ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΝΙΚΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

Προεπεξεργασία δειγμάτων για στατιστική επιλογή γονιδιακής υπογραφής

ΑΚΑΔΗΜΑΪΚΟ ΕΤΟΣ: 2008 – 2009

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΤΗΣ ΦΟΙΤΗΤΡΙΑΣ
ΜΑΓΙΑΤΗ ΜΑΡΓΑΡΙΤΑΣ
ΑΜ: 2001030095

ΕΠΙΒΛΕΠΩΝ
ΖΕΡΒΑΚΗΣ ΜΙΧΑΛΗΣ
Καθηγητής Πολυτεχνείου Κρήτης

ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΝΙΚΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

Προεπεξεργασία δειγμάτων για στατιστική επιλογή γονιδιακής υπογραφής

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΜΑΓΙΑΤΗ ΜΑΡΓΑΡΙΤΑ
ΑΜ: 2001030095

Εξεταστική Επιτροπή

Καθηγητής Πολυτεχνείου Κρήτης	**κ. ΖΕΡΒΑΚΗΣ ΜΙΧΑΛΗΣ**
Επίκουρος Καθηγητής Πολυτεχνείου Κρήτης	**κ. ΚΑΡΥΣΤΙΝΟΣ ΓΕΩΡΓΙΟΣ**
Αναπληρωτής Καθηγητής Πολυτεχνείου Κρήτης	**κ. ΛΙΑΒΑΣ ΑΘΑΝΑΣΙΟΣ**

ΠΡΟΛΟΓΟΣ

Η παρούσα διπλωματική εργασία με τίτλο «Προεπεξεργασία δειγμάτων για στατιστική επιλογή γονιδιακής υπογραφής», αφορά στην προεπιλογή ενός συνόλου γονιδίων για τις παθολογίες της λευχαιμίας και του καρκίνου του μαστού με προεπεξεργασία των δειγμάτων εκπαίδευσης, προκειμένου να ακολουθήσει στατιστική επιλογή γονιδιακής υπογραφής. Η διπλωματική εργασία ανατέθηκε στην τελειόφοιτο Μαγιάτη Μαργαρίτα 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 τον Δόκτορα 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 Λιάβα Αθανάσιο 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.

Μαγιάτη Μαργαρίτα
Χανιά, Νοέμβρης 2008

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.

Στην παρούσα διπλωματική εργασία ασχολούμαστε με δύο είδη καρκίνου, την λευχαιμία και τον καρκίνο του μαστού. Στην πρώτη περίπτωση γίνεται ταξινόμηση των ασθενών σε δύο διαφορετικά είδη λευχαιμίας, ενώ στη δεύτερη οι ασθενείς ταξινομούνται με βάση την προοπτική ή όχι του να εμφανίσουν μετάσταση εντός πέντε ετών.

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 προεπεξεργασία των δειγμάτων που έχουμε στη διάθεσή μας για τη λευχαιμία και τον καρκίνο του μαστού, με σκοπό να ακολουθήσει στατιστική επιλογή γονιδιακής υπογραφής. Βασίζεται στην leave-one-out cross-validation μέθοδο και δημιουργεί την προσδοκία να πετύχουμε καλύτερα αποτελέσματα. Η προεπεξεργασία εφαρμόζεται για δύο μεθόδους επιλογής γονιδίων, μια filter και μια wrapper, ενώ η αξιολόγηση του αποτελέσματος γίνεται με τρεις ταξινομητές, τους Nearest Mean, k-Nearest Neighbor και SVM. Γίνεται σύγκριση με τα αποτελέσματα που έχουμε χωρίς την προεπεξεργασία.

Η μεθοδολογία αυτή έχει πολύ ικανοποιητικά αποτελέσματα για την wrapper μέθοδο, δεν οδηγεί όμως σε καλύτερα αποτελέσματα για την filter. Το συμπέρασμα αυτό ισχύει και για τις δύο περιπτώσεις καρκίνου. Εξετάζεται ο ρόλος του κάθε ταξινομητή καθώς και το κατά πόσο οι διαφορετικές παράμετροι επηρεάζουν το τελικό αποτέλεσμα. Τέλος γίνεται μια σύγκριση ανάμεσα στις δύο παθολογίες.

ΠΕΡΙΕΧΟΜΕΝΑ

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 δύο είδη καρκίνου, την λευχαιμία και 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 της λευχαιμίας θέλουμε την κατάταξη των ασθενών σε δύο διαφορετικά είδη της συγκεκριμένης ασθένειας (ALL και AML), ενώ στην περίπτωση του καρκίνου του μαστού μας ενδιαφέρει η κατάταξή τους με βάση την προοπτική ή όχι του να εμφανίσουν μετάσταση εντός πέντε ετών. Για το σκοπό αυτό γίνεται χρήση μεθόδων αναγνώρισης προτύπων, όπου τα πρότυπα είναι οι ασθενείς και τα χαρακτηριστικά οι γονιδιακές τους εκφράσεις. Επομένως, για κάθε παθολογία, 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 λευχαιμία αντιστοιχούν περίπου 7000 γονίδια των οποίων οι σχετικές εκφράσεις καθορίζουν την ασθένεια, ενώ για την περίπτωση του καρκίνου του μαστού σε κάθε ασθενή αντιστοιχούν περίπου 25000 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.

Σκοπός αυτής της εργασίας δεν είναι η πρόταση μίας ακόμη μεθόδου επιλογής γονιδίων, αλλά η πρόταση για μια προεπεξεργασία των δεδομένων μας, ούτως ώστε να καταλήξουμε σε ένα μικρότερο υποσύνολο των αρχικών γονιδίων από όπου και να αρχίσει η γονιδιακή επιλογή. 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 περίπου 25000 γονίδια, ως αρχικό βήμα επιλέγουν από αυτά τα 5000 που εμφανίζουν την 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 προεπεξεργασίας 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.

Η δική μας πρόταση στηρίζεται πάνω σε μία ευρέως αποδεκτή μεθοδολογία για την εγγύηση της σταθερότητας ενός μοντέλου ταξινόμησης, γνωστή ως leave-one-out cross-validation, με στόχο την προεπεξεργασία των δειγμάτων εκπαίδευσης και την προεπιλογή ενός συνόλου γονιδίων, που με μετέπειτα επιλογή γονιδίων να οδηγήσει σε αποδοτικότερα μοντέλα ταξινόμησης.

Για κάθε περίπτωση καρκίνου, ερευνάται κατά πόσο η διαδικασία προεπεξεργασίας έχει αποτέλεσμα τόσο σε filter όσο και σε wrapper μεθόδους επιλογής γονιδίων, σχετικά με το αποτέλεσμα που προκύπτει χωρίς την προεπεξεργασία. Για την filter επιλογή χρησιμοποιείται το Fisher's criterion ενώ για την wrapper επιλογή τα βάρη του ταξινομητή Fisher's Linear Discriminant. Γίνεται μείωση των γονιδίων κατά στάδια και σε κάθε βήμα χρειάζεται να αξιολογηθεί κατά πόσο τα γονίδια που προκύπτουν είναι ικανά για την δημιουργία ενός αποδοτικού μοντέλου ταξινόμησης.

Για το σκοπό αυτό υπολογίζονται νέα train και test sets από τα παλιά σύνολα, τα οποία περιέχουν μόνο τα εναπομείναντα γονίδια. Χρησιμοποιούνται τρεις διαφορετικοί ταξινομητές, οι Nearest Mean, k-Nearest Neighbor και SVM – για διάφορες τιμές των παραμέτρων των δύο τελευταίων – οι οποίοι εκπαιδεύονται με το train set και των οποίων η απόδοση ελέγχεται με το test set. Επιπλέον, κάθε ταξινομητής παίζει το δικό του ρόλο στο στάδιο της προεπεξεργασίας, οπότε εξετάζεται τελικά ποιος φέρει το καλύτερο αποτέλεσμα για κάθε σύνολο δεδομένων.

Παράλληλα με τα συμπεράσματα που προκύπτουν για κάθε παθολογία, γίνεται μια σύγκριση ανάμεσα στις δύο περιπτώσεις καρκίνου, μελετώντας με αυτόν τον τρόπο την αποτελεσματικότητα της μεθόδου προεπεξεργασίας καθώς και την συμπεριφορά των ταξινομητών για διαφορετικής φύσεως δεδομένα.

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 περιγράφονται οι ταξινομητές οι οποίοι χρησιμοποιούνται. 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.

2 Μικροσυστοιχίες DNA – Συσχέτιση με την πάθηση του καρκίνου

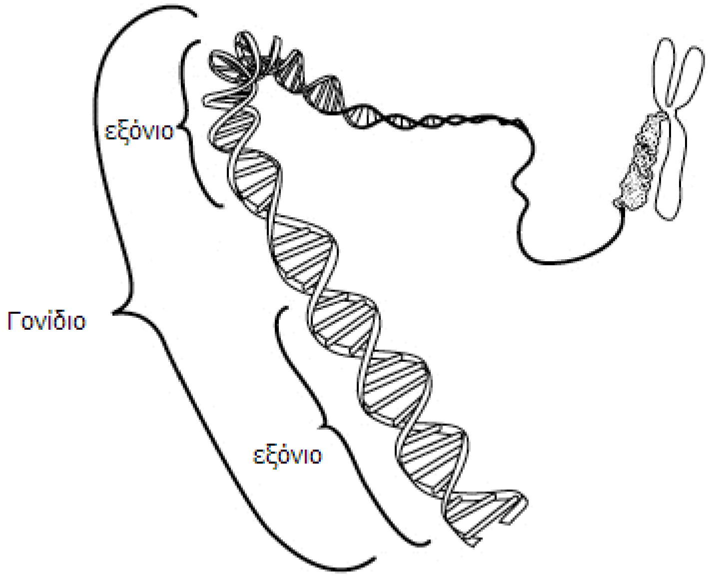
Το αντικείμενο μελέτης αυτής της διπλωματικής εργασίας βασίζεται στην εξαγωγή συμπερασμάτων για τον καρκίνο, στηριζόμενη σε πληροφορία που μπορεί να αντληθεί από την έκφραση των γονιδίων. Ακολουθεί λοιπόν μια αναφορά στο τι είναι γονιδιακή έκφραση, πως αυτή σχετίζεται με την ασθένεια του καρκίνου ενώ 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.

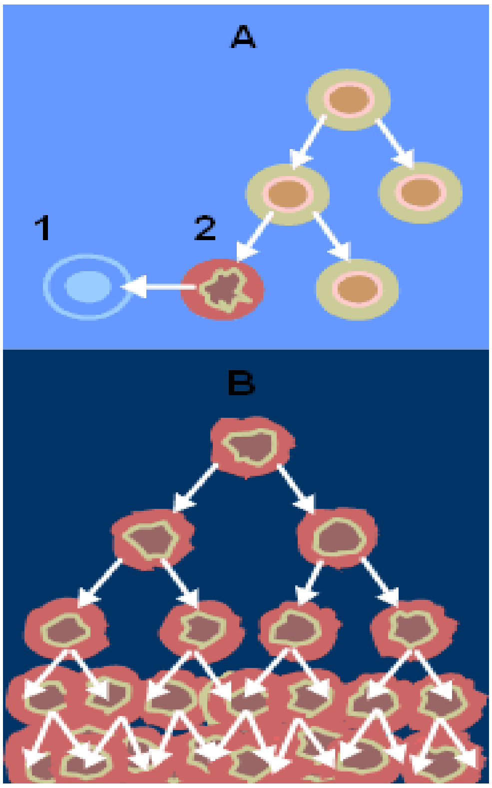
Κατά συνέπεια, το είδος και η ποσότητα των mRNAs που παράγονται από ένα κύτταρο αποτελεί αντικείμενο επιστημονικής μελέτης, ούτως ώστε να γίνει γνωστό ποια γονίδια εκφράζονται και κατ’ επέκταση πώς το κύτταρο αντιδρά στα διάφορα περιβαλλοντικά ερεθίσματα και στις δικές του μεταβαλλόμενες ανάγκες.

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 (Α) και καρκινική διαίρεση κυττάρων (Β), 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.

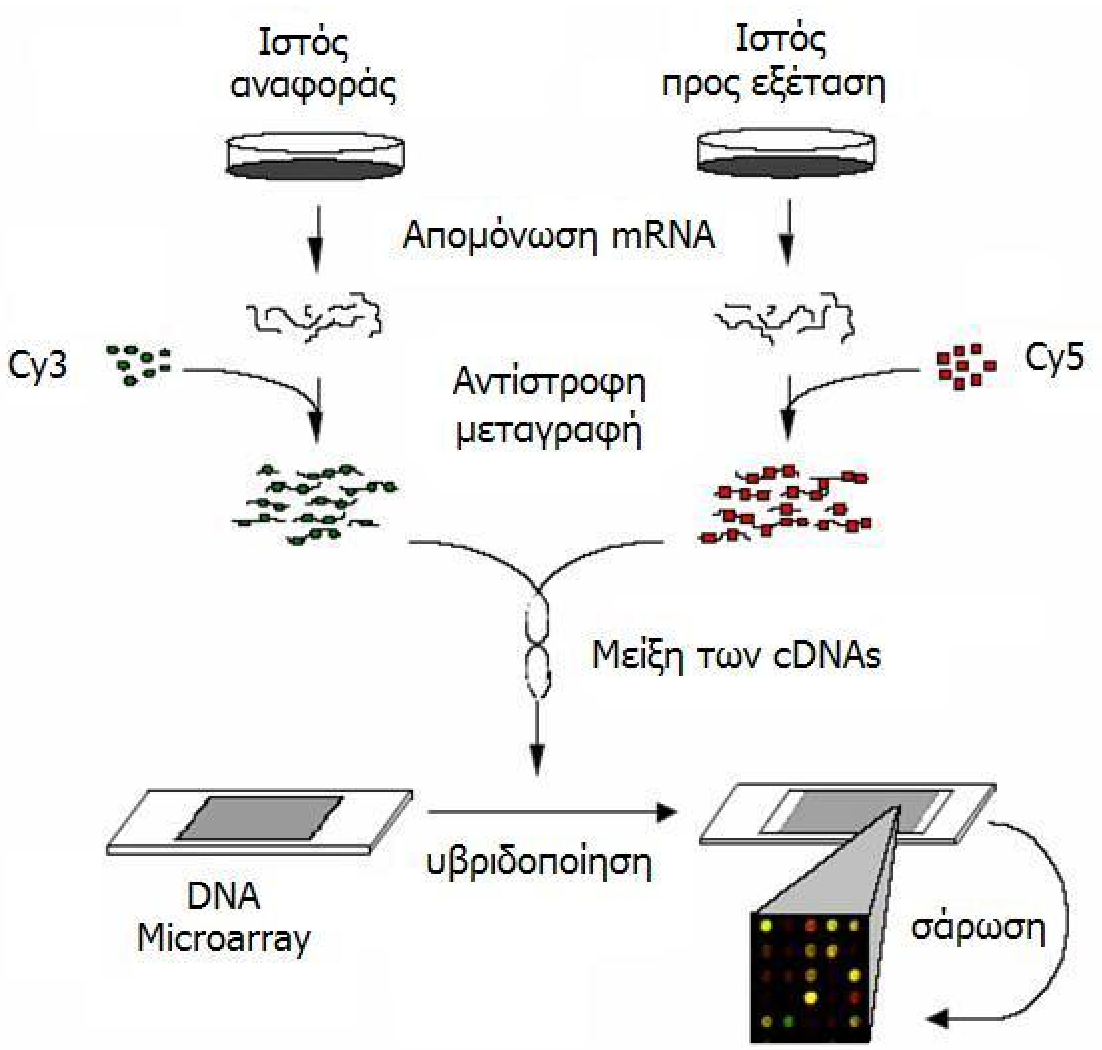
2.2 Τεχνολογία μικροσυστοιχιών DNA

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 δύο είδη καρκίνου, την λευχαιμία και τον καρκίνο του μαστού.

Η λευχαιμία είναι ένας καρκίνος του αίματος ή του μυελού των οστών και χαρακτηρίζεται από αφύσικο πολλαπλασιασμό των κυττάρων του αίματος, συνήθως των λευκών αιμοσφαιρίων. Δηλώνει την εμφάνιση στο αίμα άχρωμων στοιχείων, που εξαιτίας τους το αίμα παίρνει λευκάζουσα όψη, έτσι προκύπτει και ο όρος λευχαιμία [19]. Η κανονική αναλογία των λευκών αιμοσφαιρίων προς τα ερυθρά είναι 1:500. Με τη λευχαιμία τα λευκά αιμοσφαίρια αυξάνονται με αποτέλεσμα η αναλογία λευκών προς ερυθρών να είναι πλέον 1:20 ή και λιγότερο ακόμη. Η λευχαιμία είναι πάθηση κυρίως του μυελού των οστών, οπότε και ονομάζεται μυελογενής λευχαιμία, ή του λεμφικού συστήματος οπότε και λέγεται λεμφογενής λευχαιμία.

Εξετάζονται ιστοί ασθενών που πάσχουν από οξεία λεμφογενή λευχαιμία (Acute lymphocytic leukemia, ALL) και ιστοί ασθενών που πάσχουν από οξεία μυελογενή λευχαιμία (Acute myelogenous leukemia, AML). Όταν εφαρμόστηκε η τεχνολογία μικροσυστοιχιών DNA από ερευνητές [20], κατέστη δυνατό να διαχωριστούν οι δύο αυτοί διαφορετικοί τύποι λευχαιμίας, χωρίς προηγούμενη γνώση για κάθε περίπτωση. Αυτό είναι πολύ σημαντικό διότι οι δύο λευχαιμίες απαιτούν εντελώς διαφορετικό είδος χημειοθεραπείας.

Το δεύτερο πείραμα αφορά 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 [21]. Τα κύτταρα αυτά έχουν την δυνατότητα να διασπάσουν κάποιους φυσιολογικούς ανατομικούς φραγμούς του μαστού και να διασπαρθούν στο υπόλοιπο σώμα.

Εξετάζονται ιστοί ασθενών που έπασχαν από καρκίνο του μαστού και δεν επανεμφάνισαν την ασθένεια σε διάστημα τουλάχιστον πέντε ετών, και ιστοί ασθενών που υποτροπίασαν σε διάστημα πέντε ετών. Ερευνητές [6] χρησιμοποιούν την τεχνολογία μικροσυστοιχιών DNA για να προβλέψουν ποιοι ασθενείς έχουν όγκους οι οποίοι έχουν προοπτική να εξελιχθούν σε μεταστατικοί.

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) [22]. 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) [23]. 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 απριόρι 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) [24].

Ένα σύνολο εκπαίδευσης λοιπόν αποτελείται από ένα πλήθος *M* προτύπων x_i, $i=1,2,\ldots,M$, 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 [23]. 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

$$\begin{aligned} D(x) > 0 \Rightarrow x \in \omega_1 \\ D(x) < 0 \Rightarrow x \in \omega_2 \end{aligned} \qquad (1)$$

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

$$g_i(x) > g_j(x) \Rightarrow x \in \omega_i \ , \ i, j = 1,2,\ldots,C \ , i \neq j \qquad (2)$$

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

$$D(x) = g_1(x) - g_2(x) \qquad (3)$$

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) [23]. Μια τέτοια συνάρτηση έχει τη μορφή

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

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_0 την 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.

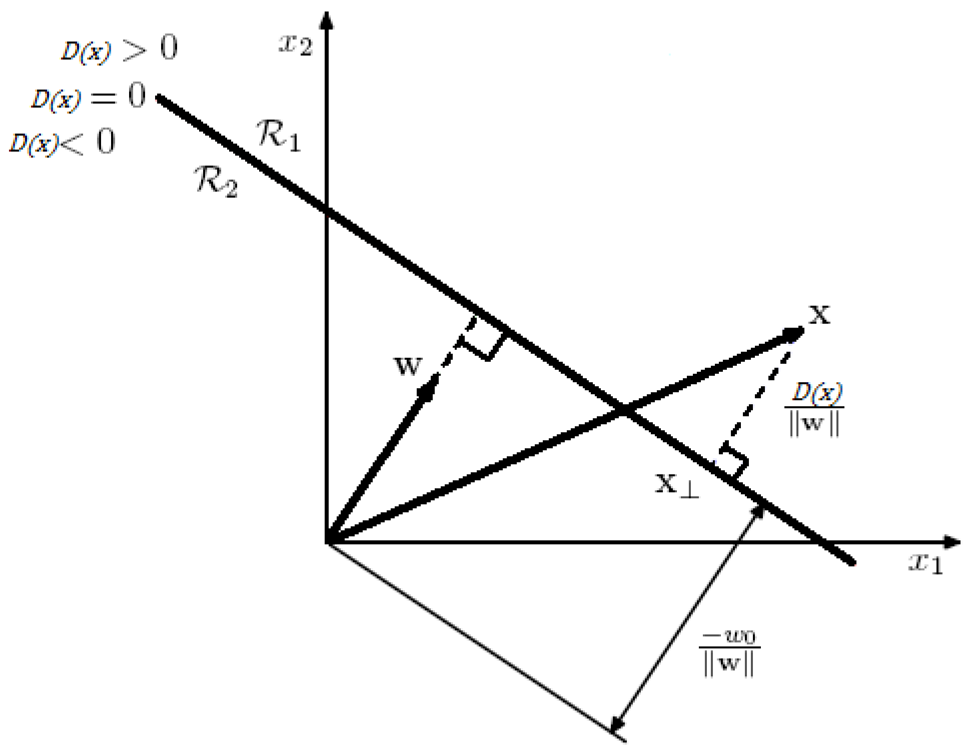
Εδώ χρειάζεται να επισημάνουμε ότι υπάρχουν περισσότερα του ενός υπερεπίπεδα τα οποία είναι ικανά να διαχωρίσουν τις κλάσεις μεταξύ τους. Μέσω της διαδικασίας της εκπαίδευσης εντοπίζεται και το βέλτιστο υπερεπίπεδο, σύμφωνα πάντα και με τον τρόπο λειτουργίας του κάθε ταξινομητή.

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

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

$$\Rightarrow w^t (x_1 - x_2) = 0 \tag{5}$$



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

Είναι προφανές ότι το διάνυσμα $x1-x2$ 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 υποχώρους 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 [25].

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 [23]. Αυτό σημαίνει ότι μπορεί να ταξινομεί με επιτυχία τα πρότυπα εκπαίδευσης αλλά να αδυνατεί να ταξινομήσει νέα πρότυπα. Αυτό είναι το φαινόμενο του over-fitting. 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" [22]. Σε περιπτώσεις όπου το πλήθος των χαρακτηριστικών είναι πολύ μεγάλο σχετικά με το πλήθος των προτύπων είναι πιθανό να εμφανιστεί το φαινόμενο του over-fitting. Για ένα δεδομένο μοντέλο ταξινόμησης, το πρόβλημα του over-fitting 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) [25].

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 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) [26].

