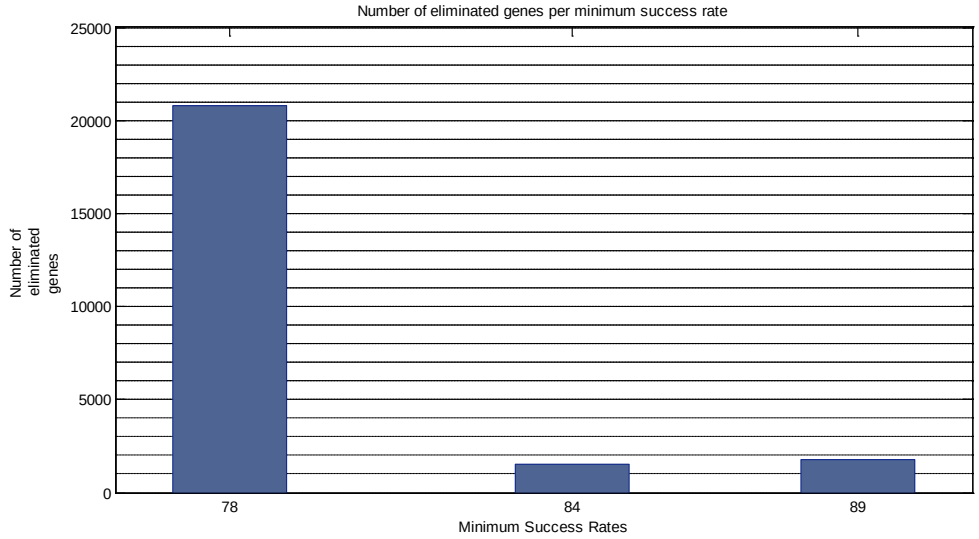
Να υπενθυμίσουμε ότι ξεκινώντας είχαμε στη διάθεσή μας τα δεδομένα του Van`t Veer για τον καρκίνο του μαστού τα οποία αποτελούνταν από 24188 γονίδια, 78 δείγματα για εκπαίδευση και 19 δείγματα για δοκιμή και σκοπός μας ήταν να καταλήξουμε σε μια γονιδιακή υπογραφή με όσο το δυνατόν λιγότερα γονίδια και καλύτερο ποσοστό επιτυχίας.

Ακολουθώντας τη μεθοδολογία επιλογής γονιδίων που περιγράφηκε στην παράγραφο 4.2 αφαιρούσαμε γονίδια. Ξεκινήσαμε στην πρώτη φάση εισάγοντας στον ταξινομητή μας τα 24188 γονίδια που είχαμε αρχικά. Βρήκαμε τα ποσοστά επιτυχίας των συνόλων των γονιδίων και κατόπιν έπρεπε να δημιουργήσουμε νέα train και test sets για να προχωρήσουμε στην επόμενη φάση. Αποφασίσαμε να κρατήσουμε μόνο τα σύνολα των γονιδίων που πέτυχαν ποσοστό επιτυχίας από 78 και πάνω. Όσο κάποιο σετ γονιδίων έδινε κάτω από 78, έστω και αν αυτό αποτελείτο από 1 και μόνο γονίδιο, συνεχίζαμε στην επόμενη φάση κρατώντας το 78 ως το κατώτατο όριο ποσοστού επιτυχίας. Αυτό το όριο το κρατήσαμε μέχρι και τη φάση 24. Σε εκείνο το σημείο κανένα σύνολο γονιδίων δε μας έδωσε ποσοστό επιτυχίας κάτω από 78. Μέχρι εκείνο το σημείο είχαμε απορρίψει το μεγαλύτερο μέρος των γονιδίων. Συγκεκριμένα αφαιρέθηκαν μέχρι εκείνο το σημείο 20788 γονίδια.

Κατόπιν για να μπορέσουμε να συνεχίσουμε έπρεπε να αυξήσουμε το κατώτατο όριο ποσοστού επιτυχίας. Αποφασίσαμε να το αυξήσουμε στο 84. Ακολουθώντας την ίδια διαδικασία περιμέναμε να φτάσουμε είτε σε μια γονιδιακή υπογραφή με πολύ λίγα γονίδια, είτε σε σημείο που κανένα σετ γονιδίων δε θα μας έδινε ποσοστό επιτυχίας κάτω από 84. Στη φάση 32 λοιπόν πετύχαμε το δεύτερο από τους δύο στόχους μας. Το ελάχιστο ποσοστό επιτυχίας που επετεύχθη ήταν 84. Έπρεπε λοιπόν πάλι να αυξήσουμε το κατώτατο όριο ποσοστού επιτυχίας. Αποφασίσαμε να το

ορίσουμε αυτή τη φορά στο 89. Κατά τη διάρκεια που απορρίπταμε γονίδια που έδιναν ποσοστό επιτυχίας κάτω από 84, 1551 γονίδια κρίθηκαν απορριπτέα.
Ξεκινήσαμε τη φάση 33 λοιπόν αφαιρώντας τα γονίδια που έδιναν ποσοστό επιτυχίας κάτω από 89. Στη φάση 61 και ενώ είχαν απομείνει μόνο 97 γονίδια (με 97 ξεκίνησε η φάση) πήραμε ποσοστό επιτυχίας 94 (94.74 για την ακρίβεια), χωρίς να αφαιρεθεί κανένα γονίδιο. Από τη φάση 33 έως και την 61 είχαν αφαιρεθεί 1751 γονίδια.