3.2.1 Εξαγωγή χαρακτηριστικών

Κατά την εξαγωγή χαρακτηριστικών, νέα χαρακτηριστικά , λιγότερα από τα αρχικά, προκύπτουν με συνδυασμό – γραμμικό ή μη γραμμικό – των αρχικών [23]. 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 [27].

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 [25]. 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.

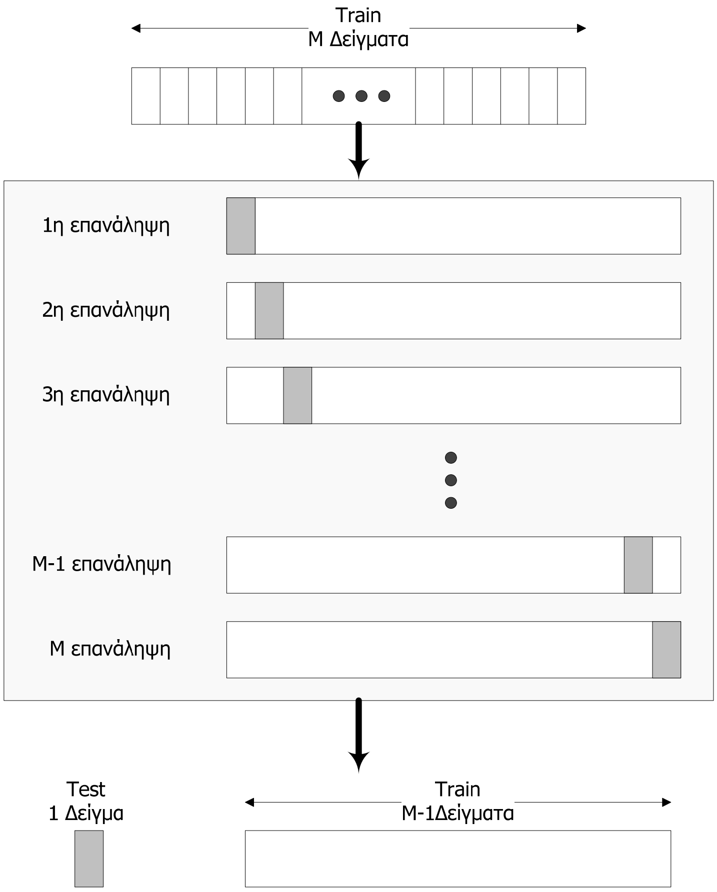
Σε αντίθεση με την εξαγωγή χαρακτηριστικών, χαρακτηριστικά τα οποία δεν δίνουν χρήσιμη πληροφορία μπορούν να απορριφθούν και έτσι καταλήγουμε σε πιο πληροφοριακά σύνολα χαρακτηριστικών με τα οποία επιτυγχάνεται καλύτερη απόδοση του ταξινομητή. Οι αλγόριθμοι επιλογής χαρακτηριστικών χωρίζονται σε δύο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

Οι filter μέθοδοι βασίζονται σε γενικά γνωρίσματα των δεδομένων εκπαίδευσης ώστε να γίνει επιλογή χαρακτηριστικών χωρίς να περιλαμβάνεται οποιαδήποτε διαδικασία μάθησης ταξινομητή [28]. 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 over-fitting, όπως επίσης τείνουν να είναι υπολογιστικά πιο δαπανηρές από τις filter μεθόδους.

3.3 Leave-one-out cross-validation

Γενικά η μέθοδος cross validation αφορά στον διαχωρισμό ενός συνόλου δειγμάτων εκπαίδευσης σε υποσύνολα, από τα οποία το ένα χρησιμοποιείται για εκπαίδευση ενός ταξινομητή και τα υπόλοιπα για εξακρίβωση της απόδοσής του [29]. Αυτή η μέθοδος χρησιμοποιείται για την εκτίμηση του μοντέλου ταξινόμησης.

Η πιο διαδεδομένη μορφή της μεθόδου προκειμένου να μελετηθεί η εκτίμηση του μοντέλου ταξινόμησης είναι η leave-one-out cross-validation (LOOCV). Σύμφωνα με αυτή, από το σύνολο εκπαίδευσης, το οποίο αποτελείται από M δείγματα, κάθε φορά επιλέγεται το ένα δείγμα ως test και τα υπόλοιπα $M-1$ ως train. Επομένως δημιουργούνται M διαφορετικά ζεύγη train-test. Η διαδικασία του LOOCV φαίνεται στο Σχήμα 5.



Σχήμα 5. Διαδικασία LOOCV

Με τον τρόπο αυτό για κάθε διαφορετικό ζεύγος train και test υπολογίζεται ένα ποσοστό σφάλματος. Ο μέσος όρος του σφάλματος ανά τα M διαφορετικά ζεύγη train και test αποτελεί και το συνολικό σφάλμα ταξινόμησης. Όσο πιο μικρό είναι το σφάλμα ταξινόμησης, τόσο πιο ικανός είναι ο ταξινομητής να γενικεύεται σε άγνωστα σε αυτόν δεδομένα.

3.4 Ταξινομητές που χρησιμοποιούνται κατά την υλοποίηση

Κατά την υλοποίηση της παρούσας διπλωματικής εργασίας γίνεται χρήση τεσσάρων διαφορετικών ταξινομητών οι οποίοι περιγράφονται στις υποενότητες που ακολουθούν. Το πρόβλημά μας έχει να κάνει με ταξινόμηση σε δύο κλάσεις οπότε δεδομένου αυτού και χάριν απλούστευσης της ανάλυσης περιοριζόμαστε σε περιγραφή των ταξινομητών για δύο κλάσεις.

3.4.1 Ταξινομητής κοντινότερης μέσης τιμής – Nearest Mean Classifier

Έστω $X=\{x_1, x_2, \ldots, x_M\}$ το σύνολο των M δειγμάτων 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 $y_i = \{-1, +1\}$. Επιπλέον, έστω ότι με μ_1 συμβολίζουμε το διάνυσμα μέσων τιμών των M_1 αντικειμένων που απαρτίζουν το σύνολο $X_1 \subset X$ με ετικέτα κλάσης -1 και με μ_2 το διάνυσμα μέσων τιμών των M_2 αντικειμένων που απαρτίζουν το σύνολο $X_2 \subset X$ με ετικέτα κλάσης +1. Τότε θα έχουμε

$$\mu_1 = \frac{1}{M_1} \sum_{\{i:y_i=-1\}} x_i \tag{6}$$

και $$\mu_2 = \frac{1}{M_2} \sum_{\{i:y_i=+1\}} x_i \tag{7}$$

Μας ενδιαφέρει η μέτρηση της απόστασης ενός προτύπου x προς ταξινόμηση από τα διανύσματα μ_1 και μ_2. Γενικά χρησιμοποιούνται διάφοροι τρόποι μέτρησης απόστασης, εδώ θα γίνει χρήση της ευκλείδειας απόστασης.

Η ευκλείδεια απόσταση δύο διανυσμάτων $a=\{a_1, a_2, \ldots, a_n\}$ και $b=\{ b_1, b_2, \ldots, b_n\}$ δίνεται από την σχέση:

$$d_E(a,b) = \sqrt{(a_1-b_1)^2 + (a_2-b_2)^2 + \cdots + (a_n-b_n)^2} \tag{8}$$

Έστω τώρα ότι $D(x)$ είναι η συνάρτηση απόφασης, τότε θα είναι της μορφής [30]:

$$D(x) = {d_E}^2(x,\mu_1) - {d_E}^2(x,\mu_2) \tag{9}$$

Για $D(x)=0$ προκύπτει το υπερεπίπεδο που διαχωρίζει τις δύο κλάσεις στον N-διάστατο χώρο χαρακτηριστικών. Εάν τώρα $f(x)$ είναι μία συνάρτηση που επιστρέφει την ετικέτα κλάσης για ένα πρότυπο x, τότε

$$f(x) = \begin{Bmatrix} -1, D(x) > 0 \\ +1, D(x) < 0 \end{Bmatrix} \tag{10}$$

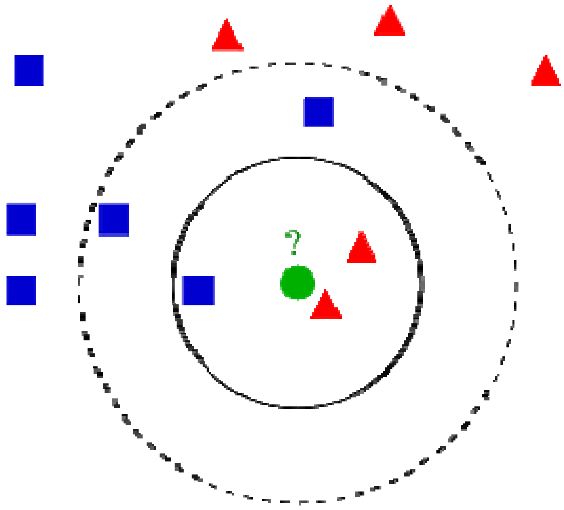
Επομένως, για κάθε πρότυπο x προς ταξινόμηση, υπολογίζεται η τετραγωνισμένη ευκλείδεια απόστασή του από τα διανύσματα μέσων τιμών μ_1 και μ_2 τα οποία αντιπροσωπεύουν την πρώτη και την δεύτερη κλάση αντίστοιχα. Υπολογίζουμε λοιπόν σε ποιο διάνυσμα μέσης τιμής είναι κοντινότερο το x, ούτως ώστε να το κατατάξουμε στην αντίστοιχη κλάση. Εάν η απόσταση από το μ_1 είναι μικρότερη από την απόσταση από το μ_2 τότε το πρότυπο ταξινομείται στην πρώτη κλάση (ετικέτα κλάσης -1) διαφορετικά ταξινομείται στην δεύτερη κλάση (ετικέτα κλάσης +1).

3.4.2 Ταξινομητής *k* κοντινότερων γειτόνων – *k*-Nearest Neighbor Classifier

Έστω $X=\{x_1, x_2, \ldots, x_M\}$ 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. Επίσης έστω ότι C είναι το σύνολο των κλάσεων ω_i , $i=1,2,...C$. Το x θα είναι ένα πρότυπο προς ταξινόμηση και έστω ότι έχουμε ένα συγκεκριμένο μέτρο μέτρησης απόστασης (π.χ. ευκλείδεια απόσταση), τότε ο αλγόριθμος του k-Nearest Neighbor ταξινομητή έχει να κάνει ως εξής [22]:

- Από το σύνολο M των προτύπων εκπαίδευσης, ορίζονται οι k κοντινότεροι γείτονες του x (δηλαδή τα k πρότυπα από τα οποία απέχει λιγότερο το x), ανεξαρτήτως της ετικέτας κλάσης που αυτά έχουν.
- Από τα k αυτά δείγματα, προσδιορίζεται το πλήθος k_i των προτύπων που ανήκουν στην κλάση ω_i, $i=1,2,\ldots,C$. Προφανώς $\sum_{i=1}^{C} k_i = k$
- Το x ταξινομείται στην κλάση ω_i με το μέγιστο πλήθος k_i δειγμάτων.

Επομένως η δουλειά του k-Nearest Neighbor ταξινομητή είναι να ταξινομεί ένα πρότυπο x στην κλάση εκείνη που έχει την πλειοψηφία ανάμεσα στους k γείτονές του (Σχήμα 6). Η επιλογή του k πρέπει να είναι τέτοια ώστε να αποφεύγεται η σύγχυση κατά την ταξινόμηση. Για παράδειγμα, σε περιπτώσεις ταξινόμησης σε δύο κλάσεις, όπως στην περίπτωσή μας, το k είναι καλό να είναι περιττός αριθμός γιατί υπάρχει περίπτωση οι $k/2$ γείτονες ενός προτύπου να ανήκουν στην μία κλάση και οι υπόλοιποι $k/2$ στην άλλη, οπότε να μην είναι ξεκάθαρο σε ποια κλάση το συγκεκριμένο πρότυπο ανήκει. Γενικά αν έχουμε *C* κλάσεις είναι προτιμότερο το *k* να μην είναι πολλαπλάσιο του *C*.

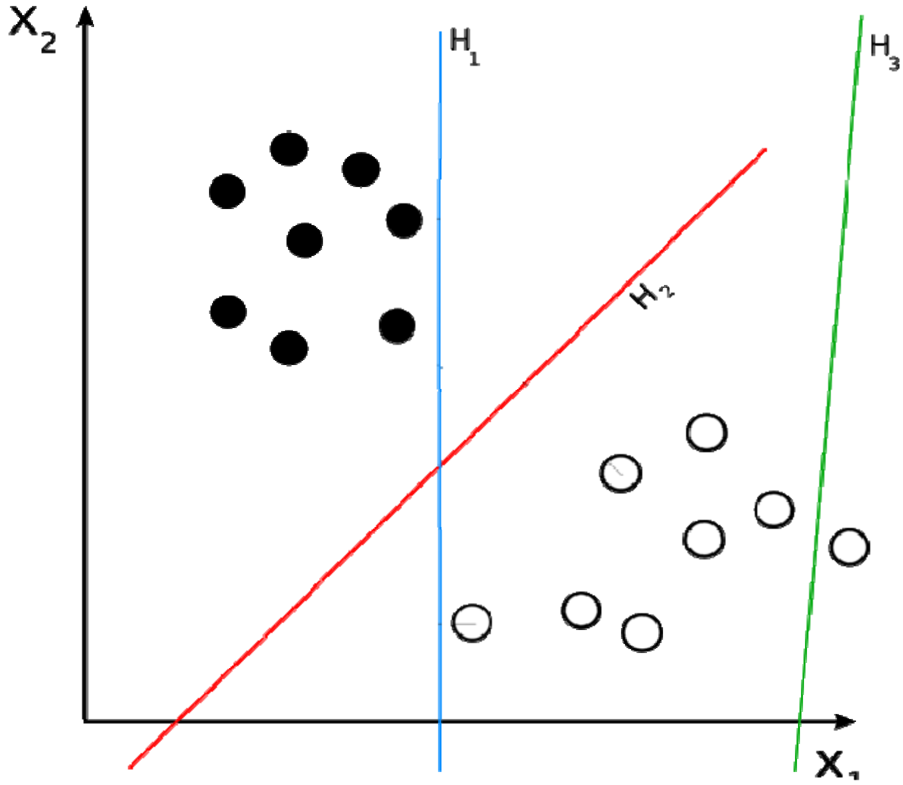


Σχήμα 6. Παράδειγμα ταξινόμησης με k-NN ταξινομητή σε δύο κλάσεις. Εάν k=3, το δείγμα προς ταξινόμηση (πράσινος κύκλος) ταξινομείται στην κλάση των κόκκινων τριγώνων, αν k=5 τότε ταξινομείται στην κλάση των μπλε τετραγώνων.

Ένα μειονέκτημα αυτού του αλγορίθμου είναι η υπολογιστική ισχύς που χρειάζεται για την εύρεση των *k* κοντινότερων γειτόνων ανάμεσα στα *M* διαφορετικά δείγματα εκπαίδευσης. Ειδικά όταν το *M* είναι πολύ μεγάλο το πρόβλημα γίνεται ακόμα πιο σοβαρό.

3.4.3 Support Vector Machines (SVMs)

Όπως έχουμε αναφέρει και παραπάνω, υπάρχουν διάφορα υπερεπίπεδα τα οποία 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 στον *N*-διάστατο χώρο των χαρακτηριστικών, τέτοια ώστε να μεγιστοποιείται το περιθώριο ανάμεσα στις κλάσεις. Η απόσταση ενός τέτοιου υπερεπιπέδου από το κοντινότερο αντικείμενο κάθε κλάσης είναι η μέγιστη δυνατή [31]. Υπάρχουν και μη γραμμικά SVMs αλλά δε θα ασχοληθούμε με αυτά σε αυτήν την διπλωματική εργασία.



Σχήμα 7. Δύο γραμμικώς διαχωριζόμενα σύνολα δεδομένων. Το υπερεπίπεδο H_3 δε διαχωρίζει τις δύο κλάσεις. Τα υπερεπίπεδα H_1 και H_2 διαχωρίζουν τις κλάσεις όμως το H_1 επιτυγχάνει μικρό περιθώριο ανάμεσα στις κλάσεις ενώ το H_2 μέγιστο.

Ας υποθέσουμε ότι έχουμε διαχωρισμό σε δύο κλάσεις, καθώς επίσης και ότι τα δεδομένα των δύο αυτών κλάσεων είναι γραμμικώς 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 την μεγαλύτερη απόσταση από τα κοντινότερα αντικείμενα των δύο κλάσεων, δηλαδή το περιθώριο είναι το μέγιστο, οπότε έχουμε και μικρότερο σφάλμα γενίκευσης. Στο Σχήμα 7 φαίνεται η διαφορά ανάμεσα σε ένα οποιοδήποτε υπερεπίπεδο που διαχωρίζει δύο κλάσεις και σε αυτό που επιτυγχάνει το μέγιστο περιθώριο ανάμεσα στις κλάσεις.

Εκείνα τα σημεία που βρίσκονται στα όρια του περιθωρίου, δηλαδή βρίσκονται πάνω στα υπερεπίπεδα εκατέρωθεν του διαχωριστικού υπερεπιπέδου ονομάζονται support vectors [22]. 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:

$$D(x) = w^T x + w_0 \tag{11}$$

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

$$y_i(w^T x_i + w_0) \geq 0 \quad \text{για} \quad i = 1, 2, \ldots, M \tag{12}$$

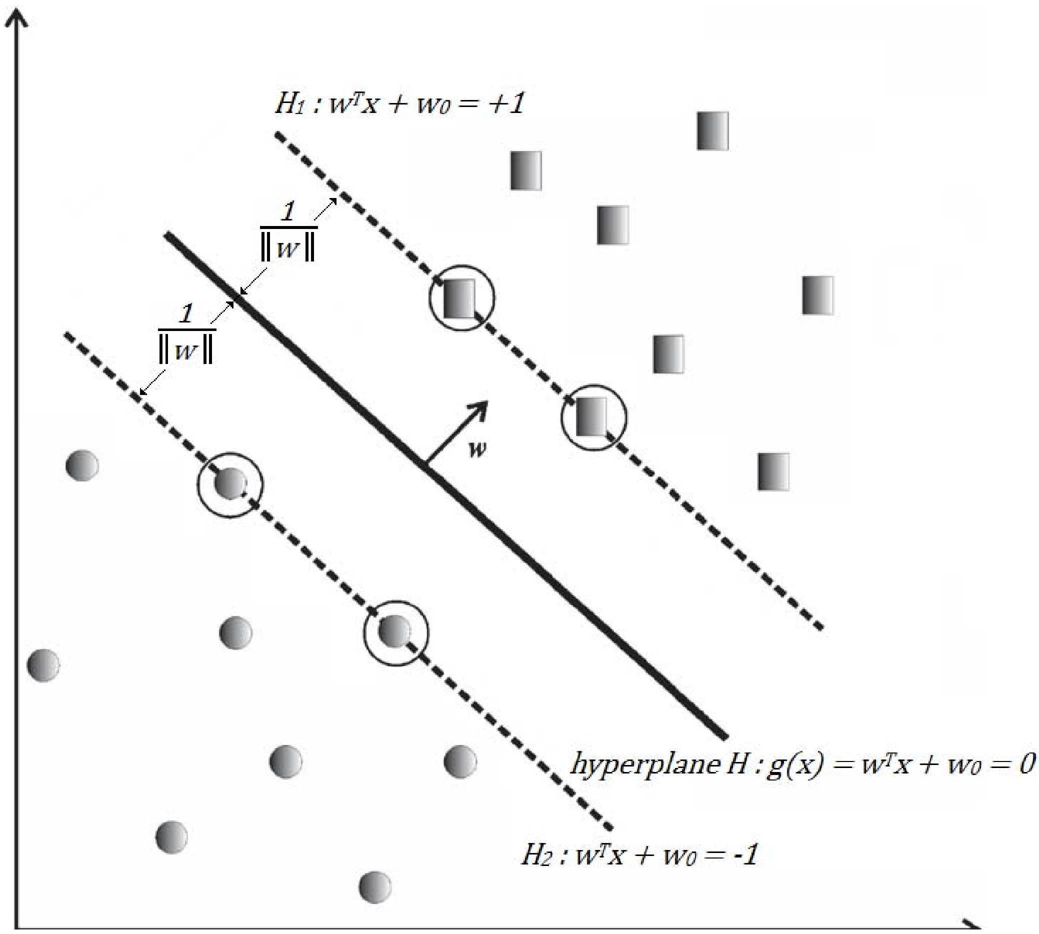
Έστω *H* το διαχωριστικό υπερεπίπεδο για το οποίο ισχύει ότι *D(x)* = 0. Ορίζουμε δύο ισαπέχοντα υπερεπίπεδα H_1 και H_2 εκατέρωθεν του *H* για τα οποία ισχύει:

$$\begin{aligned} H_1 &: w^T x + w_0 = +1 \\ H_2 &: w^T x + w_0 = -1 \end{aligned} \tag{13}$$

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

$$w^T x_i + w_0 \geq +1 \quad \text{για} \quad y_i = +1 \tag{14}$$

και $$w^T x_i + w_0 \leq -1 \quad \text{για} \quad y_i = -1 \tag{15}$$



Σχήμα 8. Υπερεπίπεδα και support vectors για δύο γραμμικά διαχωριζόμενα σύνολα δεδομένων

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 8 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. 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:

$$J(w) = \frac{1}{2} \| w \|^2 \tag{16}$$

η οποία πρέπει να ελαχιστοποιηθεί υπό τον όρο ότι:

$$y_i(w^T x_i + w_0) \geq 1 \quad \text{για} \quad i = 1, 2, ..., M \tag{17}$$

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 [22]. 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.

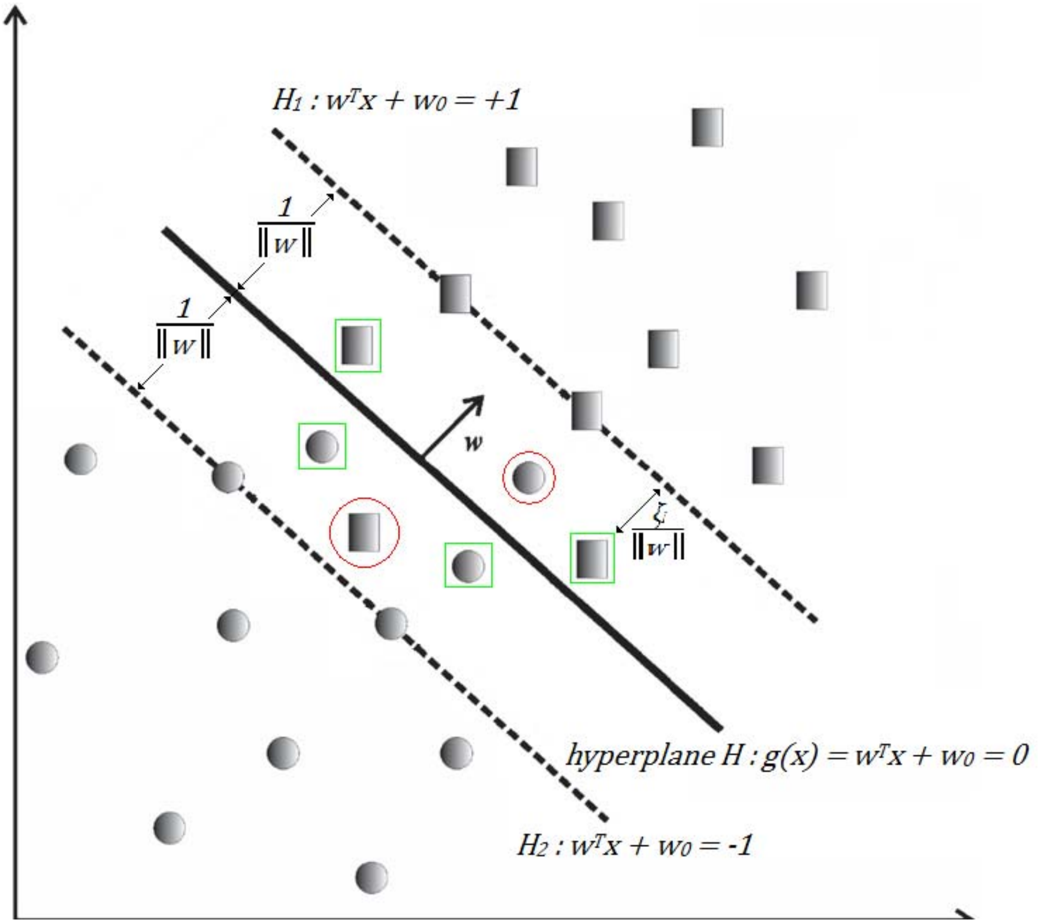
Σε αυτές τις περιπτώσεις τα δείγματα εκπαίδευσης μπορεί να ανήκουν σε μία από τις παρακάτω τρεις κατηγορίες:

- Να βρίσκονται εκτός περιθωρίου και να είναι σωστά ταξινομημένα.
- Να βρίσκονται εντός περιθωρίου και να είναι σωστά ταξινομημένα (στο Σχήμα 9 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 9 τα σημεία με το κυκλικό περίγραμμα). Αυτό σημαίνει ότι ικανοποιούν την ανισότητα

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



Σχήμα 9. 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

$$y_i(w^T x_i + w_0) \geq 1 - \xi_i \quad \text{για} \quad i = 1, 2, \ldots, M \tag{18}$$

σχέση η οποία 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 $\xi_i / ||w||$, όπως φαίνεται στο παραπάνω σχήμα (Σχήμα 9). 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

$$J(w, w_0, \xi) = \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{M} \xi_i \tag{19}$$

υπό τον περιορισμό της σχέσης 18 και για

$$\xi_i \geq 0 \quad \text{για} \quad i = 1, 2, \ldots, M \tag{20}$$

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 [32]. 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 [22]. 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:

$$\frac{\partial L}{\partial w} = 0 \Rightarrow w = \sum_{i=1}^{M} \lambda_i y_i x_i \tag{21}$$

$$\frac{\partial L}{\partial w_0} = 0 \Rightarrow \sum_{i=1}^{M} \lambda_i y_i = 0 \tag{22}$$

$$\frac{\partial L}{\partial \xi_i} = 0 \Rightarrow C - \mu_i - \lambda_i = 0\,,\; i = 1, 2, \ldots, M \tag{23}$$

$$\lambda_i[y_i(w^T x_i + w_0) - 1 + \xi_i] = 0,\ i = 1, 2, \ldots, M \tag{24}$$

$$\mu_i \xi_i = 0,\ i = 1, 2, \ldots, M \tag{25}$$

$$\mu_i \geq 0,\ \lambda_i \geq 0,\ i = 1, 2, \ldots, M \tag{26}$$

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

$$L(w, w_0, \xi, \lambda, \mu) = \frac{1}{2}\| w \|^2 + C\sum_{i=1}^{M}\xi_i - \sum_{i=1}^{M}\mu_i\xi_i - \sum_{i=1}^{M}\lambda_i[y_i(w^T x_i + w_0) - 1 + \xi_i] \tag{27}$$

Η δυαδική αναπαράσταση της συνάρτησης Lagrange είναι η εξής:

μεγιστοποίηση του $L(w, w_0, \lambda, \xi, \mu)$ (28)

υπό τον όρο ότι $w = \sum_{i=1}^{M}\lambda_i y_i x_i$ (29)

$$\sum_{i=1}^{M}\lambda_i y_i = 0 \tag{30}$$

$$C - \mu_i - \lambda_i = 0 \tag{31}$$

και $\lambda_i \geq 0, \mu_i \geq 0,\ i = 1, 2, \ldots, M$ (32)

Αντικαθιστώντας τους παραπάνω περιορισμούς στην συνάρτηση Lagrange προκύπτει

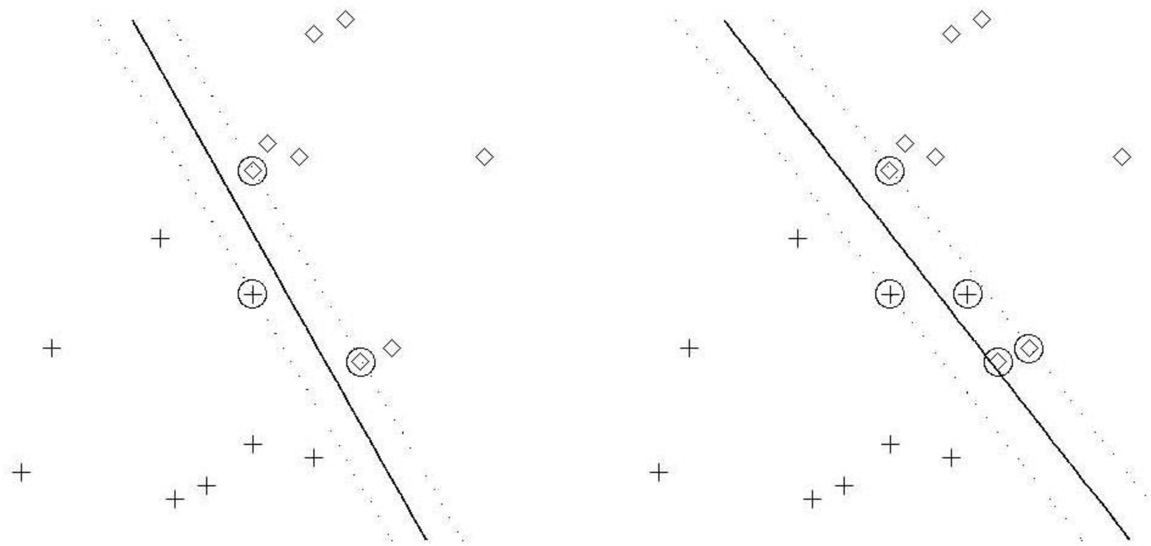
$$\max_{\lambda}\left(\sum_{i=1}^{M}\lambda_i - \frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{M}\lambda_i\lambda_j y_i y_j x_i^T x_j\right) \tag{33}$$

υπό τον όρο ότι

$$0 \leq \lambda_i \leq C,\ i = 1, 2, \ldots, M \tag{34}$$

$$\text{και} \quad \sum_{i=1}^{M} \lambda_i y_i = 0 \qquad (35)$$

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.



Σχήμα 10. Διαφορά μεταξύ γραμμικά διαχωριζόμενων κλάσεων (αριστερά) και μη διαχωριζόμενων γραμμικά κλάσεων (δεξιά)

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 10 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.

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

$$D(x) = \sum_{i=1}^{M} y_i \lambda_i \cdot (x^T x_i) + w_0 \quad (36)$$

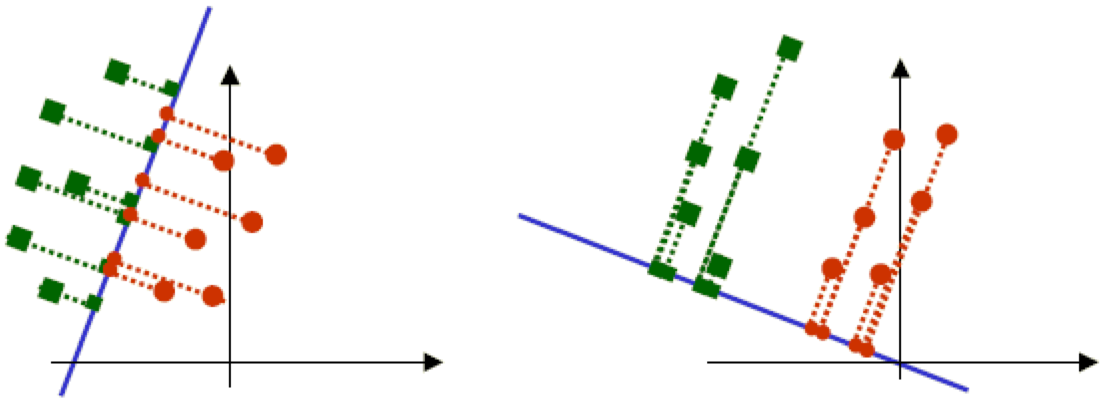
όπου για $D(x) = 0$ προκύπτει το διαχωριστικό υπερεπίπεδο. Το w_0 μπορεί να υπολογιστεί αντικαθιστώντας ένα από τα support vectors στην παρακάτω σχέση:

$$y_j((w \cdot x_j) + w_0) = 1 \quad (37)$$

3.4.4 Fisher's Linear Discriminant Analysis

Όπως αναφέρθηκε νωρίτερα, ένας τρόπος μείωσης της διάστασης του χώρου χαρακτηριστικών είναι η εξαγωγή χαρακτηριστικών (feature extraction). Μία πλέον διαδεδομένη μέθοδος για εξαγωγή χαρακτηριστικών είναι η Principal Component Analysis (PCA), σύμφωνα με την οποία εντοπίζονται κατευθύνσεις μέγιστης διασποράς των δεδομένων μέσα στον N-διάστατο χώρο χαρακτηριστικών και γίνεται προβολή των δεδομένων σε μία ευθεία αυτής της κατεύθυνσης. Η PCA είναι μία μέθοδος η οποία οδηγεί στην αναπαράσταση των δεδομένων μας σε ένα χώρο λιγότερων διαστάσεων, διατηρώντας την μέγιστη διασπορά τους, έτσι ώστε να μπορούμε να βγάλουμε συμπεράσματα για τον τρόπο κατανομής των δεδομένων στον N-διάστατο χώρο. Παρά το γεγονός ότι δίνει μια σαφή αναπαράσταση των δεδομένων σε λιγότερες διαστάσεις, δεν είναι μέθοδος που προτείνεται για ταξινόμηση [33].

Γενικά η προβολή των δεδομένων πάνω σε μία διάσταση οδηγεί σε χάσιμο πληροφορίας, με αποτέλεσμα δεδομένα κλάσεων που διαχωρίζονται καλά στον N-διάστατο χώρο να παρουσιάζουν μεγάλη επικάλυψη σε μία διάσταση. Παρόλα αυτά, προσαρμόζοντας τα βάρη του διανύσματος w, μπορεί να επιλεχθεί μια κατεύθυνση που να μεγιστοποιεί το διαχωρισμό των κλάσεων [bishop]. Αυτός είναι και ο στόχος της Fisher's Linear Discriminant Analysis (FLDA). Η διαφορά της FLDA με την PCA φαίνεται στο Σχήμα 11.



Σχήμα 11. Παράδειγμα σε δύο διαστάσεις. Η αριστερή γραμμή (PCA) δεν προτείνεται για διαχωρισμό των κλάσεων, η δεξιά διαχωρίζει βέλτιστα τις κλάσεις (FLDA)

Ας θεωρήσουμε λοιπόν ότι έχουμε ένα πρόβλημα ταξινόμησης σε δύο κλάσεις ω_1 και ω_2 με M_1 δεδομένα στην πρώτη και M_2 στην δεύτερη. Έστω ότι μ_1 και μ_2 είναι τα διανύσματα μέσων τιμών (σχέσεις 6 και 7) των κλάσεων ω_1 και ω_2 αντίστοιχα. Η προβολή ενός διανύσματος x σε μία ευθεία κατεύθυνσης w δίνεται από τη σχέση

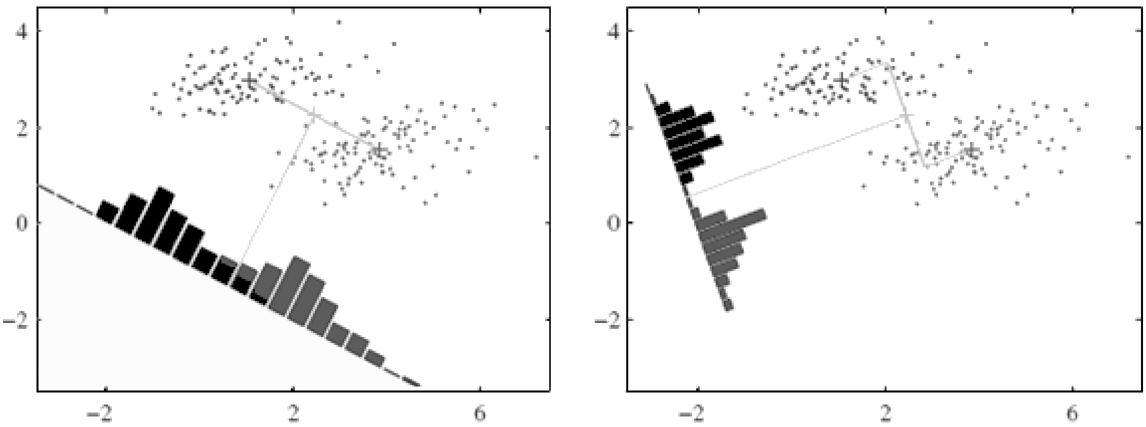
$$y = w^T x \tag{38}$$

Η πιο απλή μέθοδος εκτίμησης του πόσο καλά διαχωρίζονται οι κλάσεις μεταξύ τους όταν τα δεδομένα προβάλλονται σε μία τέτοια ευθεία είναι η διαφορά των προβαλλόμενων μέσων τιμών των κλάσεων. Τότε το w επιλέγεται τέτοιο ώστε να μεγιστοποιεί το εξής [24]:

$$\mu_{y1} - \mu_{y2} = w^T(\mu_1 - \mu_2) \tag{39}$$

όπου $$\mu_{yi} = w^T \mu_i \tag{40}$$

είναι η μέση τιμή των προβαλλόμενων δεδομένων από την κλάση ω_i, i=1,2 στην ευθεία κατεύθυνσης w.



Σχήμα 12. Η απόσταση των προβολών των μ_1 και μ_2 είναι μεγαλύτερη στην αριστερή εικόνα, παρόλα αυτά στην εικόνα δεξιά γίνεται καλύτερος διαχωρισμός των κλάσεων

Παρόλο που με μια πρώτη ματιά αυτό το κριτήριο φαίνεται ικανοποιητικό, δεν είναι πάντα, όπως φαίνεται και στο Σχήμα 12. Η διασπορά των προβαλλόμενων κλάσεων στην αριστερή εικόνα είναι μεγαλύτερη από την διασπορά των προβαλλόμενων κλάσεων στην δεξιά εικόνα, με αποτέλεσμα στην πρώτη περίπτωση να έχουμε επικάλυψη των δύο κλάσεων. Αυτό δικαιολογεί το γεγονός ότι ενώ στην πρώτη περίπτωση έχουμε μεγαλύτερη απόσταση μεταξύ των προβολών των μ_1 και μ_2, στην δεύτερη περίπτωση έχουμε καλύτερο διαχωρισμό των κλάσεων.

Η ιδέα που προτείνεται από τον Fisher είναι η μεγιστοποίηση μιας συνάρτησης που θα δίνει μεγάλο διαχωρισμό μεταξύ των προβαλλόμενων μέσων τιμών των κλάσεων, ενώ παράλληλα θα εξασφαλίζει μικρή διασπορά των κλάσεων, ελαχιστοποιώντας επομένως την επικάλυψή τους.

Η εντός-κλάσης διασπορά (within-class variance) των προβαλλόμενων δεδομένων της κλάσης ω_k σε μία διάσταση δίνεται από την σχέση

$$s_{yk}^2 = \sum_{\{i:y_i \in \omega_k\}} (y_i - \mu_{yk})^2 \quad , \quad k = 1,2 \tag{41}$$

όπου y_i οι προβολή στην ευθεία του x_i και μ_{yk} η προβολή της μέσης τιμής των δεδομένων που ανήκουν στην κλάση ω_k. Η συνολική εντός-κλάσης διασπορά είναι απλά το άθροισμα $s_{y1}^2 + s_{y2}^2$.

Το κριτήριο του Fisher (Fisher criterion) ορίζεται να είναι ο λόγος του τετραγώνου της διαφοράς των προβαλλόμενων μέσων τιμών προς την συνολική εντός-κλάσης διασπορά και δίνεται από την σχέση

$$J(w) = \frac{(\mu_{y1} - \mu_{y2})^2}{s_{y1}^2 + s_{y2}^2} \tag{42}$$

Μπορεί να γίνει σαφής η εξάρτηση της παραπάνω σχέσης από το διάνυσμα κατεύθυνσης *w*, χρησιμοποιώντας τις σχέσεις 38, 40 και 41 ώστε να ξαναγραφεί το κριτήριο Fisher στην εξής μορφή

$$J(w) = \frac{w^T S_B w}{w^T S_W w} \tag{43}$$

όπου S_B είναι ο μεταξύ των κλάσεων (between-class) πίνακας συνδιασποράς, αποτελεί μέτρο του διαχωρισμού ανάμεσα στις μέσες τιμές των δύο κλάσεων ω_1 και ω_2 πριν την προβολή τους και δίνεται από τη σχέση:

$$S_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^{\mathrm{T}} \tag{44}$$

και S_W είναι ο συνολικός εντός-κλάσης πίνακας συνδιασποράς, και δίνεται από τη σχέση

$$S_W = \sum_{\{i:x_i \in \omega_1\}} (x_i - \mu_1)(x_i - \mu_1)^T + \sum_{\{i:x_i \in \omega_2\}} (x_i - \mu_2)(x_i - \mu_2)^T \tag{45}$$

Παραγωγίζοντας τη σχέση 43 ως προς το w και εξισώνοντας με το 0, προκύπτει ότι η *J(w)* μεγιστοποιείται όταν

$$(w^T S_B w) S_w w = (w^T S_w w) S_B w \tag{46}$$

Από τη σχέση 44 συμπεραίνουμε ότι το $S_B w$ είναι πάντα στην κατεύθυνση του $(\mu_1 - \mu_2)$. Επιπλέον δε μας ενδιαφέρει το μέτρο του *w* παρά μόνο η κατεύθυνσή του οπότε μπορούμε να απαλείψουμε τις φυσικές ποσότητες (όχι διανύσματα) $w^T S_B w$ και $w^T S_W w$. Τελικά πολλαπλασιάζοντας με S_W^{-1} και τα δύο σκέλη της 9 παίρνουμε την εξής σχέση

$$w \propto S_W^{-1}(\mu_1 - \mu_2) \tag{47}$$

Δηλαδή το w είναι ανάλογο του $S_W^{-1}(\mu 1 - \mu 2)$. Από τη στιγμή που δεν μας ενδιαφέρει το μέτρο του w αλλά μόνο η κατεύθυνσή του, μπορεί να θεωρηθεί ότι

$$w = S_W^{-1}(\mu_1 - \mu_2) \tag{48}$$

Τελικά εκείνη η ευθεία στην οποία αν προβληθούν τα δεδομένα x_i θα υπάρχει βέλτιστος διαχωρισμός ανάμεσα στις κλάσεις ω_1 και ω_2 είναι η ευθεία με κατεύθυνση w. Η ευθεία η οποία διαχωρίζει τις δύο κλάσεις θα είναι κάθετη σε αυτήν [34].

Εάν λοιπόν θεωρήσουμε ότι τα δεδομένα $X=\{x_1, x_2, \ldots, x_M\}$ είναι τα πρότυπα προς εκπαίδευση, με εφαρμογή της FLDA καταλήγουμε στο w της εξίσωσης (48) και τελικά η διαχωριστική συνάρτηση θα είναι η εξής

$$D(x) = w^T x + w_0 \tag{49}$$

όπου το w_0 υπολογίζεται από την σχέση:

$$w_0 = -\frac{1}{2} w^T (\mu_1 + \mu_2) \tag{50}$$

4 Μεθοδολογία Υλοποίησης

Στο προηγούμενο κεφάλαιο παρουσιάστηκε κάποια 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, τα χαρακτηριστικά τους είναι οι σχετικές εκφράσεις των γονιδίων και πρόκειται να γίνει ταξινόμηση σε δύο κλάσεις, ανάλογα με τη σχετική έκφραση των γονιδίων.

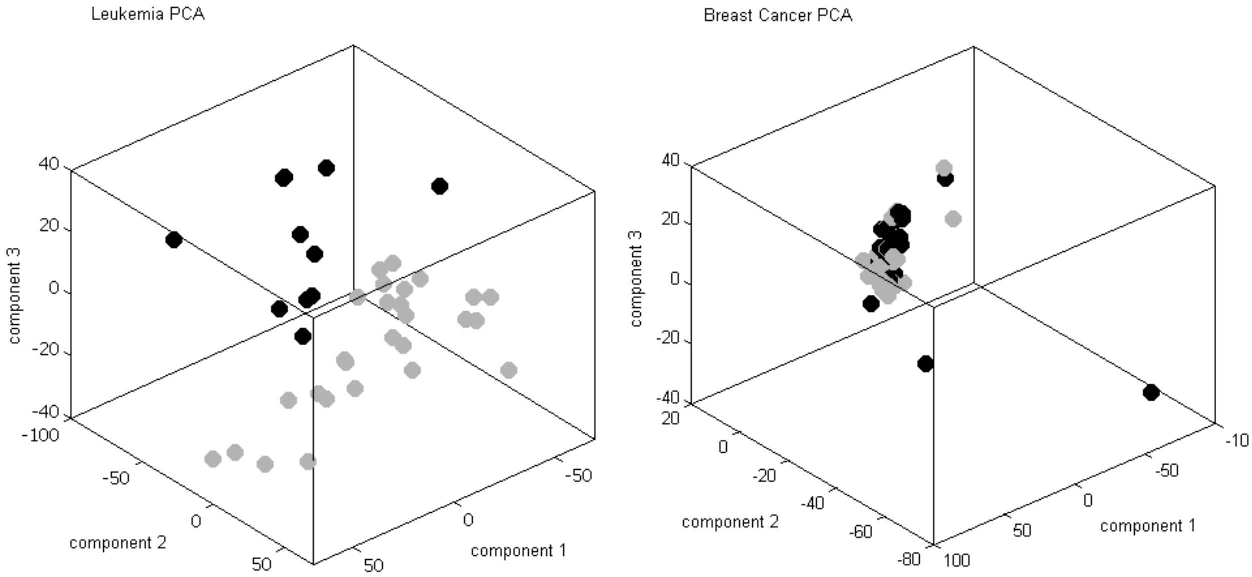
Αρχικά γίνεται μια περιγραφή των δεδομένων της λευχαιμίας και του καρκίνου του μαστού και στη συνέχεια περιγράφεται η μεθοδολογία που ακολουθείται σε αυτή τη διπλωματική εργασία.