Παρακάτω στο διάγραμμα 1 παρουσιάζεται για κάθε ένα κατώτατο ποσοστό επιτυχίας που είχαμε θέσει, πόσα γονίδια συνολικά αφαιρέθηκαν. Στον οριζόντιο άξονα παρουσιάζονται τα ποσοστά επιτυχίας, ενώ στον κάθετο ο αριθμός των γονιδίων που αφαιρέθηκαν.

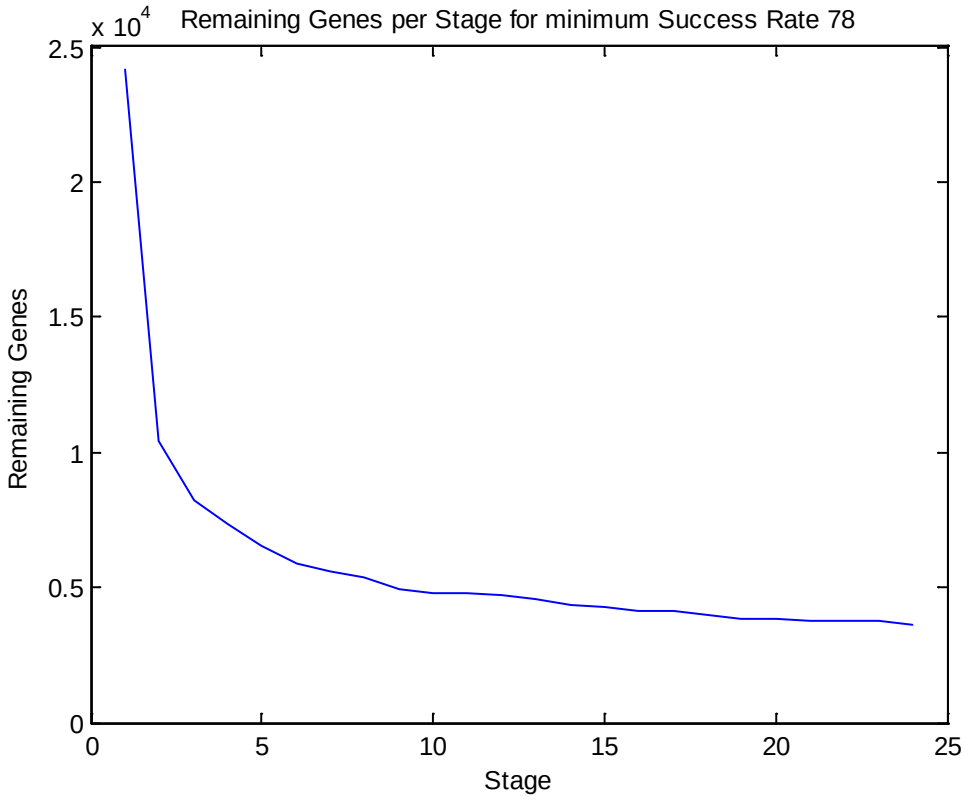


Διάγραμμα 1. Παρουσιάζεται ο αριθμός των γονιδίων που αφαιρέθηκαν (κάθετος άξονας) για κάθε κατώτατο ποσοστό επιτυχίας (οριζόντιος άξονας).

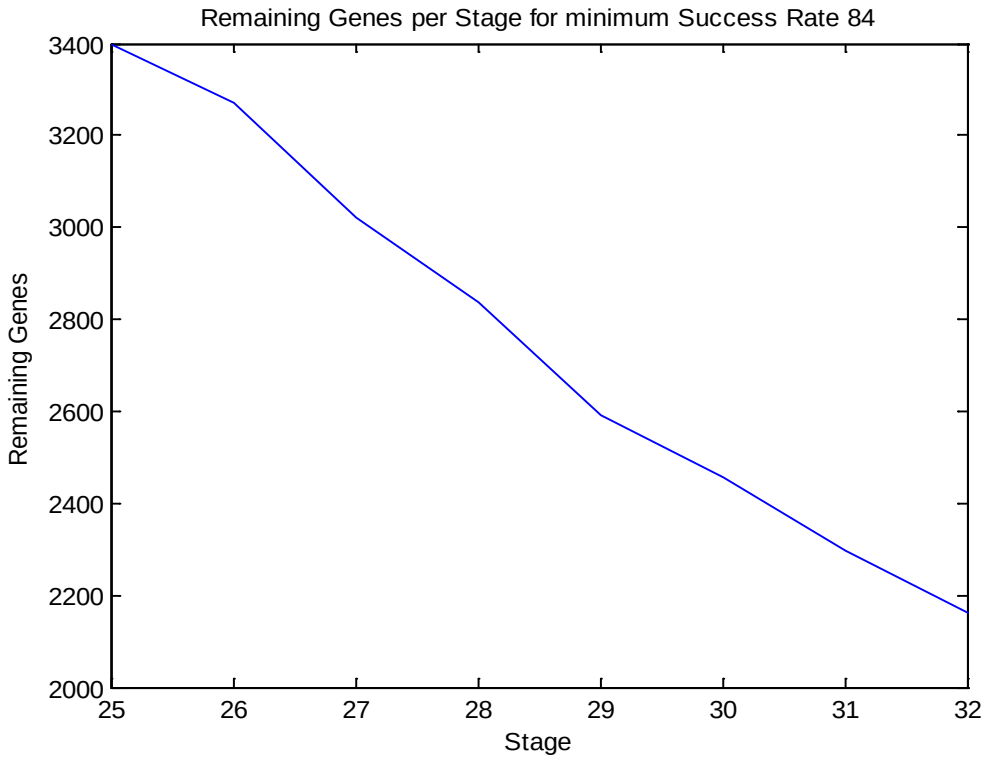
Αξίζει να αναφέρουμε ότι ενώ είχαμε περάσει το σημείο όπου κανένα γονίδιο δεν έδινε κάτω από ένα συγκεκριμένο ποσοστό επιτυχίας και ψάχναμε να βρούμε ποια γονίδια δίνουν πιο κάτω από το επόμενο ποσοστό που είχαμε ορίσει, ώστε να τα αφαιρέσουμε, ορισμένα έδιναν ποσοστό επιτυχίας χαμηλό (πιο χαμηλό και από το πρώτο ποσοστό). Για να εξηγηθεί καλύτερα αυτό, αναφέρεται το επόμενο παράδειγμα. Ενώ είχαμε περάσει το σημείο όπου κανένα γονίδιο δεν έδινε ποσοστό επιτυχίας πιο κάτω από 78, και ψάχναμε