4.1 Περιγραφή των δεδομένων

Τα δεδομένα της λευχαιμίας και του καρκίνου του μαστού τα οποία εξετάζονται στην παρούσα διπλωματική εργασία, προέρχονται από τις δημοσιεύσεις των Golub [20] και Van't Veer [6] αντίστοιχα. Και για τα δύο δεδομένα είναι στη διάθεσή μας ένα σύνολο εκπαίδευσης και ένα ανεξάρτητο σύνολο δοκιμής.

Για τα δεδομένα της λευχαιμίας έχουμε συνολικά 7129 γονίδια και το train set αποτελείται από 38 δείγματα (27 ALL και 11 AML) ενώ το test set από 34 δείγματα (20 ALL και 14 AML). 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 μείωση στις τρεις διαστάσεις μέσω PCA μεθοδολογίας, τόσο για τα δεδομένα του καρκίνου του μαστού όσο και για αυτά της λευχαιμίας, για να βγει κάποιο συμπέρασμα για την κατανομή των δεδομένων στο χώρο. Τα αποτελέσματα φαίνονται στο Σχήμα 13. Παρατηρούμε ότι όσον αφορά τα δεδομένα της λευχαιμίας, οι δύο κλάσεις εμφανίζουν μεν μεγάλη διασπορά, είναι όμως καλά διαχωρισμένες μεταξύ τους, ενώ για τα δεδομένα του καρκίνου του μαστού υπάρχει μεγάλη επικάλυψη ανάμεσα στις δύο κλάσεις και μάλιστα η διασπορά τους είναι μικρή με αποτέλεσμα να είναι δύσκολο να διαχωριστούν μεταξύ τους.



Σχήμα 13. Απεικόνιση των τριών principal components των δεδομένων της λευχαιμίας (αριστερά) και του καρκίνου του μαστού (δεξιά)

Συμπεραίνουμε λοιπόν ότι τα δεδομένα του καρκίνου του μαστού είναι πολύ πιο δύσκολο να διαχωριστούν από αυτά της λευχαιμίας. Συνεπώς, στην περίπτωση του καρκίνου του μαστού το πρόβλημα της επιλογής γονιδίων είναι πολύ πιο ενδιαφέρον και κρίσιμο ούτως ώστε να βρεθεί ένα καλύτερο υποσύνολο γονιδίων, βάση του οποίου να γίνεται καλός διαχωρισμός των δύο κλάσεων.

Επομένως 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 $y_i \in \{-1,+1\}$. Το κάθε γονίδιο θα συμβολίζεται με g_i , $i=1,2,\ldots,N$. Οπότε η έκφραση του γονιδίου g_i για τον ασθενή x_j, $j=1,2,\ldots,M$ θα είναι το i-οστό στοιχείο του διανύσματος x_j, δηλαδή το $(x_j)_i$.

4.2 Τυπική διαδικασία 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 [35].

Ακολουθείται λοιπόν επιλογή 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. Η αξιολόγηση του ποια γονίδια θα αφαιρεθούν γίνεται με δύο μεθόδους, μία filter και μία wrapper.

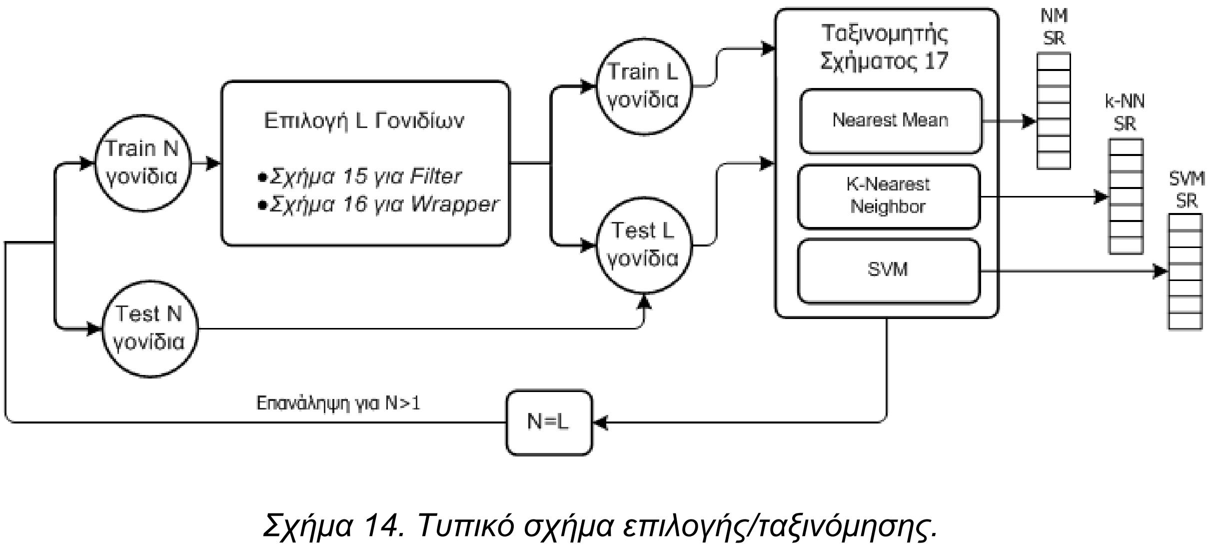
Σε κάθε στάδιο αυτής της επαναληπτικής διαδικασίας, είναι αναγκαίο να εκτιμηθεί η απόδοση κάποιου ταξινομητή χρησιμοποιώντας μόνο τα εναπομείναντα γονίδια. Έτσι θα μπορούν να προκύψουν συμπεράσματα για την πληροφορία που κρύβεται σε αυτά τα γονίδια, σχετικά με το πρόβλημα της ταξινόμησης. Στην παρούσα διπλωματική αυτό γίνεται χρησιμοποιώντας τρεις ταξινομητές, τους Nearest Mean (NM), k-Nearest Neighbor (*k*-NN) και SVM . Η ανάλυση των παραπάνω έγινε στο προηγούμενο κεφάλαιο.

Όσον αφορά τα δεδομένα της λευχαιμίας, η επιλογή της παραμέτρου *k* του *k*-Nearest Neighbor πρέπει να είναι τέτοια ώστε να μην υπερβαίνει τον μέγιστο αριθμό των επιτρεπτών γειτόνων (στην περίπτωσή μας οι περισσότεροι γείτονες που μπορεί να υπάρξουν είναι 38, όσα είναι και τα δείγματα του train set), αλλά και να καλύπτει όλο το φάσμα επιτρεπτών τιμών, ώστε να έχουμε μια σαφή εικόνα της συμπεριφοράς του ταξινομητή για διάφορες τιμές της παραμέτρου *k*. Επομένως για το *k* επιλέγονται οι τιμές 1, 5, 11 και 37. Για τα δεδομένα του καρκίνου του μαστού, η μέγιστη τιμή που μπορεί να πάρει η παράμετρος *k* του *k*-Nearest Neighbor είναι 78 (όσα είναι και τα δείγματα του train set) οπότε επιλέγονται οι τιμές 1, 5, 11, 37, 51 και 71.

Επίσης για την παράμετρο *C* του SVM επιλέγονται οι τιμές 0.001, 1, 100 και 1000. Επομένως για τα δεδομένα της λευχαιμίας εκτιμάται η απόδοση συνολικά 9 ταξινομητών (NM, k-NN για 4 τιμές της παραμέτρου *k*, και SVM για 4 τιμές της παραμέτρου *C*), ενώ για τα δεδομένα του καρκίνου του μαστού εκτιμάται η απόδοση συνολικά 11 ταξινομητών (NM, k-NN για 6 τιμές της παραμέτρου *k*, και SVM για 4 τιμές της παραμέτρου *C*).

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 των ταξινομητών και εκτίμηση της απόδοσής τους.

Έστω λοιπόν ότι έχουμε *I* επαναλήψεις της διαδικασίας επιλογής. Στην *i*-οστή επανάληψη (*i=1,2,…,I*) και για κάθε ταξινομητή, αποθηκεύεται το ποσοστό επιτυχίας που επιτυγχάνεται, στην *i*-οστή θέση ενός διανύσματος *SR* (από το success rate, ποσοστό επιτυχίας) μεγέθους *I*. Η διαδικασία που μόλις περιγράφηκε φαίνεται στο Σχήμα 15.



Σχήμα 14. Τυπικό σχήμα επιλογής/ταξινόμησης.

Μετά το τέλος της διαδικασίας, εντοπίζεται το μέγιστο ποσοστό επιτυχίας για κάθε ταξινομητή καθώς και σε ποιο βήμα της επιλογής (για ποια τιμή του i) αυτό επιτυγχάνεται.

Είναι πιθανό να μην έχουν όλοι οι ταξινομητές την μέγιστη απόδοση για το ίδιο σύνολο γονιδίων.

Ακολουθεί η ανάλυση των δύο μεθόδων, filter και wrapper, για την επιλογή των γονιδίων, καθώς και ο τρόπος με τον οποίο γίνεται η εκπαίδευση κάθε ταξινομητή και η εκτίμηση της απόδοσής του.

4.2.1 Filter μέθοδος – Fisher's Criterion

Στην filter μέθοδο τα γονίδια κατατάσσονται με βάση κάποιο κριτήριο, όπως κάποιο συντελεστή συσχέτισης [25]. 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 μεγαλύτερο είναι αυτό, 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.

Το κριτήριο που χρησιμοποιείται για την ταξινόμηση των γονιδίων είναι το Fisher's criterion, το οποίο υπολογίζεται βάση του train set και μόνο και που για κάθε γονίδιο g_i υπενθυμίζουμε ότι δίνεται από την σχέση:

$$F_i = \frac{((\mu_+)_i - (\mu_-)_i)^2}{(\sigma_+^2)_i + (\sigma_-^2)_i}, i = 1, 2, \dots, N \tag{51}$$

όπου

$$\begin{aligned} \mu_+ &= \frac{1}{M_+} \sum_{\{j:y_j=+1\}} x_j \\ \mu_- &= \frac{1}{M_-} \sum_{\{j:y_j=-1\}} x_j \end{aligned} \quad , j=1,2,\dots,M \,, \quad M_- + M_+ = M \tag{52}$$

τα διανύσματα μέσων τιμών των κλάσεων (+) και (-) αντίστοιχα και

$$\left.\begin{aligned}\sigma_+^2 &= \frac{1}{M_+ - 1}\sum_{\{j:y_j=+1\}}(x_j-\mu_+)^2\\ \sigma_-^2 &= \frac{1}{M_- - 1}\sum_{\{j:y_j=-1\}}(x_j-\mu_-)^2\end{aligned}\right. \quad , \; j=1,2,\ldots,M\,, \quad M_- + M_+ = M \tag{53}$$

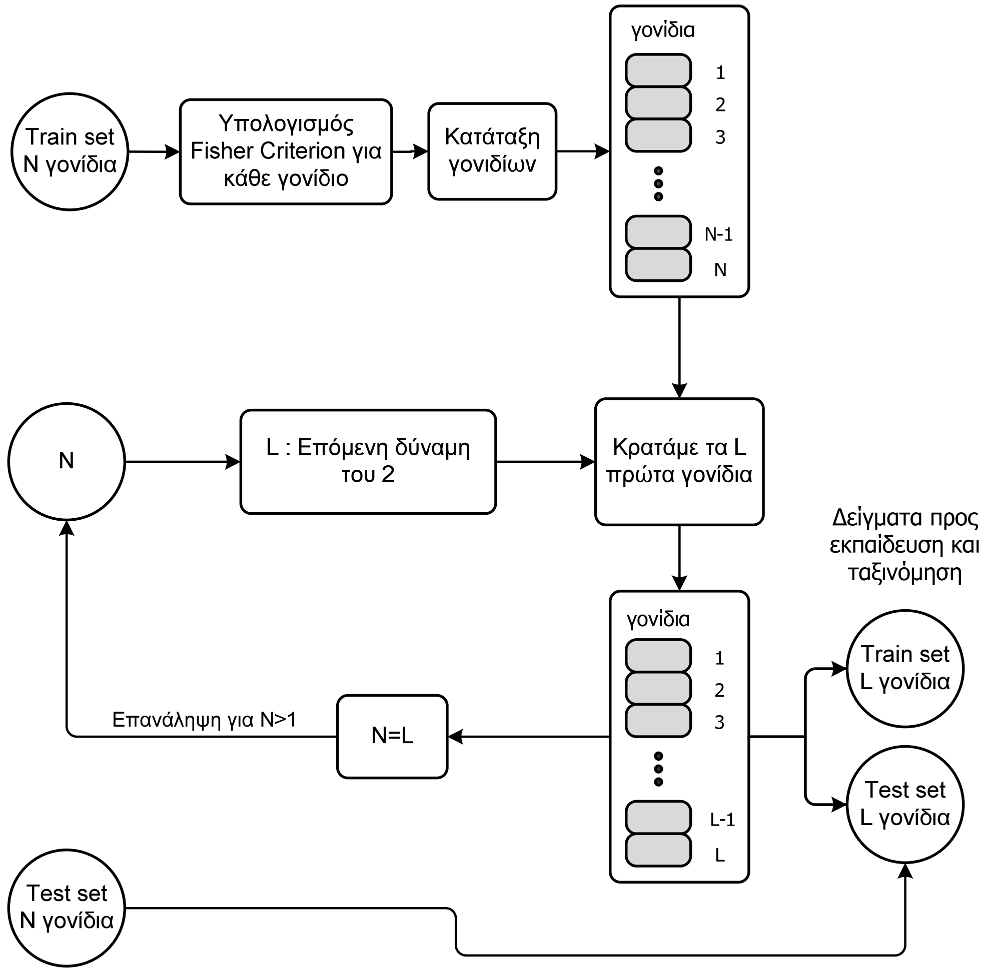
τα διανύσματα διασποράς των κλάσεων (+) και (-) αντίστοιχα. Με M_- και M_+ συμβολίζεται το πλήθος των προτύπων του train set που ταξινομούνται στην αρνητική και θετική κλάση αντίστοιχα.

Επομένως το F είναι ένα διάνυσμα N συντελεστών όπου F_i αντιστοιχεί στο Fisher's criterion για το γονίδιο g_i, $i=1,2,\ldots,N$. Τελικά τα γονίδια ταξινομούνται ξεκινώντας από αυτό με το μεγαλύτερο F_i και καταλήγοντας σε αυτό με το μικρότερο F_i.

Έχοντας κατατάξει λοιπόν τα γονίδια από το πιο σημαντικό στο λιγότερο σημαντικό, το επόμενο βήμα είναι η αφαίρεση ενός πλήθους λιγότερο σημαντικών γονιδίων (με το μικρότερο F_i). Στη συγκεκριμένη διπλωματική επιλέγουμε την αφαίρεση τόσων γονιδίων ώστε να φτάσουμε στην αμέσως μικρότερη δύναμη του 2. Οπότε καταλήγουμε σε ένα υποσύνολο των αρχικών γονιδίων. Από αυτό το νέο σύνολο γονιδίων αφαιρούνται πάλι τόσα – αυτά πάντα με το μικρότερο F_i, το οποίο δεν αλλάζει για κάθε γονίδιο – ώστε να φτάσουμε στην επόμενη δύναμη του 2. 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.

Σε κάθε επανάληψη της διαδικασίας υπολογίζονται τα νέα train και test sets για τον έλεγχο της απόδοσης των τριών ταξινομητών NM, KNN και SVM.

Στο Σχήμα 15 φαίνεται συνολικά η διαδικασία που περιγράφηκε παραπάνω.



Σχήμα 15. Επιλογή γονιδίων με filter μέθοδο

4.2.2 Wrapper μέθοδος – Fisher's Linear Discriminant Analysis

Όπως έχουμε ήδη εξηγήσει, η wrapper μέθοδος για την επιλογή γονιδίων απαιτεί την εκπαίδευση κάποιου ταξινομητή. Η μέθοδος ταξινόμησης που χρησιμοποιούμε είναι η Fisher's Linear Discriminant που περιγράφηκε στο προηγούμενο κεφάλαιο. Όπως ήδη έχουμε αναφέρει, με εφαρμογή της FLDA βρίσκεται εκείνη η ευθεία πάνω στην οποία οι προβολές των δειγμάτων της μίας κλάσης διαχωρίζονται όσο γίνεται καλύτερα από αυτές της άλλης κλάσης.

Και πάλι χρησιμοποιείται το train set για την επιλογή των γονιδίων. Με το train γίνεται η εκπαίδευση του ταξινομητή και έτσι προκύπτουν κάποια βάρη, που αποτελούν την κλίση της βέλτιστης ευθείας στην οποία τα πρότυπα προβάλλονται. Οπότε εξάγεται ένα διάνυσμα βαρών w διάστασης N – όσα είναι και τα γονίδια. Η συνάρτηση απόφασης είναι η εξής

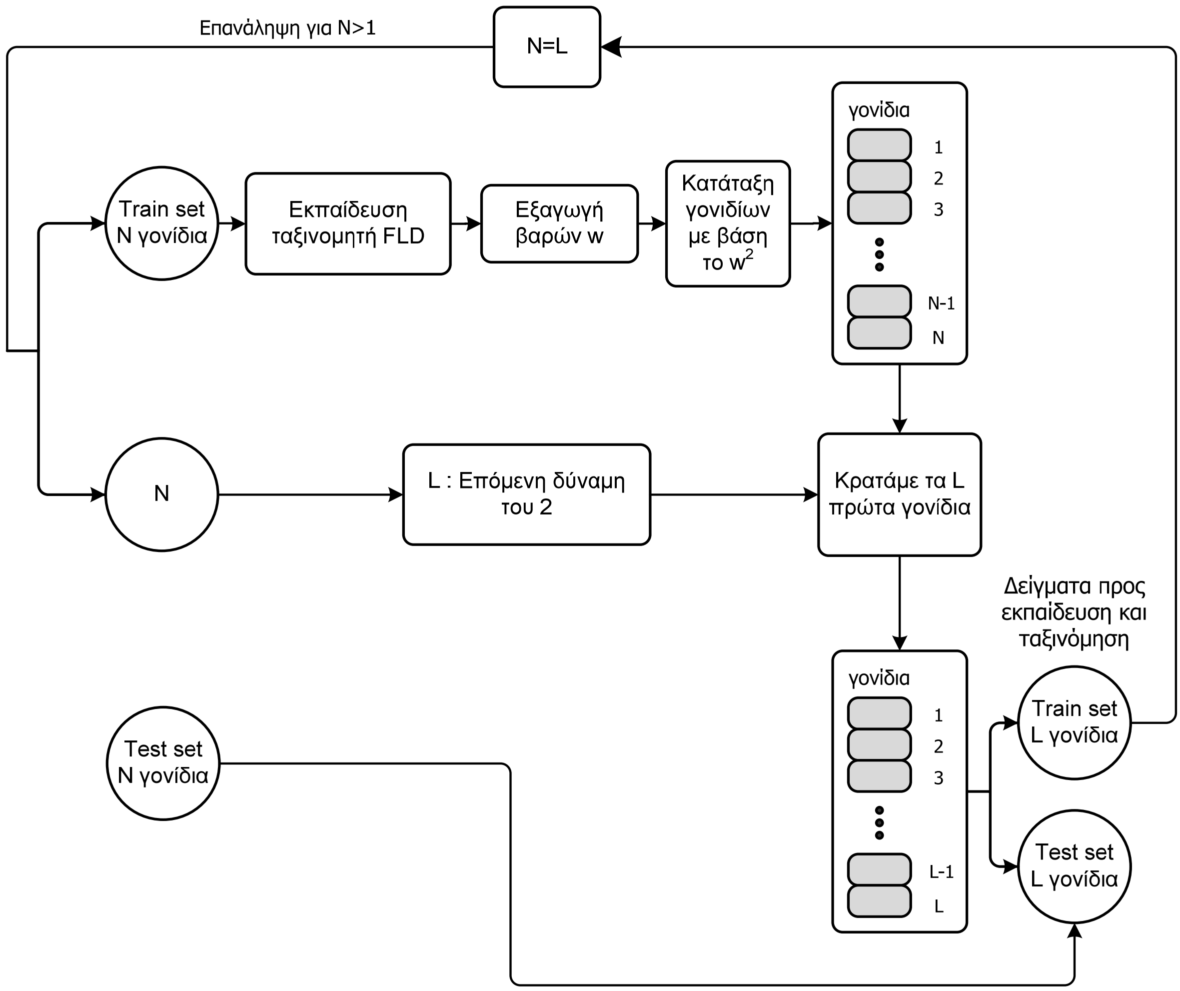
$$D(x) = w^T x + w_0 \tag{54}$$

όπου για D(x)>0 το πρότυπο x ταξινομείται στην θετική κλάση και για D(x)<0 ταξινομείται στην αρνητική κλάση.

Οπότε το *i*-οστό στοιχείο του δείγματος *x* πολλαπλασιάζεται από το *i*-οστό στοιχείο του *w*. Άρα σε κάθε γονίδιο g_i αντιστοιχεί ένα βάρος w_i. Η ταξινόμηση των γονιδίων με βάση το τετράγωνο του βάρους που αντιστοιχεί στο καθένα είναι μια βασική μέθοδος, επειδή τα βάρη προκύπτουν από γραμμικό ταξινομητή [25]. Επομένως χρησιμοποιούμε ως κριτήριο το w_i^2 και τελικά τα γονίδια ταξινομούνται ξεκινώντας από αυτό που έχει το μεγαλύτερο w_i^2 και καταλήγοντας σε αυτό με το μικρότερο w_i^2.

Η διαδικασία που ακολουθείται είναι ίδια με αυτή της filter μεθόδου, μόνο που εδώ κάθε φορά που μειώνονται τα γονίδια πρέπει να υπολογίζονται νέα βάρη. Σε κάθε επανάληψη της μεθόδου υπολογίζονται τα νέα train και test sets για τον έλεγχο της απόδοσης των τριών ταξινομητών NM, KNN και SVM. Το νέο train set που προκύπτει χρησιμοποιείται για εκπαίδευση του ταξινομητή FLD και για εξαγωγή των νέων βαρών για τα εναπομείναντα γονίδια. Η διαδικασία επαναλαμβάνεται έως ότου καταλήξουμε σε 1 γονίδιο. Όπως στην filter έτσι και εδώ τελικά προκύπτει ένα πλήθος συνόλων γονιδίων, καθένα από τα οποία είναι υποσύνολο των μεγαλυτέρων του και γενικά υποσύνολο των αρχικών γονιδίων.

Στο Σχήμα 16 φαίνεται συνολικά η διαδικασία που περιγράφηκε παραπάνω.



Σχήμα 16. Επιλογή γονιδίων με wrapper μέθοδο

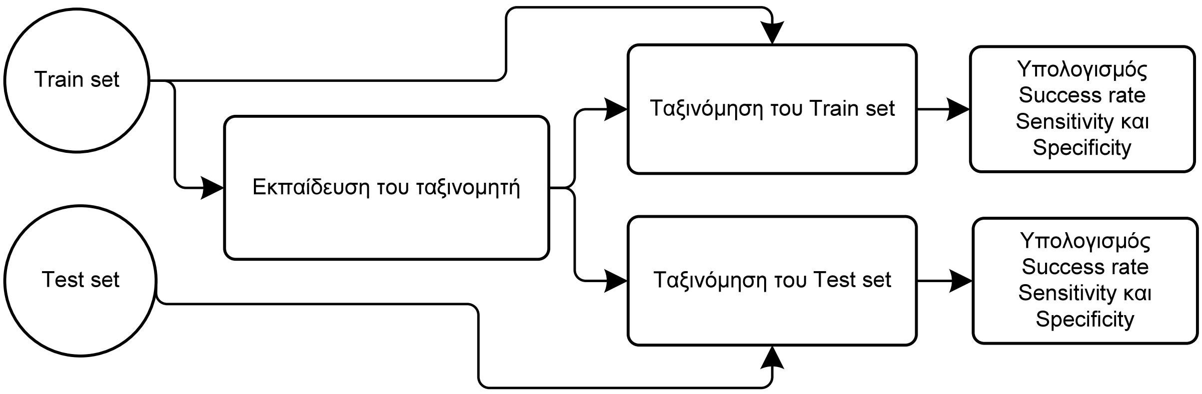
4.2.3 Ταξινόμηση των δειγμάτων σε κάθε βήμα επιλογής γονιδίων

Όπως αναφέρθηκε, σε κάθε βήμα της επιλογής γονιδίων γίνεται εκτίμηση της απόδοσης των ταξινομητών Nearest Mean, k-Nearest Neighbor και SVM.

Πιο συγκεκριμένα, με το train set εκπαιδεύεται ο κάθε ταξινομητής ξεχωριστά και στη συνέχεια ελέγχεται η απόδοσή του, 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.

Επιπλέον μας ενδιαφέρει να ελέγξουμε το κατά πόσο καλά μπορεί ο κάθε ταξινομητής να ταξινομήσει τα δεδομένα μιας συγκεκριμένης κλάσης. Για το σκοπό αυτό, πέρα από το ποσοστό επιτυχίας (success rate, SR), υπολογίζονται και δύο επιπρόσθετα ποσοστιαία μεγέθη, το Sensitivity (SE) και το Specificity (SP). Το πρώτο αφορά στο ποσοστό των δειγμάτων του set τα οποία πρέπει να ταξινομηθούν στη θετική κλάση και που όντως ταξινομούνται σε αυτήν, ενώ το δεύτερο είναι το αντίστοιχο ποσοστό για τα δεδομένα που ταξινομούνται στην αρνητική κλάση.

Στο Σχήμα 17 φαίνεται αναλυτικά η διαδικασία υπολογισμού των SR, SE και SP για δεδομένο ταξινομητή.

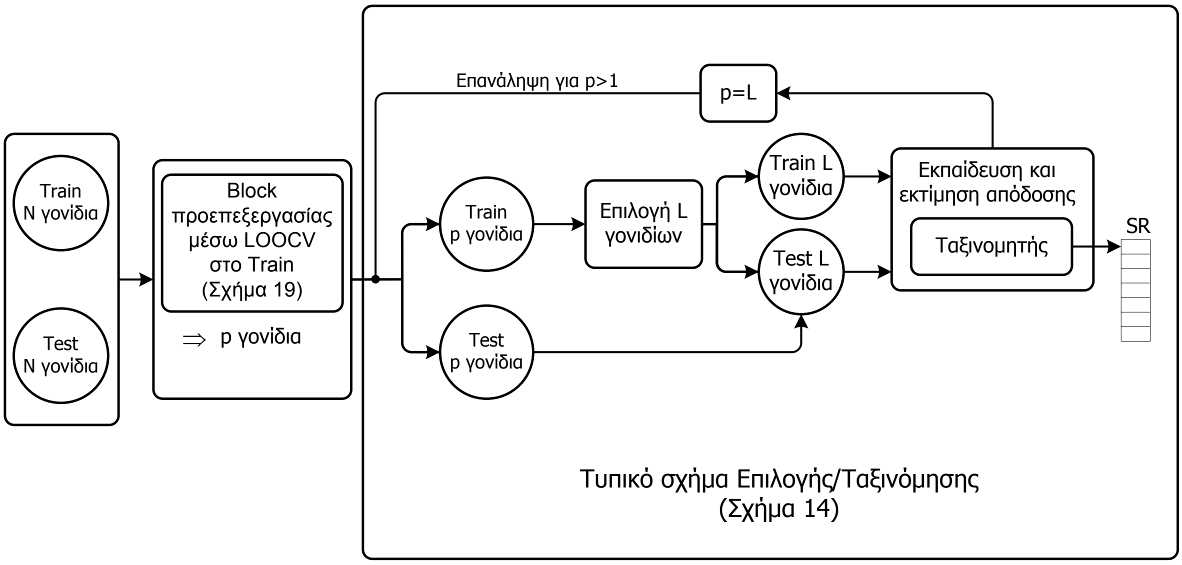


Σχήμα 17. Υπολογισμός SR, SE, SP για δεδομένο ταξινομητή

4.3 Διαδικασία επιλογής γονιδίων με προεπεξεργασία δειγμάτων

Η παραπάνω ανάλυση αφορά σε μία τυπική διαδικασία επιλογής γονιδίων. Η καινοτομία αυτής της διπλωματικής εργασίας είναι η πρόταση για μια προεπεξεργασία των δειγμάτων εκπαίδευσης, ούτως ώστε να προκύψει ένα πιθανά πιο καλό σύνολο γονιδίων, από το οποίο να ξεκινήσει η διαδικασία επιλογής. Πιο συγκεκριμένα, γίνεται αρχικά μια προεπεξεργασία βασισμένη στην Leave-one-out cross validation (LOOCV) μέθοδο, με σκοπό να βρεθεί ένα υποσύνολο των αρχικών γονιδίων και να εφαρμοστεί η διαδικασία που αναλύθηκε προηγουμένως – έχοντας ως αφετηρία αυτό το σύνολο – με την προσδοκία να καταλήξουμε σε καλύτερα αποτελέσματα.

Αυτό το προεπιλεγμένο σύνολο – έστω p σε πλήθος – γονιδίων πλέον θα είναι η αφετηρία του τυπικού σχήματος επιλογής – ταξινόμησης που περιγράφηκε στην ενότητα 1.1. Από το σύνολο εκπαίδευσης, καθώς και από το ανεξάρτητο σύνολο δοκιμής, υπολογίζονται τα νέα train και test sets, των οποίων τα δείγματα θα αποτελούνται μόνο από τις εκφράσεις των p γονιδίων που προκύπτουν από την προεπιλογή. Αυτή η διαδικασία φαίνεται στο Σχήμα 18.



Σχήμα 18. Ολικό σχήμα προεπιλογής γονιδίων με LOOCV και μετέπειτα επιλογής / ταξινόμησης, για δεδομένο ταξινομητή.

Η διαδικασία επιλογής γονιδίων είναι λοιπόν η ίδια με αυτή που περιγράφηκε στην προηγούμενη ενότητα, μόνο που τώρα η μείωση ξεκινάει από λιγότερα γονίδια. Λόγω του ότι ορισμένες φορές το σύνολο αυτό των γονιδίων είναι αρκετά μικρό, γίνεται επιλογή γονιδίων τόσο μειώνοντας σε δυνάμεις του δύο όσο και μειώνοντας κατά ένα γονίδιο. Αυτό γίνεται σε περιπτώσεις που το σύνολο των γονιδίων είναι μικρότερο από 1500 γονίδια.

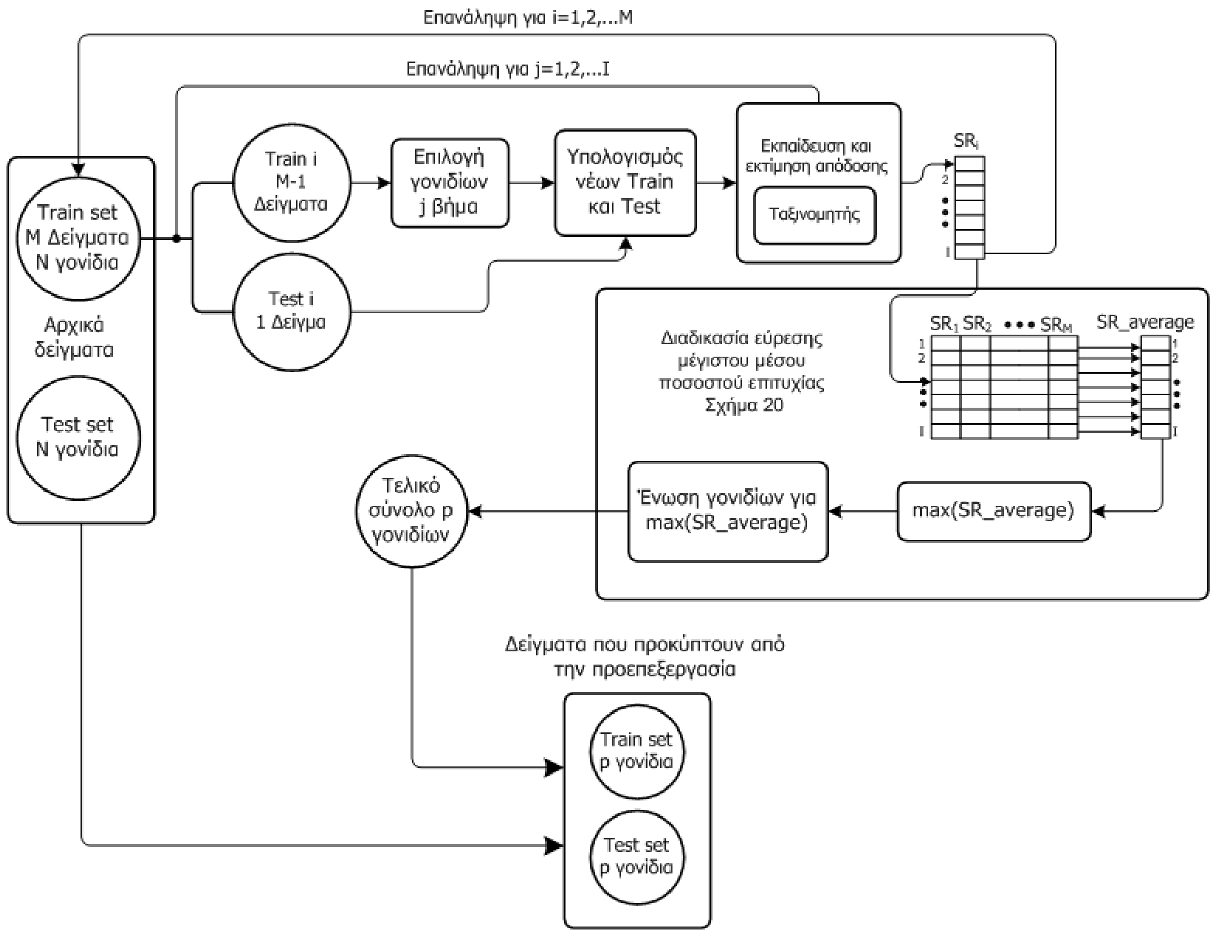
Επίσης είναι πιθανό για κάθε ταξινομητή να προκύψει διαφορετικό σύνολο γονιδίων, όπως θα εξηγήσουμε παρακάτω. Επιλέγεται για κάθε σύνολο να γίνεται εκπαίδευση μόνο του ταξινομητή από τον οποίο αυτό προέκυψε και εκτίμηση της απόδοσής του. Με αυτόν τον τρόπο είναι δυνατή μια σύγκριση των αποτελεσμάτων που έχουμε κάνοντας επιλογή γονιδίων από το αρχικό σύνολο και αυτών που έχουμε κάνοντας επιλογή από το σύνολο που προκύπτει με LOOCV στο σύνολο εκπαίδευσης, για κάθε ένα ταξινομητή ξεχωριστά.

4.3.1 Προεπεξεργασία δειγμάτων με εφαρμογή LOOCV στο σύνολο εκπαίδευσης

Εφαρμόζεται στο σύνολο εκπαίδευσης η LOOCV μέθοδος η οποία περιγράφηκε στο προηγούμενο κεφάλαιο. Συνοπτικά υπενθυμίζουμε ότι εάν το σύνολο εκπαίδευσης αποτελείται από M δείγματα, κάθε φορά ορίζεται ως test set ένα δείγμα και ως train set τα υπόλοιπα $M – 1$ δείγματα. Επομένως προκύπτουν M διαφορετικά ζεύγη train και test set. Για κάθε ζεύγος train και test ακολουθεί η τυπική διαδικασία επιλογής – ταξινόμησης που περιγράφηκε στην προηγούμενη ενότητα (1.1). Επομένως αυτή η διαδικασία επαναλαμβάνεται M φορές, μία για κάθε ζεύγος train και test.

Έστω και πάλι ότι έχουμε I επαναλήψεις της διαδικασίας επιλογής γονιδίων, τότε για το i-οστό ζεύγος train-test ($i=1,2,\ldots,M$) αποθηκεύεται η απόδοση (ποσοστό επιτυχίας) του κάθε ταξινομητή σε κάθε βήμα της επιλογής, σε ένα διάνυσμα SR_i μεγέθους I. Επειδή έχουμε μόνο ένα δείγμα για δοκιμή, το ποσοστό επιτυχίας θα είναι είτε 1 (100% επιτυχία) είτε 0. Αφού συμπληρωθεί το διάνυσμα SR_i, δηλαδή έχει ολοκληρωθεί η διαδικασία επιλογής γονιδίων, ξεκινάμε και πάλι την διαδικασία για το i+1 ζεύγος train και test. Οπότε προκύπτουν M διαφορετικά διανύσματα SR_i , $i=1,2,\ldots,M$ μεγέθους I, ένα για κάθε ζεύγος train – test set.

Κάθε φορά που συμπληρώνεται το διάνυσμα SR_i, αποθηκεύεται σε έναν πίνακα μεγέθους $I \times M$, του οποίου οι στήλες είναι τα M διανύσματα SR_i που προκύπτουν ενώ οι γραμμές εκφράζουν τα βήματα της επιλογής γονιδίων (Σχήμα 19).



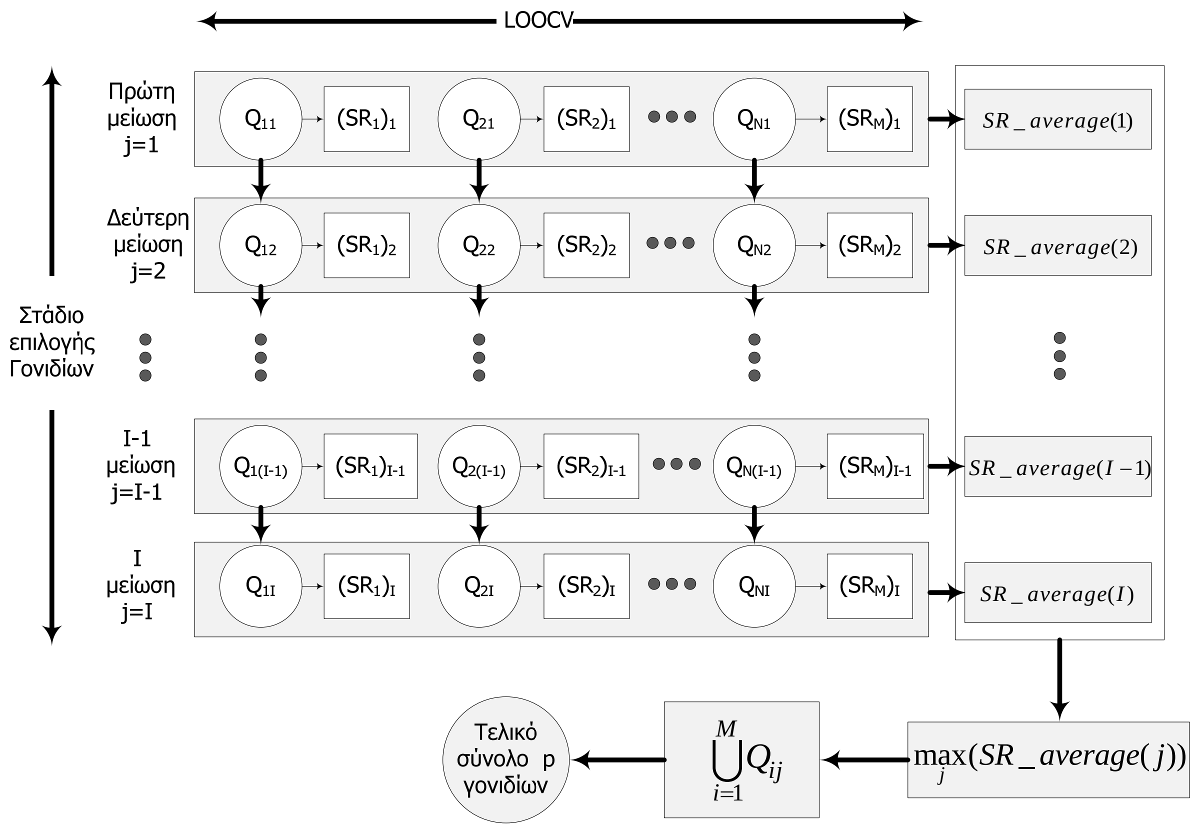
Σχήμα 19. Block προεπεξεργασίας. Η εκτίμηση της απόδοσης γίνεται και με τους τρεις ταξινομητές, για λόγους απλούστευσης στο σχήμα φαίνεται μόνο ένας.

Επομένως για κάθε ταξινομητή προκύπτει ένας πίνακας $I \times M$ με ποσοστά επιτυχίας. Το επόμενο βήμα τώρα είναι να εξετάσουμε κάθε γραμμή του πίνακα. Κάθε γραμμή περιέχει την απόδοση του εκάστοτε ταξινομητή σε ένα συγκεκριμένο βήμα επιλογής γονιδίων για τα M διαφορετικά ζεύγη train – test set που προέκυψαν από το LOOCV. Στόχος μας είναι να δούμε σε πιο βήμα επιλογής γονιδίων (δηλαδή σε ποια γραμμή του πίνακα) επιτυγχάνουμε την μέγιστη μέση απόδοση. Υπολογίζεται λοιπόν ένα νέο διάνυσμα *SR_average* μεγέθους I το οποίο περιέχει την μέση τιμή της απόδοσης σε κάθε βήμα της επιλογής γονιδίων. Πιο συγκεκριμένα η κάθε καταχώρηση αυτού του διανύσματος δίνεται από την σχέση

$$SR\_average_j = \frac{1}{M}\sum_{i=1}^{M}(SR_i)_j \quad , \quad j = 1,2,\ldots.I \qquad (55)$$

όπου με $(SR_i)_j$ συμβολίζεται η j καταχώρηση του διανύσματος SR_i .

Έστω ότι με Q_{ij} συμβολίζεται το σύνολο γονιδίων που προκύπτει στο j βήμα της επιλογής, έχοντας σαν αφετηρία το i σύνολο εκπαίδευσης. Λόγω του ότι σε κάθε στήλη έχουμε διαφορετικά σύνολα εκπαίδευσης και δοκιμής και επειδή η επιλογή των γονιδίων εξαρτάται άμεσα από το σύνολο εκπαίδευσης, είναι πιθανό από διαφορετικά train sets να προκύπτουν κατά την επιλογή διαφορετικά σύνολα γονιδίων. Δηλαδή για ένα συγκεκριμένο βήμα j επιλογής γονιδίων, τα Q_{ij} για i=1,2,…,M είναι πιθανό να μην έχουν κοινά γονίδια, ή από την άλλη να υπάρχει μερική ή και ολική επικάλυψη ανάμεσα σε κάποια από αυτά. Για κάθε σύνολο Q_{ij} επιτυγχάνεται το ποσοστό επιτυχίας $(SR_i)_j$ όπως φαίνεται και στο Σχήμα 20, το οποίο περιγράφει τον τρόπο υπολογισμού του μέγιστου μέσου ποσοστού επιτυχίας.



Σχήμα 20. Διαδικασία υπολογισμού μέσου ποσοστού επιτυχίας για κάθε μείωση γονιδίων και εύρεση τελικού συνόλου γονιδίων

Αυτό που μας ενδιαφέρει είναι να βρούμε την ένωση των M διαφορετικών συνόλων στο βήμα εκείνο της επιλογής γονιδίων όπου θα επιτυγχάνεται μέγιστο μέσο ποσοστό επιτυχίας. Επιπλέον εάν εκείνο επιτυγχάνεται σε περισσότερα του ενός βήματα, μας ενδιαφέρει το βήμα αυτό που οδηγεί σε λιγότερα γονίδια. Για εκείνο το βήμα λοιπόν, όπως φαίνεται και στο Σχήμα 21, υπολογίζεται το εξής σύνολο P

$$P = \bigcup_{i=1}^{M} Q_{ij}, \{ j : \max_{j}(SR\_average(j))\}, \; j = 1, 2, \ldots I \quad (56)$$

Εδώ χρειάζεται να πούμε ότι η εκτίμηση της απόδοσης γίνεται και με τους τρεις ταξινομητές (Nearest Mean, k-Nearest Neighbor και SVM). Για κάθε έναν προκύπτει διαφορετικός πίνακας απόδοσης, κατά συνέπεια και διαφορετικό μέγιστο ποσοστό επιτυχίας και ίσως αυτό να επιτυγχάνεται σε διαφορετικό βήμα της επιλογής γονιδίων. Επομένως για κάθε ταξινομητή ίσως να καταλήγουμε σε διαφορετικό προεπιλεγμένο σύνολο γονιδίων.