πλέον ποια γονίδια δίνουν ποσοστό κάτω από 84 ώστε να τα αφαιρέσουμε, υπήρξαν γονίδια που έδιναν πιο κάτω από 78 (ακόμα και πολύ χαμηλά). Αυτό το θεωρούμε φυσιολογικό αφού έχοντας αφαιρέσει πολλά γονίδια η εκπαίδευση μπορεί να γίνεται ελλιπέστερη, ή τα γονίδια που είχαν ήδη αφαιρεθεί να έκρυβαν την πραγματική πληροφορία που δίνουν τα γονίδια στα οποία αναφέρεται αυτή η παράγραφος.

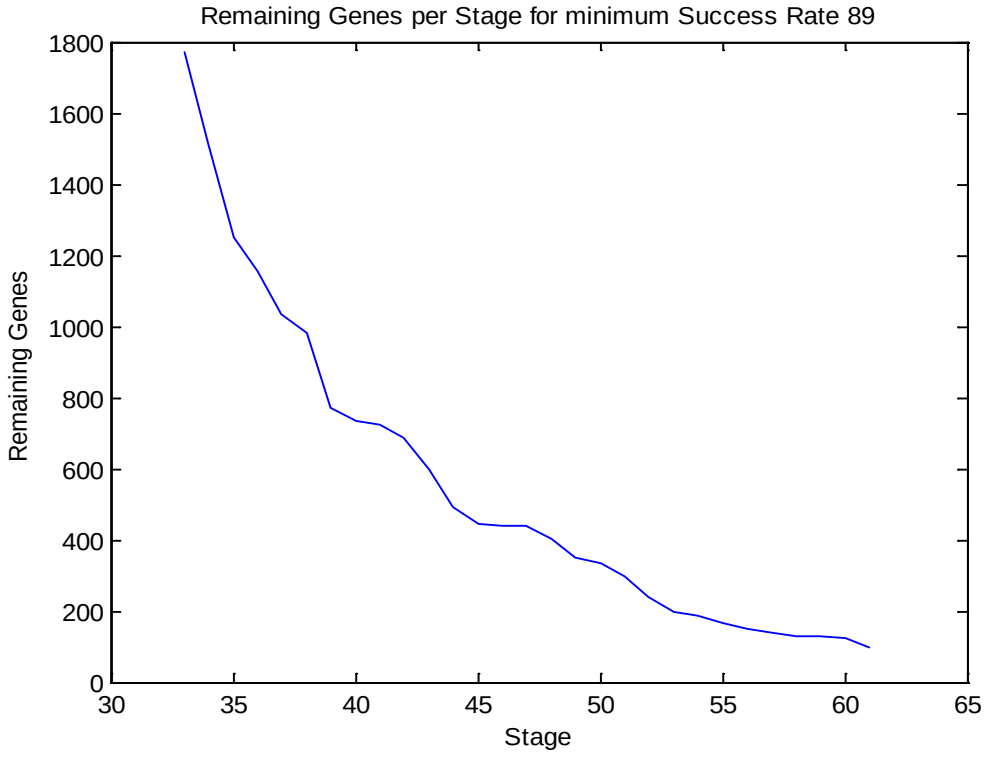
Στα παρακάτω 3 διαγράμματα παρουσιάζεται ο ρυθμός αφαίρεσης των γονιδίων από φάση σε φάση για τα τρία ελάχιστα ποσοστά επιτυχίας 78, 84 και 89 αντίστοιχα.



Διάγραμμα 2. Ο αριθμός γονιδίων με τα οποία ξεκινούσε κάθε μία από τις 24 φάσεις κατά τις οποίες ελέγχαμε για κατώτατο ποσοστό επιτυχίας 78.

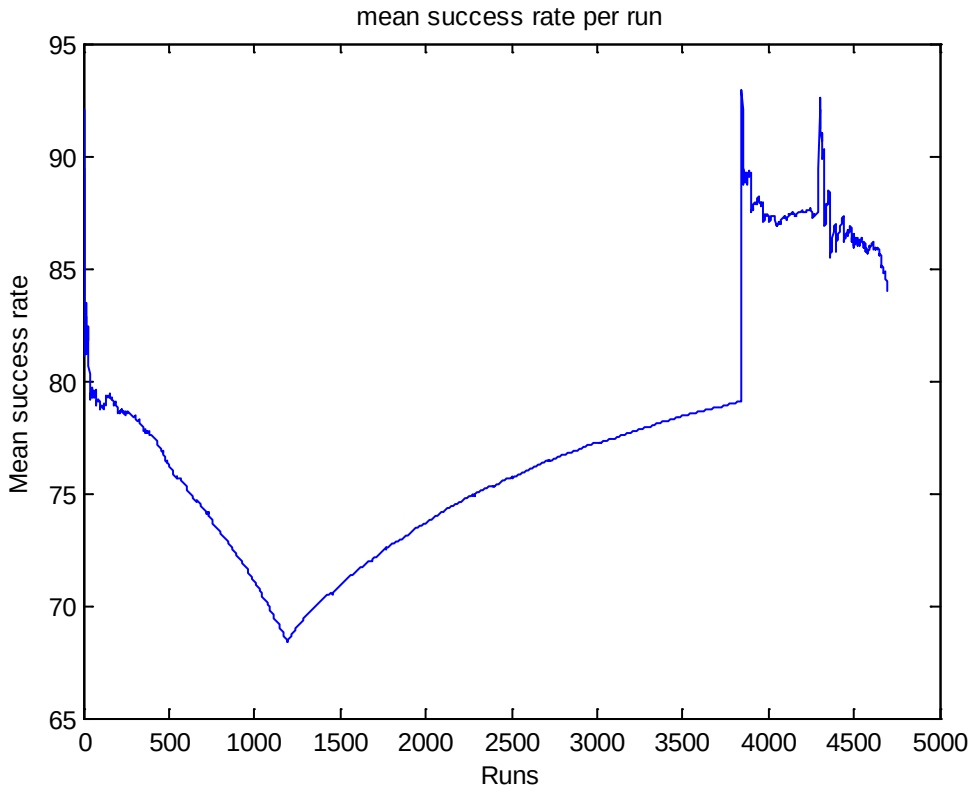


Διάγραμμα 3. Ο αριθμός γονιδίων με τα οποία ξεκινούσε κάθε μία από τις 7 φάσεις κατά τις οποίες ελέγχαμε για κατώτατο ποσοστό επιτυχίας 84.



Διάγραμμα 4. Ο αριθμός γονιδίων με τα οποία ξεκινούσε κάθε μία από τις 29 φάσεις κατά τις οποίες ελέγχαμε για κατώτατο ποσοστό επιτυχίας 89.

Στο επόμενο διάγραμμα παρουσιάζεται το μέσο ποσοστό επιτυχίας όπως αυτό διαμορφωνόταν κατά τη διάρκεια της όλης διαδικασίας.

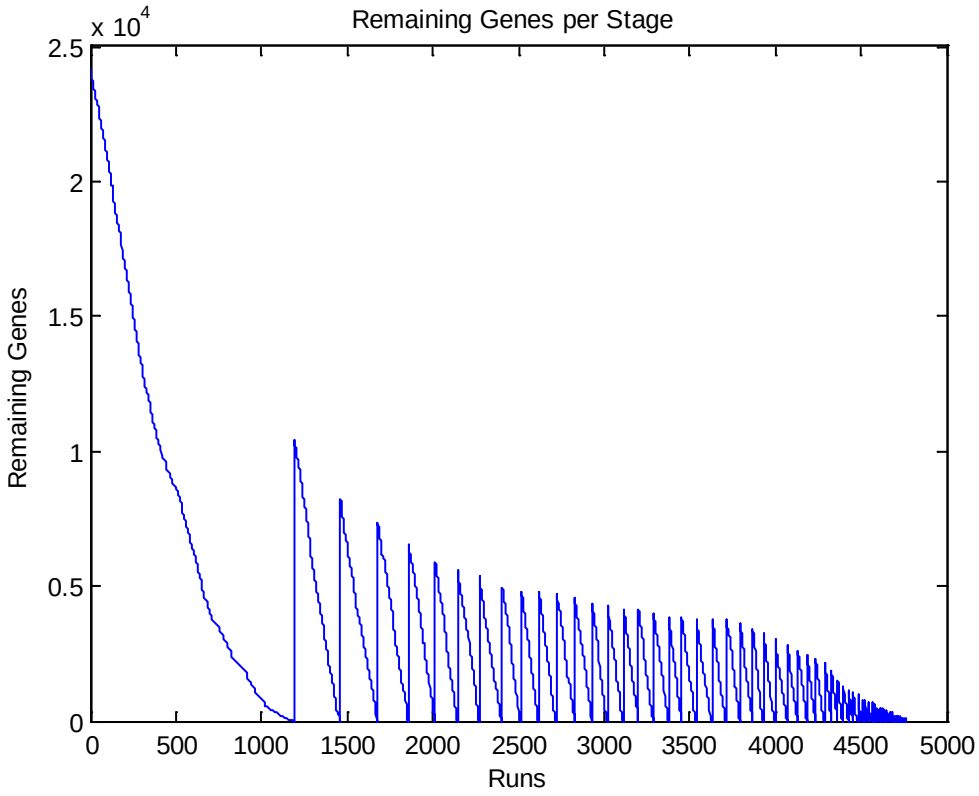


Διάγραμμα 5. Το μέσο ποσοστό επιτυχίας καθ` όλη τη διάρκεια της διαδικασίας

Το πρώτο ελάχιστο, στο run 1192, είναι το τέλος της πρώτης φάσης, όπου είχαν αφαιρεθεί για πρώτη φορά όλα τα γονίδια ως σημαντικά, ώστε να δούμε το ποσοστό επιτυχίας τους και να κρατήσουμε αυτά που έδιναν από 78 και πάνω στην επόμενη φάση. Στο run 3844 είναι η αρχή της φάσης 25 όπου και έχουμε πλέον αποκλείσει όλα τα γονίδια που δίνουν ποσοστό επιτυχίας κάτω από 78 και αυξάνουμε το όριο σε 84. Στο run 4296 είναι η αρχή της φάσης 33 και το σημείο όπου έχουμε αφαιρέσει όλα τα γονίδια που δίνουν ποσοστό επιτυχίας κάτω από 84 και αυξάνουμε το όριο σε 89.