5 Παρουσίαση των Αποτελεσμάτων

Να υπενθυμίσουμε ότι για τα δεδομένα της λευχαιμίας έχουμε στη διάθεσή μας 7129 γονίδια, 38 δείγματα για εκπαίδευση και 34 δείγματα για δοκιμή. Η εκτίμηση της απόδοσης γίνεται με 9 ταξινομητές, τους NM, *k*-NN (για *k*=1, 5, 11, 37) και SVM (για *C*=0.001, 1, 100, 1000). Για τα δεδομένα του καρκίνου του μαστού διατίθενται 24188 γονίδια, 78 δείγματα για εκπαίδευση και 19 δείγματα για δοκιμή. Η εκτίμηση της απόδοσης γίνεται με 11 ταξινομητές, τους NM, *k*-NN (για *k*=1, 5, 11, 37, 51, 71) και SVM (για *C*=0.001, 1, 100, 1000).

Τα αποτελέσματα αναλύονται ξεχωριστά για κάθε περίπτωση καρκίνου η οποία εξετάζεται, καθώς και ξεχωριστά για κάθε μέθοδο που χρησιμοποιείται για την μείωση των γονιδίων (filter ή wrapper μέθοδος). Για συγκεκριμένη παθολογία και συγκεκριμένη μέθοδο επιλογής γονιδίων παρουσιάζονται πρώτα τα αποτελέσματα χωρίς την προεπεξεργασία (τυπικό σχήμα επιλογής/ταξινόμησης, Σχήμα 14), στη συνέχεια τα αποτελέσματα μετά την προεπεξεργασία (τελικό σχήμα επιλογής, Σχήμα 19), καθώς και μία σύγκριση μεταξύ των δύο. Αφού παρουσιάζονται τα αποτελέσματα για κάθε μέθοδο επιλογής, τελικά γίνεται μια σύγκριση μεταξύ της filter και της wrapper μεθοδολογίας για τη συγκεκριμένη παθολογία.

Αφού παρουσιαστούν αναλυτικά τα αποτελέσματα για κάθε παθολογία, συγκεντρώνονται οι αποδόσεις των ταξινομητών για την τιμή της παραμέτρου τους που φέρει το καλύτερο αποτέλεσμα και γίνεται μια τελευταία συγκεντρωτική σύγκριση ανάμεσα σε filter και wrapper μεθοδολογία για κάθε παθολογία. Τελικά γίνονται κάποιες παρατηρήσεις συγκριτικά για τις δύο παθολογίες.

5.1 Αποτελέσματα για τα δεδομένα της Λευχαιμίας

5.1.1 Απόδοση των ταξινομητών στα αρχικά γονίδια

Προκειμένου να γίνει μια εκτίμηση της βελτίωσης που επιτυγχάνεται αργότερα στην απόδοση των 9 ταξινομητών με επιλογή γονιδίων, παραθέτουμε αρχικά την απόδοση τους στο σύνολο των 7129 γονιδίων. Τα αποτελέσματα φαίνονται στον Πίνακα 1. Έτσι προκύπτει ένα πρώτο συμπέρασμα για την φύση των δεδομένων μας και πόσο καλά μπορούν οι ταξινομητές να διαχωρίσουν τις δύο κλάσεις.

Το ποσοστό επιτυχίας δεν είναι τόσο ικανοποιητικό αν εξαιρέσουμε τους Nearest Mean και 1-Nearest Neighbor.

Ταξινομητής	Αριθμός Γονιδίων	Train set			Test set		
		SR %	SE %	SP %	SR %	SE %	SP %
Nearest Mean	7129	97.37	90.91	100	82.35	57.14	100
1-Nearest Neighbor	7129	100	100	100	82.35	64.28	95
5-Nearest Neighbor	7129	92.1	72.73	100	73.53	42.86	95
11-Nearest Neighbor	7129	78.95	27.27	100	67.65	21.43	100
37-Nearest Neighbor	7129	71.05	0	100	58.82	0	100
SVM C=0.001	7129	100	100	100	79.41	50	100
SVM C=1	7129	100	100	100	79.41	50	100
SVM C=100	7129	100	100	100	79.41	50	100
SVM C=1000	7129	100	100	100	79.41	50	100

Πίνακας 1. Λευχαιμία: Απόδοση ταξινομητών στο αρχικό σύνολο γονιδίων

Όλοι οι ταξινομητές ταξινομούν με μεγάλη επιτυχία (95-100%) τα δείγματα της αρνητικής κλάσης (ALL) τόσο για το train όσο και για το test set. Αδυνατούν όμως να ταξινομήσουν με επιτυχία τα δείγματα της θετικής κλάσης (AML) κυρίως για το test set.

Όσον αφορά τις τιμές των παραμέτρων *k* και *C* έχουμε να παρατηρήσουμε τα εξής:

- Για τον *k*-Nearest Neighbor όσο αυξάνεται η τιμή της παραμέτρου *k*, τόσο μειώνεται η απόδοση.

- Για *k*>13 η απόδοση παραμένει σταθερή και ίδια με αυτή για *k*=37. Αυτό συμβαίνει διότι για πολύ μεγάλο *k* η συνάρτηση απόφασης επηρεάζεται από το μοντέλο ταξινόμησης [36]. Επειδή ο *k*-NN ταξινομεί ένα δείγμα στην πλειοψηφούσα κλάση ανάμεσα στους *k* γείτονές του, για μεγάλο *k* το δείγμα τείνει να ταξινομείται στην κλάση που έχει τα πιο πολλά δείγματα στο train set.

- Δεν υπάρχει διαφορά στην απόδοση του SVM ανάμεσα στις διαφορετικές τιμές της παραμέτρου *C*.

Σε γενικές γραμμές τα αποτελέσματα που έχουμε στο αρχικό σύνολο των γονιδίων είναι ικανοποιητικά, όχι βέβαια βέλτιστα. Στόχος μας είναι να πετύχουμε μεγαλύτερα ποσοστά επιτυχίας για μικρότερο set γονιδίων.

Ακολουθούν αρχικά τα αποτελέσματα που προκύπτουν με χρήση της filter μεθόδου για την μείωση των γονιδίων και στη συνέχεια αυτά που προκύπτουν με την wrapper μέθοδο.

5.1.2 Filter μέθοδος – Τυπική διαδικασία επιλογής/ταξινόμησης στο αρχικό σύνολο γονιδίων

Εφαρμόζεται το τυπικό σχήμα επιλογής/ταξινόμησης (Σχήμα 14), που περιγράφηκε στο προηγούμενο κεφάλαιο, για filter μεθοδολογία επιλογής γονιδίων. Σε κάθε στάδιο επιλογής γονιδίων γίνεται εκπαίδευση και εκτίμηση της απόδοσης των ταξινομητών.

Μετά την διαδικασία επιλογής, για κάθε έναν από τους ταξινομητές εντοπίζεται το ελάχιστο σύνολο γονιδίων για το οποίο επιτυγχάνεται η μέγιστη απόδοση. Τα αποτελέσματα φαίνονται στον Πίνακα 2.

Σε γενικές γραμμές το αποτέλεσμα είναι πολύ ικανοποιητικό για όλους τους ταξινομητές, πλην του 37–NN. Πλέον το SE στο test είναι πιο καλό από αυτό που είχαμε αρχικά.

Στο Train set το ποσοστό επιτυχίας φτάνει πλέον στο 100% σε όλες τις υπόλοιπες περιπτώσεις, ακόμα και σε αυτή του 11-Nearest Neighbor που αρχικά ήταν χαμηλό. Έχουμε βελτίωση του ποσοστού επιτυχίας στο test set για λίγα γονίδια.

Την καλύτερη απόδοση παρουσιάζει ο SVM για τις τιμές 1, 100 και 1000 της παραμέτρου *C*, παρά το γεγονός ότι αρχικά δεν είχε τόσο καλή απόδοση για αυτές τις τιμές.

Στην περίπτωση του *k*-Nearest Neighbor για *k*=37, για όλα τα σύνολα γονιδίων που προκύπτουν, το ποσοστό επιτυχίας είναι το ίδιο ακριβώς με εκείνο που είχαμε πριν την μείωση των γονιδίων.

Ταξινομητής	**Αριθμός Γονιδίων**	**Train set**			**Test set**		
		SR %	**SE %**	**SP %**	**SR %**	**SE %**	**SP %**
Nearest Mean	64	100	100	100	88.23	71.43	100
1-Nearest Neighbor	4	100	100	100	88.23	71.43	100
5-Nearest Neighbor	128	100	100	100	85.29	64.28	100
11-Nearest Neighbor	128	100	100	100	85.29	64.28	100
37-Nearest Neighbor	1	71.05	0	100	58.82	0	100
SVM C=0.001	512	100	100	100	82.35	57.14	100
SVM C=1	128	100	100	100	91.18	78.57	100
SVM C=100	128	100	100	100	91.18	78.57	100
SVM C=1000	128	100	100	100	91.18	78.57	100

Πίνακας 2. Λευχαιμία, filter μέθοδος: Τυπική διαδικασία επιλογής/ταξινόμησης στα αρχικά γονίδια – Μέγιστο ποσοστό επιτυχίας για κάθε ταξινομητή

Όσον αφορά τις τιμές των παραμέτρων *k* και *C* έχουμε να παρατηρήσουμε τα εξής:

- Όσο το *k* αυξάνεται, τόσο μειώνεται η απόδοση του ταξινομητή *k* κοντινότερων γειτόνων. Για *k*=1 έχουμε πολύ καλή απόδοση για μόλις 4 γονίδια.

- Ο SVM έχει λιγότερη απόδοση για *C*=0.001 απ' ότι για τις άλλες τρεις τιμές της παραμέτρου, όπου η συμπεριφορά του είναι όμοια και για τις τρεις.

5.1.3 Filter μέθοδος – Επιλογή γονιδίων με προεπεξεργασία δειγμάτων

5.1.3.1 Προεπεξεργασία – επιλογή αρχικού συνόλου γονιδίων

Γίνεται η προεπεξεργασία των δειγμάτων όπως αυτή περιγράφηκε στα Σχήματα 19 και 20. Εφαρμόζεται LOOCV στο train, γίνεται τυπική επιλογή γονιδίων με filter μέθοδο για κάθε ένα από τα 38 ζεύγη train και test και υπολογίζονται τα ποσοστά επιτυχίας σε κάθε βήμα επιλογής. Στην δεύτερη και τρίτη στήλη του Πίνακα 3 παρουσιάζεται για κάθε ταξινομητή η επανάληψη της επιλογής γονιδίων, για την οποία έχουμε μέγιστο μέσο ποσοστό επιτυχίας ανά τα 38 διαφορετικά τρεξίματα και το μέγεθος καθενός από τα 38 διαφορετικά σύνολα γονιδίων που προκύπτουν σε αυτό το βήμα αντίστοιχα. Στην τέταρτη στήλη φαίνεται το μέγιστο μέσο ποσοστό επιτυχίας και τελικά στην τελευταία στήλη φαίνεται το μέγεθος του συνόλου που προκύπτει από την ένωση των 38 διαφορετικών συνόλων γονιδίων. Τελικά, για κάθε ταξινομητή προκύπτει ένα σύνολο γονιδίων, υποσύνολο των 7129.

Ταξινομητής	Επανάληψη Επιλογής γονιδίων	Μέγεθος συνόλου γονιδίων	Μέγιστο μέσο ποσοστό επιτυχίας	Μέγεθος προεπιλεγμένου συνόλου
Nearest Mean	$11^{η}$	4	97.37	10
1-Nearest Neighbor	$10^{η}$	8	97.37	18
5-Nearest Neighbor	$11^{η}$	4	97.37	10
11-Nearest Neighbor	$11^{η}$	4	97.37	10
37-Nearest Neighbor	$13^{η}$	1	71.05	2
SVM C=0.001	$4^{η}$	512	94.74	981
SVM C=1	$10^{η}$	8	97.37	18
SVM C=100	$10^{η}$	8	97.37	18
SVM C=1000	$10^{η}$	8	97.37	18

Πίνακας 3. Λευχαιμία, filter μέθοδος: Μέγιστο μέσο ποσοστό επιτυχίας και μέγεθος προεπιλεγμένου συνόλου για κάθε ταξινομητή

Επιτυγχάνουμε πολύ καλό μέσο ποσοστό επιτυχίας σε όλες τις περιπτώσεις των ταξινομητών εκτός αυτής του 37-Nearest Neighbor, όπου έχουμε το ίδιο χαμηλό μέσο ποσοστό επιτυχίας σε όλα τα στάδια της επιλογής των γονιδίων.

Όλοι οι υπόλοιποι ταξινομητές πλην του SVM για *C*=0.001 καταλήγουν και σε μικρό προεπιλεγμένο σύνολο γονιδίων.

5.1.3.2 Τελική επιλογή γονιδίων με αφετηρία το προεπιλεγμένο σύνολο γονιδίων

Για κάθε έναν ταξινομητή καταλήγουμε σε ένα προεπιλεγμένο σύνολο γονιδίων. Κάνοντας επιλογή ξεκινώντας από αυτά τα γονίδια και χρησιμοποιώντας τα αρχικά train και independent test sets, όπως περιγράφηκε στο Σχήμα 18, καταλήγουμε στα αποτελέσματα του Πίνακα 4.

Ταξινομητής	Αριθμός Γονιδίων	Train set			Test set		
		SR %	SE %	SP %	SR %	SE %	SP %
Nearest Mean	6	100	100	100	88.23	71.43	100
1-Nearest Neighbor	3	100	100	100	88.23	71.43	100
5-Nearest Neighbor	5	100	100	100	85.29	64.28	100
11-Nearest Neighbor	6	97.37	90.91	100	85.29	64.28	100
37-Nearest Neighbor	1	71.05	0	100	58.82	0	100
SVM C=0.001	336	100	100	100	82.35	57.14	100
SVM C=1	5	100	100	100	88.23	71.43	100
SVM C=100	5	100	100	100	88.23	71.43	100
SVM C=1000	5	100	100	100	88.23	71.43	100

Πίνακας 4. Λευχαιμία, filter μέθοδος: Επιλογή γονιδίων με αφετηρία το προεπιλεγμένο σύνολο γονιδίων – Μέγιστο ποσοστό επιτυχίας για κάθε ταξινομητή

Σε σχέση με την επιλογή στο αρχικό σύνολο γονιδίων, παρατηρούμε ότι οι ταξινομητές διατηρούν την ίδια απόδοση για λιγότερα γονίδια. Μόνο ο SVM για τις τιμές 1, 100 και 1000 της παραμέτρου *C* παρουσιάζει ελάχιστη μείωση στην απόδοση, ενώ ο 11-NN παρουσιάζει ελάχιστα χαμηλότερη απόδοση στο train set.

Οι παρατηρήσεις για τις διάφορες τιμές των παραμέτρων *k* και *C* είναι οι ίδιες που διαπιστώσαμε και στην επιλογή στο αρχικό σύνολο των γονιδίων.

Κατά την διαδικασία αυτή καταλήγουμε για κάθε ταξινομητή σε διαφορετικό σύνολο γονιδίων για το οποίο ο συγκεκριμένος έχει τη βέλτιστη απόδοση. Μας ενδιαφέρει να δούμε εάν υπάρχουν κοινά γονίδια μεταξύ αυτών των διαφορετικών συνόλων γονιδίων.

Λόγω του ότι ο ταξινομητής Nearest-Neighbor για *k*=37 δε δίνει ικανοποιητικά αποτελέσματα, δεν τον λαμβάνουμε υπόψη στα συμπεράσματά μας όσον αφορά τα γονίδια στα οποία καταλήγουμε. Οπότε έχουμε να κάνουμε με οκτώ διαφορετικά σύνολα γονιδίων.

Ανάμεσα στα οκτώ σύνολα γονιδίων που προκύπτουν από τους διαφορετικούς ταξινομητές, η τομή τους είναι τα 3 γονίδια στα οποία καταλήγουμε με τον 1-Nearest Neighbor. Η ένωση των γονιδίων στα οποία καταλήγουμε με αυτή τη διαδικασία είναι τα 336 γονίδια στα οποία καταλήγουμε με τον SVM για C=0.001.

Κάθε σύνολο γονιδίων είναι υποσύνολο των μεγαλυτέρων του.

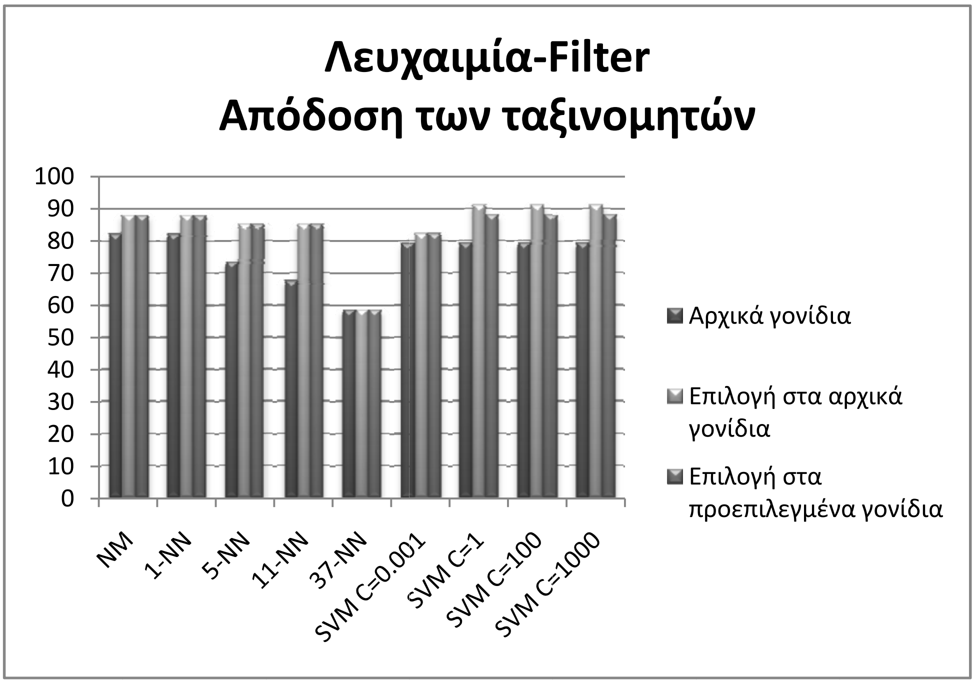
5.1.4 Filter μέθοδος – Σύγκριση επιλογής στα αρχικά γονίδια και επιλογής στο προεπιλεγμένο σύνολο γονιδίων

Συγκρίνοντας τα αποτελέσματα που έχουμε με επιλογή στα αρχικά γονίδια με αυτά που έχουμε με επιλογή στο προεπιλεγμένο σύνολο γονιδίων, μπορούμε να βγάλουμε κάποιο συμπέρασμα για το αν η προεπεξεργασία των δειγμάτων εκπαίδευσης αποδίδει στην περίπτωση της filter μεθόδου.

Κατά την επιλογή γονιδίων στο αρχικό σύνολο, έχουμε μείωση στα γονίδια (κατά μέσο όρο φτάνουμε στα 128) και αύξηση 11.61% κατά μέσο όρο στο ποσοστό επιτυχίας σε σχέση με αυτό που είχαμε στα αρχικά γονίδια. Και εδώ όσο αυξάνεται το *k* μειώνεται η απόδοση του *k*-NN και ο SVM για *C*=0.001 έχει χαμηλότερη απόδοση από αυτή για τις υπόλοιπες τρεις παραμέτρους. Όλοι οι ταξινομητές έχουν καλά ποσοστά επιτυχίας εκτός του *k*-NN για *k*=37.

Κατά την επιλογή γονιδίων στο προεπιλεγμένο σύνολο, έχουμε ακόμη μεγαλύτερη μείωση στα γονίδια (5 γονίδια κατά μέσο όρο) εκτός από τον SVM για *C*=0.001 όπου και πάλι τα γονίδια είναι πολλά. Η αύξηση στο ποσοστό επιτυχίας σε σχέση με αυτό στα αρχικά γονίδια είναι η ίδια με προηγουμένως, εκτός από τον SVM για *C*=1, 100 και 1000 όπου είναι λίγο μικρότερη, όπως φαίνεται και στο Σχήμα 21. Η συμπεριφορά του *k*-Nearest Neighbor και του SVM για τις διάφορες παραμέτρους τους είναι όμοια με αυτήν της πρώτης προσέγγισης.

Στο Σχήμα 21 φαίνεται συγκριτικά η απόδοση των ταξινομητών στα αρχικά γονίδια με την απόδοση που επιτυγχάνεται κατά την επιλογή στο αρχικό σύνολο γονιδίων και με αυτή που επιτυγχάνεται κατά την επιλογή στο προεπιλεγμένο σύνολο γονιδίων.



Σχήμα 21. Λευχαιμία: Απόδοση των ταξινομητών για την filter μέθοδο.

Παρατηρούμε ότι η προεπεξεργασία των δειγμάτων εκπαίδευσης δεν οδηγεί σε καλύτερη απόδοση, αλλά οδηγεί σε λιγότερα γονίδια.

Προκειμένου να έχουμε μία γενικότερη άποψη για το τι τελικά επιτυγχάνεται με το στάδιο της προεπεξεργασίας, μελετάμε τα σύνολα γονιδίων στα οποία καταλήγουμε με και χωρίς την προεπεξεργασία (Πίνακας 4 και Πίνακας 2 αντίστοιχα). Τα γονίδια στα οποία καταλήγουμε με επιλογή στο προεπιλεγμένο σύνολο είναι υποσύνολο αυτών που καταλήγουμε με επιλογή στο αρχικό σύνολο. Μάλιστα, στα ίδια γονίδια θα καταλήγαμε και με την επιλογή στο αρχικό σύνολο, εάν από ένα σημείο και μετά μειώναμε τα γονίδια κατά ένα κάθε φορά.

Συμπεραίνουμε λοιπόν ότι στην περίπτωση της filter μεθόδου η προεπεξεργασία των δειγμάτων εκπαίδευσης προκειμένου να βρεθεί ένα καλύτερο σύνολο γονιδίων από το οποίο να αρχίσει η επιλογή γονιδίων, δεν έχει κάποιο σημαντικό πλεονέκτημα. Είναι πιο αποδοτικό και λιγότερο χρονοβόρο να ξεκινήσει η επιλογή γονιδίων στο αρχικό σύνολο και από ένα σημείο και μετά να μειώνουμε κατά ένα τα γονίδια.

Προχωρώντας ένα βήμα παραπέρα, μία σκέψη θα ήταν να συγκεντρώσουμε τα σύνολα των γονιδίων που προκύπτουν για κάθε ταξινομητή μετά την επιλογή στο προεπιλεγμένο σύνολο γονιδίων (Πίνακας 4) χωρίς να λάβουμε υπόψη τα γονίδια που προκύπτουν από τον Nearest Neighbor για k = 37 (λόγω του ότι η απόδοσή του είναι εξαιρετικά χαμηλή) καθώς και τα γονίδια που προκύπτουν από τον SVM για C = 0.001 (λόγω του ότι είναι πάρα πολλά). Θα μπορούσε να υπολογιστεί η ένωση των υπόλοιπων συνόλων γονιδίων η οποία αποτελείται από 6 γονίδια και να γίνει επιλογή γονιδίων με βάση το τυπικό σχήμα επιλογής/ταξινόμησης όπως αυτό περιγράφεται στο Σχήμα 14. Κάνοντας κάτι τέτοιο, τα 6 γονίδια που προκύπτουν με την ένωση είναι υποσύνολο καθενός από τα προεπιλεγμένα σύνολα του Πίνακα 3 από τα οποία ξεκίνησε η τελική επιλογή. Μάλιστα εάν τα ταξινομήσουμε με βάση το Fisher's Criterion παρατηρούμε ότι διατηρούν την ίδια σειρά με πριν. Οπότε κάνοντας επιλογή σε αυτό το σύνολο των 6 γονιδίων, καταλήγουμε πάλι στα αποτελέσματα του Πίνακα 4.