Στο Διάγραμμα 6 παρουσιάζεται ο αριθμός των εναπομείναντων γονιδίων κατά τη διάρκεια της όλης διαδικασίας. Οι κορυφές που παρατηρούμε είναι οι αρχές των φάσεων όπου έχοντας αφαιρέσει κατά την προηγούμενη φάση όλα τα γονίδια ως σημαντικά ξεκινάμε την επόμενη δημιουργώντας νέα train και

test sets από τα γονίδια που έδωσαν ποσοστό επιτυχίας πάνω από το όριο που είχαμε ορίσει κάθε φορά.



Διάγραμμα 6. Ο αριθμός των εναπομείναντων γονιδίων καθ` όλη τη διάρκεια της διαδικασίας.

Καταλήξαμε λοιπόν σε μια γονιδιακή υπογραφή η οποία αποτελείται από 97 γονίδια και τα οποία δίνουν ποσοστό επιτυχίας 94.74. Πετύχαμε δηλαδή το στόχο που εξ αρχής είχαμε θέσει, καταλήξαμε σε μια γονιδιακή υπογραφή με λίγα γονίδια η οποία μάλιστα μας επιτρέπει με το συγκεκριμένο αλγόριθμο να προβλέψουμε με εξαιρετικά μεγάλη επιτυχία αν η το δείγμα (ασθενής) πρόκειται μέσα σε 5 χρόνια να παρουσιάσει μετάσταση. Όπως αναφέρθηκε στην παράγραφο 4.1 στη συγκεκριμένη παθολογία υπάρχει μεγάλη επικάλυψη ανάμεσα στις δύο κλάσεις και μάλιστα η διασπορά τους είναι μικρή με αποτέλεσμα να είναι δύσκολο να διαχωριστούν μεταξύ τους. Αυτό κάνει το αποτέλεσμα της συγκεκριμένης διπλωματικής ακόμα πιο σημαντικό. Παρακάτω στον Πίνακα 3 παρατίθενται τα 97 γονίδια στα οποία καταλήξαμε καθώς και η περιγραφή τους[40].

Πίνακας 3.

Η γονιδιακή υπογραφή 97 γονιδίων που επιλέχθηκε από τον RFE-LNW

Systematic name	Gene name	Gene Description
Contig31316_RC		ESTs
Contig7192		ESTs
NM_001669	**ARSD**	arylsulfatase D
Contig58360_RC		ESTs
AI432517_RC		th38b03.x1 NCI_CGAP_Pan1 Homo sapiens cDNA clone IMAGE:2120525 3' similar to SW:NU1M_HUMAN P03886 NADH-UBIQUINONE OXIDOREDUCTASE CHAIN 1 ; mRNA sequence.
L39061	**TAF1B**	TATA box binding protein (TBP)-associated factor, RNA polymerase I, B, 63kD
NM_002427	**MMP13**	matrix metalloproteinase 13 (collagenase 3)
NM_001778	**CD48**	CD48 antigen (B-cell membrane protein)
Contig21039_RC		ESTs
AL133101		Homo sapiens mRNA; cDNA DKFZp434O0921 (from clone DKFZp434O0921)
Contig49282_RC		Homo sapiens cDNA: FLJ21772 fis, clone COLF7808
NM_001906	**CTRB1**	chymotrypsinogen B1
AB007975	**KIAA0506**	KIAA0506 protein
NM_012114	**CASP14**	caspase 14, apoptosis-related cysteine protease
NM_003408	**ZFP37**	zinc finger protein homologous to Zfp37 in mouse
Contig49591_RC		ESTs
Contig27967_RC		ESTs
NM_012199	**EIF2C1**	eukaryotic translation initiation factor 2C, 1
Contig56857_RC	**IDE**	insulin-degrading enzyme
Contig53666_RC	**MGC3265**	ESTs, Weakly similar to
Contig9835_RC		ESTs
NM_004310	**ARHH**	ras homolog gene family, member H
Contig9446_RC		ESTs
Contig21835_RC		ESTs
AF052101		Homo sapiens clone 23872 mRNA sequence
Contig35644_RC		ESTs
NM_004419	**DUSP5**	dual specificity phosphatase 5
Contig34657_RC		ESTs
Contig31385_RC		Homo sapiens cDNA FLJ12782 fis, clone NT2RP2001869, moderately similar to ZINC FINGER PROTEIN 191
AL133641	**DKFZp586E1521**	Homo sapiens mRNA; cDNA DKFZp586E1521 (from clone DKFZp586E1521); partial cds
AI695056_RC		ESTs
NM_013326	**MIC1**	colon cancer-associated protein Mic1
NM_006059	**LAMC3**	laminin, gamma 3
Contig19877_RC		ESTs
NM_013370	**OKL38**	pregnancy-induced growth inhibitor
NM_013377	**DKFZp434B0417**	hypothetical protein

NM_020681	**HT018**	Homo sapiens HT018 protein (HT018), mRNA.
NM_006140	**CSF2RA**	colony stimulating factor 2 receptor, alpha, low-affinity (granulocyte-macrophage)
Contig30665_RC		ESTs
NM_005480	**TROAP**	trophinin associated protein (tastin)
Contig20214_RC		ESTs
Contig35079_RC		ESTs
Contig27288_RC		ESTs
Contig50719_RC		ESTs
Contig11737_RC		ESTs
NM_006459	**KEO4**	similar to Caenorhabditis elegans protein C42C1.9
Contig41110_RC		ESTs
Contig34163_RC		ESTs
NM_007245	**A2LP**	ataxin 2 related protein
D86974	**KIAA0220**	KIAA0220 protein
AF055033	**IGFBP5**	insulin-like growth factor binding protein 5
Contig6832_RC		ESTs
NM_006641	**CCR9**	chemokine (C-C motif) receptor 9
NM_013951	**PAX8**	paired box gene 8
NM_013987	**PARK2**	Parkinson disease (autosomal recessive, juvenile) 2, parkin
NM_014753	**KIAA0187**	KIAA0187 gene product
Contig58527_RC	**KIAA0670**	KIAA0670 protein/acinus
AL137416	**DKFZp434O192**	Homo sapiens mRNA; cDNA DKFZp434O192 (from clone DKFZp434O192); partial cds
NM_016351	**ADAM22**	a disintegrin and metalloproteinase domain 22
NM_006933	**SLC5A3**	solute carrier family 5 (inositol transporters), member 3
AF006061	**GH2**	growth hormone 2
AB037728	**KIAA1307**	KIAA1307 protein
Contig38628_RC		ESTs, Weakly similar to
NM_015929	**LOC51601**	lipoyltransferase
NM_015934	**NOP5/NOP58**	nucleolar protein NOP5/NOP58
Contig62901_RC		ESTs
NM_015966	**LOC51614**	hypothetical 43.2 Kd protein
NM_015987	**HEBP**	heme-binding protein
U18919	**FKBP3**	FK506-binding protein 3 (25kD)
Contig32627_RC		ESTs
NM_000055	**BCHE**	butyrylcholinesterase
NM_019002	**ETAA16**	ETAA16 protein
X78817	**ARHGAP4**	Rho GTPase activating protein 4
Contig21849_RC		ESTs
NM_018383	**FLJ11294**	hypothetical protein FLJ11294
Contig23646_RC		ESTs
NM_018428	**HCA66**	hepatocellular carcinoma-associated antigen 66
NM_000381	**MID1**	midline 1 (Opitz/BBB syndrome)
AB023163	**KIAA0946**	KIAA0946 protein; Huntingtin interacting protein H
Contig23620_RC		ESTs

NM_000431	**MVK**	mevalonate kinase (mevalonic aciduria)
NM_017880	**FLJ20558**	hypothetical protein FLJ20558
Contig54414_RC		ESTs
Contig38493_RC		ESTs
Contig35256		ESTs
NM_017996	**FLJ10103**	hypothetical protein FLJ10103
Contig11415_RC		ESTs
NM_000624	**SERPINA5**	Homo sapiens serine (or cysteine) proteinase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 5 (SERPINA5), mRNA.
NM_000688	**ALAS1**	aminolevulinate, delta-, synthase 1
Contig33909_RC		ESTs
NM_000746	**CHRNA7**	cholinergic receptor, nicotinic, alpha polypeptide 7
Contig46755_RC		ESTs
Contig28190_RC		ESTs
NM_002270	**KPNB2**	karyopherin (importin) beta 2
NM_000825	**GNRH1**	gonadotropin-releasing hormone 1 (leutinizing-releasing hormone)
NM_000873	**ICAM2**	intercellular adhesion molecule 2

Κεφάλαιο 6

Συμπεράσματα – μελλοντική εργασία

Σκοπός αυτής της διπλωματικής είναι χρησιμοποιώντας τον ταξινομητή RFE-LNW να καταλήξουμε σε μια γονιδιακή υπογραφή με τον ελάχιστο αριθμό γονιδίων τα οποία μας δίνουν χρήσιμη πληροφορία για το αν πρόκειται η ασθενής να παρουσιάσει μετάσταση μέσα στα επόμενα 5 χρόνια ή όχι. Να βρούμε δηλαδή τα γονίδια που μας δίνουν την πιο χρήσιμη πληροφορία όσων αφορά τον καρκίνο του μαστού, ώστε να μειώσουμε τις διαστάσεις της γονιδιακής έκφρασης που προέκυψαν από τα πειράματα με Microarrays του Van`t Veer και να καταλήξουμε σε ένα αποδοτικότερο μοντέλο ταξινόμησης, με μικρό αριθμό γονιδίων, ελπίζοντας έτσι να δώσουμε στους βιολόγους τη δυνατότητα να τα μελετήσουν και να αποκτήσουν σημαντική γνώση των μηχανισμών που είναι υπεύθυνοι για τη συγκεκριμένη ασθένεια, κάτι το οποίο μπορεί να οδηγήσει σε ανακάλυψη της κατάλληλης θεραπείας και σε έγκαιρη διάγνωση .