5.1.5 Wrapper μέθοδος – Τυπική διαδικασία επιλογής/ταξινόμησης στο αρχικό σύνολο γονιδίων

Εφαρμόζεται το τυπικό σχήμα επιλογής/ταξινόμησης (Σχήμα 14) που περιγράφηκε στο προηγούμενο κεφάλαιο για wrapper μεθοδολογία επιλογής γονιδίων. Σε κάθε στάδιο επιλογής γονιδίων γίνεται εκπαίδευση και εκτίμηση της απόδοσης των ταξινομητών.

Μετά την διαδικασία επιλογής, για κάθε έναν από τους ταξινομητές εντοπίζεται το ελάχιστο σύνολο γονιδίων για το οποίο επιτυγχάνεται η μέγιστη απόδοση. Τα αποτελέσματα φαίνονται στον Πίνακα 5.

Ταξινομητής	Αριθμός Γονιδίων	Train set			Test set		
		SR %	SE %	SP %	SR %	SE %	SP %
Nearest Mean	4	100	100	100	85.29	64.28	100
1-Nearest Neighbor	4	100	100	100	82.35	57.14	100
5-Nearest Neighbor	8	100	100	100	79.41	50	100
11-Nearest Neighbor	8	100	100	100	73.53	35.71	100
37-Nearest Neighbor	1	71.05	0	100	58.82	0	100
SVM C=0.001	512	100	100	100	82.35	57.14	100
SVM C=1	4	100	100	100	82.35	57.14	100
SVM C=100	4	100	100	100	82.35	57.14	100
SVM C=1000	4	100	100	100	82.35	57.14	100

Πίνακας 5. Λευχαιμία, wrapper μέθοδος: Τυπική διαδικασία επιλογής/ταξινόμησης στα αρχικά γονίδια– Μέγιστο ποσοστό επιτυχίας για κάθε ταξινομητή

Στο Train set το ποσοστό επιτυχίας φτάνει το 100% σε όλες τις περιπτώσεις (εκτός του 37-NN), ακόμα και σε αυτή του 11-Nearest Neighbor που αρχικά ήταν χαμηλό. Έχουμε μικρή βελτίωση του ποσοστού επιτυχίας στο test set για πολύ λιγότερα γονίδια.

Παρόλα αυτά παρατηρούμε ότι ο Nearest Neighbor για k=5 και k=11 συνεχίζει να μην έχει τόσο ικανοποιητικό ποσοστό επιτυχίας, ενώ για k=37 συνεχίζει να έχει πολύ χαμηλή απόδοση.

Επίσης ο SVM για C=0.001 έχει μεν καλή απόδοση αλλά την πετυχαίνει για περισσότερα γονίδια από τους υπόλοιπους ταξινομητές. Μάλιστα, στα 4 γονίδια όπου οι περισσότεροι ταξινομητές έχουν την μεγαλύτερη απόδοση, το ποσοστό επιτυχίας του είναι μόλις 58.82% στο test και 71.05% στο train.

Παρόλα αυτά μπορούμε να πούμε ότι τα 4 αυτά γονίδια είναι γονίδια υψηλής σημασίας αφού για αυτά στις περισσότερες των περιπτώσεων των classifiers έχουμε μεγάλο ποσοστό επιτυχίας.

Επίσης να σημειώσουμε ότι οι ταξινομητές αδυνατούν να ταξινομήσουν με μεγάλη επιτυχία τα δείγματα που ανήκουν στην θετική κλάση, ενώ ταξινομούν χωρίς σφάλμα τα δείγματα της αρνητικής κλάσης.

Όσον αφορά τις τιμές των παραμέτρων *k* και *C* μπορούμε να παρατηρήσουμε τα εξής:

- Όσο αυξάνεται το *k*, μειώνεται και το ποσοστό επιτυχίας του *k*-NN, κάτι το οποίο διαπιστώσαμε και με τα αρχικά γονίδια.
- Ο SVM έχει καλύτερη απόδοση για τις τιμές 1,100 και 1000 της παραμέτρου *C*, όπου η συμπεριφορά του είναι ίδια.

5.1.6 Wrapper μέθοδος – Επιλογή γονιδίων με προεπεξεργασία δειγμάτων

5.1.6.1 Προεπεξεργασία – επιλογή αρχικού συνόλου γονιδίων

Γίνεται η προεπεξεργασία των δειγμάτων όπως αυτή περιγράφηκε στα Σχήματα 19 και 20. Εφαρμόζεται LOOCV στο train, γίνεται τυπική επιλογή γονιδίων με wrapper μέθοδο για κάθε ένα από τα 38 ζεύγη train και test και υπολογίζονται τα ποσοστά επιτυχίας σε κάθε βήμα επιλογής. Στην δεύτερη και τρίτη στήλη του Πίνακα 6 παρουσιάζεται για κάθε ταξινομητή η επανάληψη της επιλογής γονιδίων, για την οποία έχουμε μέγιστο μέσο ποσοστό επιτυχίας ανά τα 38 διαφορετικά τρεξίματα και το μέγεθος καθενός από τα 38 διαφορετικά σύνολα γονιδίων που προκύπτουν σε αυτό το βήμα αντίστοιχα. Στην τέταρτη στήλη φαίνεται το μέγιστο μέσο ποσοστό επιτυχίας και τελικά στην τελευταία στήλη φαίνεται το μέγεθος του συνόλου, που προκύπτει από την ένωση των 38 διαφορετικών συνόλων γονιδίων. Τελικά, για κάθε ταξινομητή προκύπτει ένα σύνολο γονιδίων, υποσύνολο των 7129.

Οι τιμές των μέσων ποσοστών επιτυχίας είναι πολύ ικανοποιητικές για όλους τους ταξινομητές, εκτός του Nearest Neighbor για *k*=37, κάτι το οποίο ήταν αναμενόμενο από τα προηγούμενα αποτελέσματα.

Ειδικά ο Nearest Mean δίνει πολύ καλό ποσοστό επιτυχίας και τελικά καταλήγουμε και σε μικρό σχετικά σύνολο γονιδίων (517). Το ίδιο συμβαίνει και για τον SVM για *C*=1, 100, 1000 (όπου και πάλι η συμπεριφορά του ταξινομητή για τις 3 αυτές τιμές της παραμέτρου *C* είναι πανομοιότυπη). Για *C*=0.001 έχουμε καλή απόδοση όμως για πολλά γονίδια.

Για τον Nearest Neighbor όσο αυξάνεται η τιμή του *k* μειώνεται και η απόδοση. Παρόλα αυτά, όσο μεγαλύτερο είναι το *k* η μέγιστη απόδοση παρατηρείται για λιγότερα γονίδια.

Ταξινομητής	Επανάληψη Επιλογής γονιδίων	Μέγεθος συνόλου γονιδίων	Μέγιστο μέσο ποσοστό επιτυχίας (%)	Μέγεθος προεπιλεγμένου συνόλου
Nearest Mean	6^{η}	128	97.37	517
1-Nearest Neighbor	1^{η}	4096	92.11	6017
5-Nearest Neighbor	3^{η}	1024	86.84	1890
11-Nearest Neighbor	13^{η}	1	86.84	15
37-Nearest Neighbor	13^{η}	1	71.05	15
SVM C=0.001	3^{η}	1024	94.74	1890
SVM C=1	8^{η}	32	94.74	239
SVM C=100	8^{η}	32	94.74	239
SVM C=1000	8^{η}	32	94.74	239

Πίνακας 6. Λευχαιμία, wrapper μέθοδος: Μέγιστο μέσο ποσοστό επιτυχίας και μέγεθος προεπιλεγμένου συνόλου για κάθε ταξινομητή

5.1.6.2 Τελική επιλογή γονιδίων με αφετηρία το προεπιλεγμένο σύνολο γονιδίων

Για κάθε έναν ταξινομητή καταλήγουμε σε ένα προεπιλεγμένο σύνολο γονιδίων. Κάνοντας επιλογή ξεκινώντας από αυτά τα γονίδια και χρησιμοποιώντας τα αρχικά train και independent test sets όπως περιγράφηκε στο Σχήμα 18, καταλήγουμε στα αποτελέσματα του Πίνακα 7.

Ταξινομητής	**Αριθμός Γονιδίων**	**Train set**			**Test set**		
		SR %	**SE %**	**SP %**	**SR %**	**SE %**	**SP %**
Nearest Mean	2	97.37	90.91	100	94.12	85.71	100
1-Nearest Neighbor	1	100	100	100	85.29	71.43	95
5-Nearest Neighbor	1	97.37	90.91	100	91.18	92.86	90
11-Nearest Neighbor	4	89.47	63.64	100	85.29	64.28	100
37-Nearest Neighbor	1	71.05	0	100	58.82	0	100
SVM C=0.001	1890	100	100	100	82.35	57.14	100
SVM C=1	4	100	100	100	97.06	92.86	100
SVM C=100	4	100	100	100	97.06	92.86	100
SVM C=1000	4	100	100	100	97.06	92.86	100

Πίνακας 7. Λευχαιμία, wrapper μέθοδος: Επιλογή γονιδίων με αφετηρία το προεπιλεγμένο σύνολο γονιδίων – Μέγιστο ποσοστό επιτυχίας για κάθε ταξινομητή

Παρατηρούμε ότι με την διαδικασία αυτή καταλήγουμε σε καλύτερα αποτελέσματα από την τυπική διαδικασία επιλογής στα αρχικά γονίδια, κυρίως όσον αφορά την απόδοση των ταξινομητών και όχι τα γονίδια, τα οποία είναι επίσης λίγα σε αριθμό. Μάλιστα πλέον έχουμε πολύ ικανοποιητική απόδοση για όλους τους ταξινομητές εκτός του 37-Nearest Neighbor. Εξαίρεση επίσης αποτελεί και ο SVM για *C*=0.001 όχι όμως ως προς την απόδοση αλλά ως προς το πλήθος των γονιδίων για τα οποία την πετυχαίνει.

Να παρατηρήσουμε ότι πλέον η τιμή του sensitivity είναι πολύ ικανοποιητική για την πλειοψηφία των ταξινομητών.

Όσον αφορά τις τιμές των παραμέτρων *k* και *C* μπορούμε να παρατηρήσουμε τα εξής:

- Πλέον την μέγιστη απόδοση ο *k*-NN την έχει για *k*=5, ακολουθεί για *k*=1 και 11 ενώ για *k*=37 η απόδοση παραμένει σταθερά χαμηλή.
- Ο SVM έχει καλύτερη απόδοση για τις τιμές 1,100 και 1000 της παραμέτρου *C*, όπου η συμπεριφορά του είναι ίδια.

Κατά την διαδικασία αυτή καταλήγουμε για κάθε ταξινομητή σε διαφορετικό σύνολο γονιδίων για το οποίο ο συγκεκριμένος έχει τη βέλτιστη απόδοση. Όπως έγινε και με την filter μέθοδο έτσι και εδώ μας ενδιαφέρει να δούμε εάν υπάρχουν κοινά γονίδια μεταξύ αυτών των διαφορετικών συνόλων γονιδίων.

Λόγω του ότι ο ταξινομητής Nearest-Neighbor για *k*=37 δε δίνει ικανοποιητικά αποτελέσματα, δεν τον λαμβάνουμε υπόψη στα συμπεράσματά μας όσον αφορά τα γονίδια στα οποία καταλήγουμε. Οπότε έχουμε να κάνουμε με οκτώ διαφορετικά σύνολα γονιδίων.

Δεν υπάρχει τομή ανάμεσα στα διαφορετικά σύνολα γονιδίων που καταλήγουμε με τους διάφορους ταξινομητές, δεν υπάρχει δηλαδή τουλάχιστον ένα γονίδιο κοινό ανάμεσά τους. Υπάρχουν όμως γονίδια με σχετικά υψηλή συχνότητα (δύο γονίδια εμφανίζονται σε πέντε των οκτώ περιπτώσεων).

Η ένωση των συνόλων είναι τα 1890 γονίδια στα οποία καταλήγουμε με τον SVM για *C*=0.001. Εννέα γονίδια εμφανίζονται σε περισσότερες από δύο περιπτώσεις. Τα υπόλοιπα 1881 απαντώνται μόνο στο σύνολο στο οποίο καταλήγουμε με τον SVM για *C*=0.001.

5.1.7 Wrapper μέθοδος – Σύγκριση επιλογής στα αρχικά γονίδια και επιλογής στο προεπιλεγμένο σύνολο γονιδίων

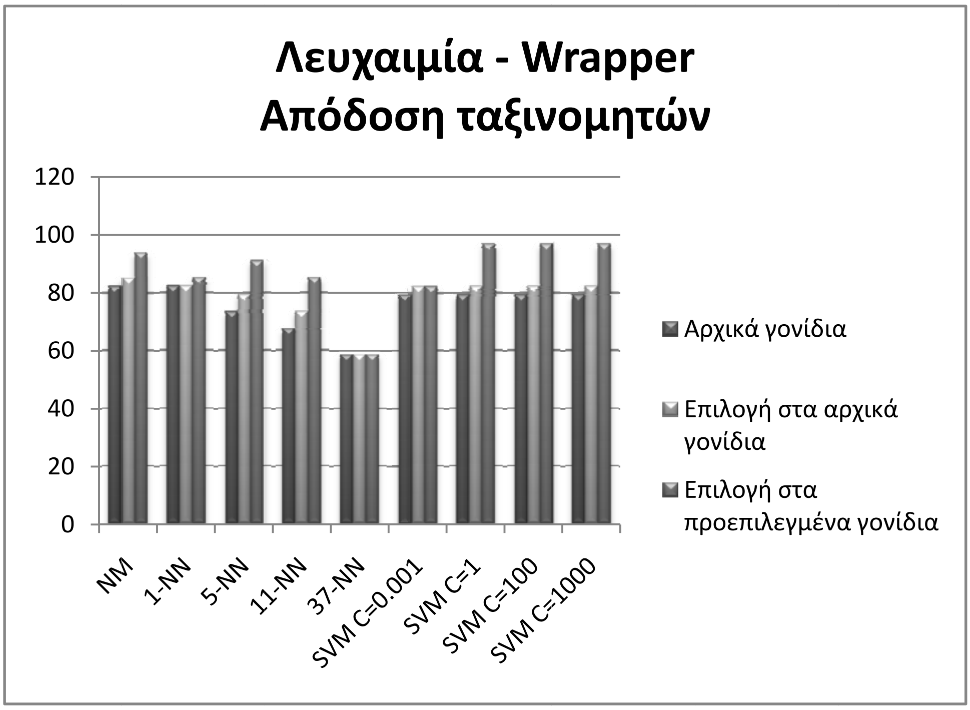
Συγκρίνοντας τα αποτελέσματα που έχουμε με επιλογή στα αρχικά γονίδια με αυτά που έχουμε με επιλογή στο προεπιλεγμένο σύνολο γονιδίων, μπορούμε να βγάλουμε κάποιο συμπέρασμα για το αν η προεπεξεργασία των δειγμάτων εκπαίδευσης αποδίδει στην περίπτωση της wrapper μεθόδου.

Κατά την τυπική διαδικασία επιλογής/ταξινόμησης πετυχαίνουμε μεγάλη μείωση στα γονίδια εκτός από την περίπτωση του SVM για *C*=0.001 και αύξηση στο ποσοστό επιτυχίας σε σχέση με τα αρχικά γονίδια κατά 3.9% στο test set κατά μέσο όρο. Για τον *k*-Nearest Neighbor όσο αυξάνεται το *k* μειώνεται η απόδοση και για πολύ μεγάλο *k* δεν είναι καθόλου ικανοποιητική. Η απόδοση του SVM για *C*=1, 100 και 1000 είναι καλύτερη από αυτή για *C*=0.001 και γενικά η συμπεριφορά του SVM για τις τρεις πρώτες τιμές της παραμέτρου *C* είναι όμοια. Όλοι οι ταξινομητές έχουν καλή απόδοση πλην του *k*-Nearest Neighbor για *k*=11 και 37. Παρατηρείται χαμηλό sensitivity στο Test set.

Κατά την επιλογή γονιδίων στο προεπιλεγμένο σύνολο γονιδίων έχουμε ακόμα μεγαλύτερη μείωση στα γονίδια εκτός και πάλι από τον SVM για *C*=0.001. Η αύξηση στο ποσοστό επιτυχίας είναι κατά μέσο όρο 15.36% σε σχέση με αυτό που είχαμε στα αρχικά γονίδια και πλέον όλοι οι ταξινομητές, πλην του *k*-Nearest Neighbor για *k*=37, έχουν καλή απόδοση. Για τον *k*-Nearest Neighbor την μέγιστη απόδοση την έχουμε πλέον για *k*=5, ενώ μετά όσο το *k* αυξάνεται η απόδοση μειώνεται. Η συμπεριφορά του SVM είναι ίδια με την πρώτη προσέγγιση με μεγαλύτερο όμως ποσοστό επιτυχίας. Το sensitivity στο test είναι πλέον πολύ καλό.

Στο Σχήμα 22 φαίνεται συγκριτικά η απόδοση των ταξινομητών στα αρχικά γονίδια με την απόδοση που επιτυγχάνεται κατά την επιλογή στο αρχικό σύνολο γονιδίων και με αυτή που επιτυγχάνεται κατά την επιλογή στο προεπιλεγμένο σύνολο γονιδίων.

Παρατηρούμε ότι με την προεπεξεργασία των δειγμάτων εκπαίδευσης καταλήγουμε σε υψηλότερα ποσοστά επιτυχίας από την επιλογή στα αρχικά γονίδια.



Σχήμα 22. Λευχαιμία: Απόδοση των ταξινομητών για την wrapper μέθοδο.

Προκειμένου να έχουμε μία γενικότερη άποψη για το τι τελικά επιτυγχάνεται με το στάδιο της προεπεξεργασίας, μελετάμε τα σύνολα γονιδίων στα οποία καταλήγουμε με και χωρίς την προεπεξεργασία (Πίνακας 7 και Πίνακας 5 αντίστοιχα). Ήδη το γεγονός ότι η απόδοση είναι μεγαλύτερη με επιλογή στο προεπιλεγμένο σύνολο γονιδίων για σχεδόν ίδιο αριθμό γονιδίων προϋποθέτει ότι δεν καταλήγουμε στα ίδια γονίδια στις δύο αυτές προσεγγίσεις. Πράγματι, σε όλες τις περιπτώσεις πρόκειται για διαφορετικά γονίδια εκτός από την περίπτωση του Nearest Neighbor για k=5, όπου έχουμε μόνο ένα κοινό γονίδιο. Επίσης στην περίπτωση του SVM για C=0.001 τα 1890 γονίδια, στα οποία καταλήγουμε με επιλογή στο προεπιλεγμένο σύνολο γονιδίων, είναι ένα υπερσύνολο των 512 που καταλήγουμε με επιλογή στα αρχικά γονίδια. Γι’ αυτό πιθανά να μην παρατηρείται καμία βελτίωση στο ποσοστό επιτυχίας για τον συγκεκριμένο ταξινομητή.

Επομένως το στάδιο της προεπεξεργασίας ευνοεί την απόδοση των ταξινομητών στην περίπτωση της wrapper μεθοδολογίας για επιλογή γονιδίων. Προχωρώντας ένα βήμα παραπέρα, συγκεντρώνουμε τα σύνολα των γονιδίων που προκύπτουν για κάθε ταξινομητή μετά την επιλογή στο προεπιλεγμένο σύνολο γονιδίων (Πίνακας 7).

Δεν λαμβάνονται υπόψη τα γονίδια που προκύπτουν από τον Nearest Neighbor για *k* = 37 (λόγω του ότι η απόδοσή του είναι εξαιρετικά χαμηλή) καθώς και τα γονίδια που προκύπτουν από τον SVM για *C* = 0.001 (λόγω του ότι είναι πάρα πολλά).

Υπολογίζεται η ένωση των υπόλοιπων συνόλων γονιδίων η οποία αποτελείται από 9 γονίδια. Τελικά γίνεται επιλογή γονιδίων σε αυτό το σύνολο σύμφωνα με την τυπική επιλογή γονιδίων του Σχήματος 14 για wrapper μεθοδολογία, μειώνοντας κατά ένα γονίδιο τη φορά και ελέγχοντας την απόδοση με όλους τους ταξινομητές. Τα αποτελέσματα που προκύπτουν φαίνονται στον Πίνακα 8.