Ακολουθώντας τη διαδικασία που περιγράφηκε, συνεχώς μειώναμε τον αριθμό των γονιδίων, αφαιρώντας σε κάθε βήμα αυτά που κρίναμε ότι έδιναν τη λιγότερο χρήσιμη πληροφορία για το πεδίο έρευνάς μας, καταλήγοντας τελικά σε 97 γονίδια. Κατά τη διάρκεια της διαδικασίας τα γονίδια που απέμεναν όντως ήταν αυτά που έδιναν πιο χρήσιμη πληροφορία από αυτά που είχαν αφαιρεθεί. Αυτό το συμπέρασμα εξάγεται αν παρατηρήσουμε το μέσο ποσοστό επιτυχίας κατά τη διάρκεια της διαδικασίας το οποίο παρουσιάζεται στο Διάγραμμα 5. Παρατηρούμε ότι το μέσο ποσοστό επιτυχίας αρχικά μειώνεται. Αυτό συμβαίνει γιατί αρχικά περιέχονται πολλά γονίδια που περιέχουν άχρηστη πληροφορία. Στη συνέχεια παρατηρούμε ότι το μέσο ποσοστό επιτυχίας αυξάνεται, αφού τα γονίδια που απομένουν δίνουν πιο χρήσιμη πληροφορία. Προς το τέλος της διαδικασίας παρατηρούμε μία μείωση (παραμένει όμως σε ψηλά επίπεδα) η οποία οφείλεται στο μικρό αριθμό γονιδίων και όχι στην ποιότητα της πληροφορίας που αυτά μας δίνουν. Καταλήγουμε έτσι στα 97 προαναφερθέντα γονίδια με

το πολύ ψηλό ποσοστό επιτυχίας 94.74 και στο συμπέρασμα ότι η μέθοδος που ακολουθήσαμε όντως αφαιρεί τα γονίδια που δε μας δίνουν χρήσιμη πληροφορία.

Αυτή η διπλωματική έχει ως απώτερο σκοπό να βοηθήσει τους βιολόγους να κατανοήσουν τους μηχανισμούς που κρύβονται πίσω από τη νόσο του καρκίνου του μαστού. Στο μέλλον θα μπορούσαν να μελετηθούν αποτελέσματα και άλλων παρόμοιων μελετών, ούτως ώστε να γίνει προσπάθεια να συγκριθούν τα αποτελέσματα τους και πιθανώς να καταλήξουν σε κάποια κοινά γονίδια τα οποία ευθύνονται για τη νόσο αυτή και να εστιάσουν σε αυτά ώστε να είναι πιο εύκολη η εύρεση κατάλληλης θεραπείας και η έγκαιρη διάγνωση.

Επίσης στο μέλλον θα μπορούσε να υλοποιηθεί ο ταξινομητής RFE-LNW σε hardware, γιατί αυτό θα μείωνε κατά πολύ το χρόνο εξαγωγής αποτελεσμάτων, μιας και η software προσέγγιση είναι πολύ χρονοβόρα.

Τέλος θα ήταν καλό να δοκιμαστεί η διαδικασία αυτή και για άλλα δεδομένα καρκίνου, ή ίσως και για άλλες περιπτώσεις επιλογής χαρακτηριστικών που δεν έχουν να κάνουν με γονιδιακή επιλογή, ούτως ώστε να ερευνηθεί κατά πόσο αυτή η μέθοδος θα μπορούσε να αποδώσει και για άλλης φύσεως δεδομένα, πέραν του καρκίνου.

Βιβλιογραφία

[1] WHO, Cancer, *World Health Organization*, July 2008, http://www.who.int/mediacentre/factsheets/fs297/en/, accessed September 2008.

[2] American Cancer Society, Report sees 7.6 million global 2007 cancer deaths, *Reuters*, December 2008, http://www.reuters.com/article/healthNews/idUSN1633064920071217, accessed September 2008.

[3] Kinzler, Kenneth W.; Vogelstein, Bert (2002). "Introduction". *The genetic basis of human cancer* (2nd, illustrated, revised ed.). New York: McGraw-Hill, Medical Pub. Division. p. 5.

[4] Fodor, S. A., Massively parallel genomics, *Science, 277* (1997) 393–395.

[5] H. Hu, J. Li, H. Wang, G. Daggard, Combined Gene Selection Methods for Microarray Data Analysis, in *10th International Conference on Knowledge-Based & Intelligent Information & Engineering Systems*, Springer, vol. 4253 (20066) pp. 976-983

[6] L.J. Van't Veer, H. Dai, M.J. Van de Vijver, Y.D. He, et al., Gene expression profiling predicts clinical outcome of breast cancer, *Letters to Nature*, 415 (2002), pp 530–536.

[7] S. Y. Kim, T. Hamasaki, Evaluation of Clustering based on Preprocessing in Gene Expression Data, *International Journal of Biomedical Sciences,* vol3:1 (2008) pp 48-53.

[8] Yuchun Tang, Yan-Qing Zhang, Zhen Huang, "Development of Two-Stage SVM-RFE Gene Selection Strategy for Microarray Expression Data Analysis", *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, IEEE Computer Society Press, Vol 4:3 (2007) pp 365-381.

[9] Μαργαρίτης, Λ.Χ., *Κυτταρική Βιολογία*, Ιατρικές εκδόσεις Λίτσας, (Αθήνα 1989)

[10] Γ. Π. Φύσσας, Βασικά στοιχεία γενετικής, http://www.mastology.gr/gr/dna.asp, πρόσβαση Σεπτέμβριος 2008.

[11] Θωμόπουλος, Γ. Ν., *Βιολογία Κυττάρου*, University Studio Press, (Θεσσαλονίκη 1990)

[12] Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, Korbel JO, Emanuelsson O, Zhang ZD, Weissman S, Snyder M (2007). "What is a gene, post-ENCODE? History and updated definition". *Genome Research* **17** (6): 669–681.

[13] Anastasios Koutsos, Alexandra Manaia, Julia Willingale-Theune, The virtual microarray: Introduction to DNA Microarrays, *EMBL,* May 2007, http://www.embl.org/training/ells/teachingbase/project3/intro.pdf, accessed September 2008.

[14] Magic Z, Radulovic S, Brankovic-Magic M (2007). "cDNA microarrays: identification of gene signatures and their application in clinical practice". *J BUON* 12 Suppl 1: S39–44.

[15] Seer's training, Cancer as a disease: Categories of Cancer: Cancer classification, http://training.seer.cancer.gov/module_cancer_disease/unit3_categories2_by_histology.html, accessed September 2008.

[16] David Olle, DNA Microarrays: A New Tool for Cancer Diagnosis, December 2000, http://www.suite101.com/article.cfm/new_cancer_treatments/55304, accessed September 2008.

[17] Αναστασία Αναλυτή, Χαρίδημος Κονδυλάκης, Δημήτρης Μανακανάτας, Μάνος Καλαϊτζάκης, Δημήτρης Πλεξουσάκης, PrognoChip-BASE: το Γονιδιακό Πληροφοριακό Σύστημα του PrognoChip, *Ινστιτούτο Πληροφορικής, Ίδρυμα Τεχνολογίας & Έρευνας (ΙΤΕ-ΙΠ), Κρήτη, Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης, Κρήτη*.

[18] Jeremy Buhler, Anatomy of a Comparative Gene Expression Study, *Washington university in St.Louis,* August 2002, http://www.cs.wustl.edu/~jbuhler/research/array/, accessed September 2008.

[19] Π. Κ. Γλεντής, Καρκίνος του μαστού, http://www.surgeon.gr/110/3342.aspx, accessed September 2008.

[20] Sergios Theodoridis, Konstantinos Koutroumbas, Pattern recognition (3rd Edition), *Academic Press* (2006).

[21] Andrew R. Webb, Statistical Pattern Recognition (2nd Edition), *John Wiley & Sons* (2002).

[22] Christopher M. Bishop, Pattern Recognition and Machine Learning, *Springer* (2006).

[23] I. Guyon, J. Weston, S. Barnhill and V. Vapnik, Gene selection for cancer classification using Support vector machines, *machine learning*, 46 (2002) 389-422.

[24] A. Gorban, B. Kegl, D. Wunsch, A. Zinovyev (Eds.), Principal Manifolds for Data Visualisation and Dimension Reduction, LNCSE 58, Springer, Berlin - Heidelberg - New York, 2007.

[25] Peng, H.C., Long, F., and Ding, C., Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 27, No. 8, pp.1226-1238, 2005.

[26] Y. Wang, F. Makedon, J. Ford, J. Pearlman, HykGene: a hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data, Bioinformatics 21 (2004) 1530–1537.

[27] X. Zhou, K.Z. Mao, LS bound based gene selection for DNA microarray data, Bioinformatics 21 (2005) 1559–1564.

[28] I. Inza, P. Larranaga, R. Blanco, A.J. Cerrolaza, Filter versus wrapper gene selection approaches in DNA microarray domains, Artif. Intell.Med. 31 (2004) 91–103.

[29] L. Yu and H. Liu, Feature selection for high-dimensional data: A fast correlation-based filter solution, *In Proceedings of the International Conference on Machine Learning* (2003) 856 – 863.

[30] Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, 2000.

[31] SVM Parameters, http://www.svms.org/parameters/, accessed September 2008.

[32] X. Zhou, K.Z. Mao, LS bound based gene selection for DNA microarraydata, Bioinformatics 21 (8) (2004) 1559–1564.

[33] E. Ke Tank, P.N. Suganthan, X. Yao, Gene selection algorithms for microarray data based on least square support vector machine, BMC Bioinformatics 7 (95) (2006).

[34] A. Statnikov, C.F. Aliferis, I. Tsamardinos, D. Hardin, S. Levy, A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis, Bioinformatics 21 (5) (2005) 631–643.

[35] A. Statnikov, I. Tsamardinos, Y. Dosbayev, C.F. Aliferis, GEMS: a system for automated cancer diagnosis and biomarker discovery from microarray gene expression data, Int. J. Med. Informatics 74 (2005) 491–503.

[36] M.E. Blazadonakis, M. Zervakis, Wrapper filtering criteria via linear neuron and kernel approaches, Comp. Biol. Med. (2008),doi: 10.1016/j.compbiomed.2008.05.005.

[37] M. Riedmiller, H. Braun, A direct adoptive method for faster backpropagation learning: the RPROP algorithm, in: Proceedings of the IEEE International Conference on Neural Networks (ICNN), 1993, pp.586–591.

[38] R. Simon, M.D. Radmacher, K. Dobbin, L.M. McShane, Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification, J. Nat. Cancer Inst. 95 (2003) 14–18.

[39] M. E. Blazadonakis and M. Zervakis, “The Linear Neuron as a Marker Selector and Clinical predictor in Cancer Gene Analysis”, *Computer Methods and Programs in Biomedicine,* ELSEVIER, Vol 91:1 (2008) pp 22-35.

[40]http://a248.e.akamai.net/7/248/430/20051017192336/www.rii.com/publications/data/ArrayData_BRCA1.zip