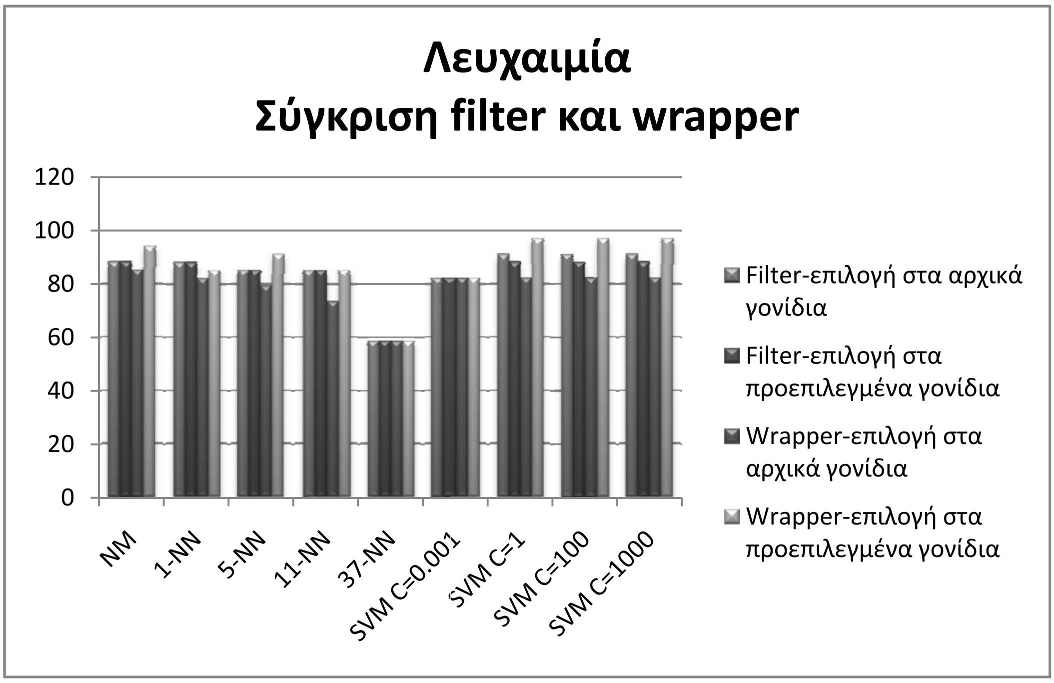
Ταξινομητής	Αριθμός Γονιδίων	Train set			Test set		
		SR %	SE %	SP %	SR %	SE %	SP %
Nearest Mean	3	100	100	100	94.12	85.71	100
1-Nearest Neighbor	4	100	100	100	94.12	85.71	100
5-Nearest Neighbor	1	97.37	90.91	100	91.18	92.86	90
11-Nearest Neighbor	2	92.10	72.73	100	85.29	64.28	100
37-Nearest Neighbor	1	71.05	0	100	58.82	0	100
SVM C=0.001	1	71.05	0	100	58.82	0	100
SVM C=1	4	100	100	100	97.06	92.86	100
SVM C=100	2	100	100	100	94.12	85.71	100
SVM C=1000	2	100	100	100	94.12	85.71	100

Πίνακας 8. Λευχαιμία, wrapper μέθοδος: Επιλογή γονιδίων με αφετηρία την ένωση συνόλων γονιδίων από τους διαφορετικούς ταξινομητές

Παρατηρούμε ότι πλέον όλοι οι ταξινομητές επιτυγχάνουν απόδοση μεγαλύτερη από 90% εκτός του 11-Nearest Neighbor που όμως έχει επίσης καλή απόδοση και των 37-NN και SVM για *C*=0.001 όπου η απόδοση είναι χαμηλή. Όσο μεγαλώνει η τιμή της παραμέτρου *k* μειώνεται η απόδοση του *k*-Nearest Neighbor όμως η μέγιστη απόδοσή του παρατηρείται για λιγότερα γονίδια. Το ίδιο συμβαίνει και για την τιμή της παραμέτρου *C* του SVM.

5.1.8 Σύγκριση filter και wrapper μεθόδων επιλογής

Στο Σχήμα 23 φαίνεται συνολικά η απόδοση των ταξινομητών για τα δεδομένα της λευχαιμίας, τόσο με επιλογή (filter και wrapper) στα αρχικά γονίδια, όσο και με επιλογή (filter και wrapper) στο προεπιλεγμένο σύνολο γονιδίων.



Σχήμα 23. Σύγκριση απόδοσης ταξινομητών με wrapper και filter για τα δεδομένα της Λευχαιμίας

Όσον αφορά την επιλογή στα αρχικά γονίδια, η filter μέθοδος δίνει καλύτερα αποτελέσματα από την wrapper. Παρατηρούμε ότι έχουμε πολύ πιο ικανοποιητικά αποτελέσματα κυρίως όσον αφορά το ποσοστό επιτυχίας. Πιο συγκεκριμένα δεν συναντάμε το τόσο χαμηλό SE που έχουμε στην wrapper μέθοδο (πλέον το SE κυμαίνεται από 65%-80% ενώ το SP συνεχίζει να είναι στο 100%). Βέβαια αυτό το τόσο ικανοποιητικό SR το πετυχαίνουμε για περισσότερα γονίδια από ότι στην wrapper μέθοδο, όμως σε καμία περίπτωση δε μπορούμε να χαρακτηρίσουμε χειρότερο το αποτέλεσμα αφού για λίγα γονίδια έχουμε τόσο μεγάλο ποσοστό επιτυχίας. Μπορούμε να πούμε ότι με την wrapper μέθοδο καταλήγουμε σε όσο το δυνατόν λιγότερα γονίδια αλλά με την filter σε όσο το δυνατόν μεγαλύτερο ποσοστό επιτυχίας.

Όσον αφορά την επιλογή στα προεπιλεγμένα γονίδια, η wrapper μέθοδος δίνει καλύτερα αποτελέσματα από την filter, τόσο ως προς το ποσοστό επιτυχίας όσο και ως προς τα γονίδια στα οποία αυτό επιτυγχάνεται.

Γενικά μπορούμε να παρατηρήσουμε ότι η προεπεξεργασία ωφελεί την wrapper μέθοδο και όχι την filter. Μάλιστα ενώ με την τυπική επιλογή γονιδίων η filter μέθοδος υπερτερεί της wrapper, με την προεπεξεργασία η wrapper παίρνει το προβάδισμα.

Για την τιμή 0.001 της παραμέτρου *C*, με οποιαδήποτε μέθοδο, παρατηρείται μικρότερη απόδοση και για περισσότερα γονίδια από ότι για *C*=1, 100 και 1000 – όπου έχουμε πάντα την ίδια απόδοση.

Σε όλες τις περιπτώσεις όσο το *k* αυξάνεται η απόδοση μειώνεται. Εξαίρεση αποτελεί η επιλογή στο προεπιλεγμένο σύνολο γονιδίων με wrapper μέθοδο όπου για *k*=5 παρατηρείται η καλύτερη απόδοση. Για πολύ μεγάλο *k* η απόδοση είναι χαμηλή και αυτό οφείλεται στο γεγονός ότι όσο μεγαλύτερο το *k* τόσο η συνάρτηση απόφασης επηρεάζεται από το μοντέλο ταξινόμησης, δηλαδή από τον τρόπο με τον οποίο είναι ταξινομημένα τα δείγματα εκπαίδευσης.

Επίσης να παρατηρήσουμε ότι με τις δύο διαφορετικές μεθόδους (wrapper και filter) καταλήγουμε σε δεκαέξι σύνολα γονιδίων για τα οποία δεν υπάρχει έστω ένα κοινό γονίδιο. Ένα γονίδιο απαντάται σε δέκα των δεκαέξι περιπτώσεων και δύο γονίδια απαντώνται σε εννέα των δεκαέξι περιπτώσεων. Ακολουθούν γονίδια με μικρότερη συχνότητα.

5.2 Αποτελέσματα για τα δεδομένα του καρκίνου του μαστού

5.2.1 Απόδοση των ταξινομητών στα αρχικά Train και Independent Test sets

Προκειμένου να γίνει μια εκτίμηση της βελτίωσης που επιτυγχάνεται αργότερα στην απόδοση των 11 ταξινομητών με επιλογή γονιδίων, παραθέτουμε αρχικά την απόδοση τους στο σύνολο των 24188 γονιδίων. Τα αποτελέσματα φαίνονται στον Πίνακα 9. Έτσι προκύπτει ένα πρώτο συμπέρασμα για την φύση των δεδομένων μας και πόσο καλά μπορούν οι ταξινομητές να τα διαχωρίσουν.

Ταξινομητής	Αριθμός Γονιδίων	Train set			Test set		
		SR %	SE %	SP %	SR %	SE %	SP %
Nearest Mean	24188	82.05	76.47	86.36	84.21	83.33	85.71
1-Nearest Neighbor	24188	100	100	100	63.16	58.33	71.43
5-Nearest Neighbor	24188	73.07	47.06	93.18	47.37	33.33	71.43
11-Nearest Neighbor	24188	67.95	35.29	93.18	47.37	25	85.71
37-Nearest Neighbor	24188	66.67	79.41	56.82	73.68	91.67	42.86
51-Nearest Neighbor	24188	56.41	91.17	29.54	63.16	100	0
71-Nearest Neighbor	24188	56.41	0	100	36.84	0	100
SVM C=0.001	24188	96.15	94.12	97.73	78.95	83.33	71.43
SVM C=1	24188	100	100	100	68.42	91.67	28.57
SVM C=100	24188	100	100	100	68.42	91.67	28.57
SVM C=1000	24188	100	100	100	68.42	91.67	28.57

Πίνακας 9. Καρκίνος του μαστού: Απόδοση ταξινομητών στο αρχικό σύνολο γονιδίων

Διαπιστώνουμε ότι ως επί των πλείστων οι ταξινομητές αδυνατούν να ταξινομήσουν επιτυχώς δεδομένα άγνωστα σε αυτούς. Παρατηρούμε ότι μόνο ο Nearest Mean μπορεί να ταξινομήσει με αρκετά καλή επιτυχία τα δείγματα του test set και ακολουθεί ο SVM για C=0.001 με όχι τόσο ικανοποιητική απόδοση.

Στις περιπτώσεις του 1-Nearest Neighbor και του SVM για C=1, 100, 1000 παρατηρούμε ότι το train ταξινομείται με 100% επιτυχία αλλά το test με επιτυχία μικρότερη του 70%. Έχουμε δηλαδή στις περιπτώσεις αυτές το φαινόμενο του over-fitting, δηλαδή το μοντέλο μας αδυνατεί να γενικευτεί σε άγνωστα σε αυτό δεδομένα και περιορίζεται μόνο σε αυτά που χρησιμοποιήθηκαν για να εκπαιδευτεί.

Σε όλες τις υπόλοιπες περιπτώσεις (Nearest Neighbor για k=5, 11, 37, 51 και 71) τόσο το train όσο και το test ταξινομούνται με μικρό ποσοστό επιτυχίας (μικρότερο του 75%). Μάλιστα για k>57, το ποσοστό επιτυχίας είναι το ίδιο μικρό ποσοστό που επιτυγχάνουμε για *k*=71. Αυτό όπως εξηγήσαμε και στα δεδομένα της λευχαιμίας συμβαίνει διότι για πολύ μεγάλο *k* η συνάρτηση απόφασης επηρεάζεται από το μοντέλο ταξινόμησης.

Όσον αφορά τις τιμές των παραμέτρων *k* και *C* έχουμε να παρατηρήσουμε τα εξής:

- Ο *k*-NN παρουσιάζει καλύτερη απόδοση για ενδιάμεσες τιμές του *k*, δηλαδή ούτε για πολύ μικρές, ούτε για πολύ μεγάλες.
- Για την τιμή 0.001 της παραμέτρου *C* έχουμε καλύτερη απόδοση του SVM απ' ότι για τις τιμές 1, 100 και 1000 όπου η απόδοση είναι η ίδια.

Με μια πρώτη ματιά λοιπόν, τα αποτελέσματα δεν είναι καθόλου ικανοποιητικά. Αυτό είναι και ένα συμπέρασμα που προέκυψε και στο προηγούμενο κεφάλαιο (Σχήμα 14), όπου είδαμε ότι τα δεδομένα του καρκίνου του μαστού είναι δύσκολο να διαχωριστούν. Στόχος μας είναι να βρεθεί σαφώς ένα μικρότερο σύνολο γονιδίων για το οποίο να έχουμε – όσο αυτό είναι εφικτό – καλύτερη απόδοση.

Ακολουθούν αρχικά τα αποτελέσματα που προκύπτουν με χρήση της filter μεθόδου για την μείωση των γονιδίων και στη συνέχεια αυτά που προκύπτουν με την wrapper μέθοδο.

5.2.2 Filter Μέθοδος – Τυπική διαδικασία επιλογής/ταξινόμησης στο αρχικό σύνολο γονιδίων

Εφαρμόζεται το τυπικό σχήμα επιλογής/ταξινόμησης (Σχήμα 14) που περιγράφηκε στο προηγούμενο κεφάλαιο για filter μεθοδολογία επιλογής γονιδίων. Σε κάθε στάδιο επιλογής γονιδίων γίνεται εκπαίδευση και εκτίμηση της απόδοσης των ταξινομητών. Μετά την διαδικασία επιλογής, για κάθε έναν από τους ταξινομητές εντοπίζεται το ελάχιστο σύνολο γονιδίων για το οποίο επιτυγχάνεται η μέγιστη απόδοση. Τα αποτελέσματα φαίνονται στον Πίνακα 10.

Ταξινομητής	Αριθμός Γονιδίων	Train set			Test set		
		SR %	SE %	SP %	SR %	SE %	SP %
Nearest Mean	1024	82.05	82.35	81.82	89.47	91.67	85.71
1-Nearest Neighbor	512	100	100	100	78.95	75	85.71
5-Nearest Neighbor	8	92.31	91.18	93.18	78.95	83.33	71.43
11-Nearest Neighbor	512	84.61	94.12	77.27	89.47	100	71.43
37-Nearest Neighbor	512	80.77	85.29	77.27	84.21	91.67	71.43
51-Nearest Neighbor	4096	76.92	67.65	84.09	84.21	83.33	85.71
71-Nearest Neighbor	1	56.41	0	100	36.84	0	100
SVM C=0.001	24188	96.15	94.12	97.73	78.95	83.33	71.43
SVM C=1	512	100	100	100	89.47	91.67	85.71
SVM C=100	512	100	100	100	89.47	91.67	85.71
SVM C=1000	512	100	100	100	89.47	91.67	85.71

Πίνακας 10 . Καρκίνος του μαστού, filter μέθοδος: Τυπική διαδικασία επιλογής/ταξινόμησης στα αρχικά γονίδια – Μέγιστο ποσοστό επιτυχίας για κάθε ταξινομητή

Γενικά το πλήθος των γονιδίων στα οποία καταλήγουμε δεν είναι μεγάλο, δεν είναι όμως και ικανοποιητικό. Εξαίρεση αποτελεί ο Nearest Neighbor για *k*=5 που έχει πλέον ικανοποιητική απόδοση στο test set (κοντά στο 80%) για λίγα γονίδια.

Επίσης για *k*=1 η απόδοση είναι ικανοποιητική (κοντά στο 80%). Ο Nearest Neighbor για *k*=11, 37 και 51 παρουσιάζει σημαντική βελτίωση και πλέον το ποσοστό επιτυχίας είναι πολύ καλό. Για *k*=51 όμως αυτό επιτυγχάνεται για πολλά γονίδια. Και εδώ ο Nearest Mean έχει σταθερά καλή απόδοση.

Πολύ σημαντική βελτίωση παρουσιάζει ο SVM για *C*=1, 100 και 1000. Για *C*=0.001 δεν παρατηρείται καλύτερο ποσοστό επιτυχίας πέραν αυτού που έχουμε στα αρχικά γονίδια.

Όσον αφορά τις τιμές των παραμέτρων *k* και *C* μπορούμε να παρατηρήσουμε τα εξής:

- Η καλύτερη απόδοση του *k*-NN επιτυγχάνεται για *k*=11 και γενικά για ενδιάμεσες τιμές του *k* – ούτε πολύ μικρές ούτε πολύ μεγάλες. Σταθερά η απόδοση του 71-Nearest Neighbor είναι πολύ χαμηλή.
- Η καλύτερη απόδοση του SVM επιτυγχάνεται για τις τιμές 1, 100 και 1000 της παραμέτρου *C*, όπου η συμπεριφορά του ταξινομητή είναι όμοια. Για *C*=0.001 η απόδοση είναι πολύ κακή – το καλύτερο ποσοστό επιτυχίας σημειώνεται στα 24188 γονίδια.

5.2.3 Filter Μέθοδος – Επιλογή γονιδίων με προεπεξεργασία δειγμάτων εκπαίδευσης

5.2.3.1 Προεπεξεργασία – Επιλογή αρχικού συνόλου γονιδίων

Γίνεται η προεπεξεργασία των δειγμάτων όπως αυτή περιγράφηκε στα Σχήματα 19 και 20. Εφαρμόζεται LOOCV στο train, γίνεται τυπική επιλογή γονιδίων με filter μέθοδο για κάθε ένα από τα 78 ζεύγη train και test και υπολογίζονται τα ποσοστά επιτυχίας σε κάθε βήμα επιλογής. Στην δεύτερη και τρίτη στήλη του Πίνακα 11 παρουσιάζεται για κάθε ταξινομητή η επανάληψη της επιλογής γονιδίων, για την οποία έχουμε μέγιστο μέσο ποσοστό επιτυχίας ανά τα 78 διαφορετικά τρεξίματα και το μέγεθος καθενός από τα 78 διαφορετικά σύνολα γονιδίων που προκύπτουν σε αυτό το βήμα αντίστοιχα. Στην τέταρτη στήλη φαίνεται το μέγιστο μέσο ποσοστό επιτυχίας και τελικά στην τελευταία στήλη φαίνεται το μέγεθος του συνόλου που προκύπτει από την ένωση των 78 διαφορετικών συνόλων γονιδίων. Τελικά, για κάθε ταξινομητή προκύπτει ένα σύνολο γονιδίων, υποσύνολο των 24188.

Παρατηρούμε ότι το μέγιστο μέσο ποσοστό επιτυχίας για κάθε ταξινομητή δεν είναι πολύ καλό. Γενικά όμως καταλήγουμε σε μικρά set γονιδίων εκτός από την περίπτωση του 11-Nearest Neighbor και του SVM για *C*=0.001 όπου κυρίως για τον δεύτερο το σύνολο γονιδίων που προκύπτει είναι πολύ μεγάλο. Επίσης παρατηρείται μια διαφοροποίηση όσον αφορά την παράμετρο *C*, όπου για την τιμή 1 ο SVM πλέον καταλήγει σε λιγότερα γονίδια από ότι για 100 και 1000 – μέχρι τώρα η συμπεριφορά του SVM ήταν ίδια για τις τρεις αυτές παραμέτρους. Επίσης για μεγάλα *k* (μεγαλύτερα η ίσα του 37) καταλήγουμε σε μικρά σύνολα γονιδίων ενώ για μικρά *k* σε μεγάλα σύνολα γονιδίων.

Ταξινομητής	**Επανάληψη Επιλογής γονιδίων**	**Μέγεθος συνόλου γονιδίων**	**Μέγιστο μέσο ποσοστό επιτυχίας (%)**	**Μέγεθος προεπιλεγμένου συνόλου**
Nearest Mean	14^{η}	2	71.79	5
1-Nearest Neighbor	9^{η}	64	69.23	205
5-Nearest Neighbor	7^{η}	256	70.51	701
11-Nearest Neighbor	4^{η}	2048	66.67	4016
37-Nearest Neighbor	14^{η}	2	73.08	5
51-Nearest Neighbor	13^{η}	4	71.79	11
71-Nearest Neighbor	15^{η}	1	56.41	3
SVM C=0.001	2^{η}	8192	64.1	13368
SVM C=1	14^{η}	2	74.36	5
SVM C=100	9^{η}	64	78.2	205
SVM C=1000	9^{η}	64	78.2	205

Πίνακας 11. Καρκίνος του μαστού, filter μέθοδος: Μέγιστο ποσοστό επιτυχίας και μέγεθος προεπιλεγμένου συνόλου για κάθε ταξινομητή

5.2.3.2 Τελική επιλογή γονιδίων με αφετηρία το προεπιλεγμένο σύνολο γονιδίων

Για κάθε έναν ταξινομητή καταλήγουμε σε ένα προεπιλεγμένο σύνολο γονιδίων. Κάνοντας επιλογή ξεκινώντας από αυτά τα γονίδια και χρησιμοποιώντας τα αρχικά train και independent test sets όπως περιγράφηκε στο Σχήμα 18, καταλήγουμε στα αποτελέσματα του Πίνακα 12.

Επιτυγχάνονται πολύ καλά ποσοστά επιτυχίας και μάλιστα αυτά παρατηρούνται για λίγα γονίδια. Και πάλι εξαίρεση αποτελεί ο SVM για *C*=0.001 που έχει χαμηλή απόδοση για πολλά γονίδια. Όσον αφορά τον Nearest Mean σχετικά με την επιλογή στα αρχικά γονίδια (Πίνακας 10), παρατηρείται μείωση στην απόδοση τόσο στο test όσο και στο train set. Παρόλα αυτά το σφάλμα γενίκευσης εξακολουθεί να είναι μικρό.

Ταξινομητής	Αριθμός Γονιδίων	Train set			Test set		
		SR %	SE %	SP %	SR %	SE %	SP %
Nearest Mean	2	74.36	70.59	77.27	84.21	83.33	85.71
1-Nearest Neighbor	10	100	100	100	89.47	91.6	85.71
5-Nearest Neighbor	11	87.18	91.17	84.09	94.74	100	85.71
11-Nearest Neighbor	512	84.61	94.12	77.27	89.47	100	71.43
37-Nearest Neighbor	2	76.92	67.65	84.09	78.95	75	85.71
51-Nearest Neighbor	7	75.64	64.71	84.09	78.95	75	85.71
71-Nearest Neighbor	1	56.41	0	100	36.84	0	100
SVM C=0.001	8192	85.90	76.47	93.18	73.68	66.67	85.71
SVM C=1	2	76.92	67.65	84.09	84.21	83.33	85.71
SVM C=100	183	100	100	100	84.21	75	100
SVM C=1000	183	100	100	100	84.21	75	100

Πίνακας 12. Καρκίνος του μαστού, filter μέθοδος: Επιλογή γονιδίων με αφετηρία το προεπιλεγμένο σύνολο γονιδίων – Μέγιστο ποσοστό επιτυχίας για κάθε ταξινομητή

Όσον αφορά τις τιμές των παραμέτρων *k* και *C* μπορούμε να πούμε τα εξής:

- Η απόδοση του *k*-NN πλέον είναι μεγαλύτερη για μικρά *k*, με μέγιστη για *k*=5. Για μεγάλα *k* η απόδοση πλησιάζει το 80%, οπότε μπορεί να χαρακτηριστεί καλή αλλά όχι ικανοποιητική. Σταθερά η απόδοση του 71-Nearest Neighbor είναι πολύ χαμηλή.
- Την καλύτερη απόδοση ο SVM την επιτυγχάνει για τις τιμές 100 και 1000 της παραμέτρου *C* (όπου η συμπεριφορά του SVM είναι όμοια), κυρίως όσον

αφορά το ποσοστό επιτυχίας στο train. Ακολουθεί η απόδοση για *C*=1 η οποία είναι ίδια στο test με τις προηγούμενες τιμές αλλά στο train είναι χαμηλή ακόμα και αν επιτυγχάνεται για λιγότερα γονίδια. Τέλος για *C*=0.001 δεν σημειώνεται καλή απόδοση.

Κατά μέσο όρο η αύξηση στο ποσοστό επιτυχίας είναι ίδιο με αυτό που είχαμε κατά την επιλογή στα αρχικά γονίδια, μόνο που εδώ το επιτυγχάνουμε για λιγότερα γονίδια.

Κατά την διαδικασία αυτή καταλήγουμε για κάθε ταξινομητή σε διαφορετικό σύνολο γονιδίων για το οποίο ο συγκεκριμένος έχει τη βέλτιστη απόδοση. Μας ενδιαφέρει να δούμε εάν υπάρχουν κοινά γονίδια μεταξύ αυτών των διαφορετικών συνόλων γονιδίων.

Λόγω του ότι ο ταξινομητής Nearest-Neighbor για *k*=71 έχει πολύ χαμηλή απόδοση, δεν τον λαμβάνουμε υπόψη στα συμπεράσματά μας όσον αφορά τα γονίδια στα οποία καταλήγουμε. Οπότε έχουμε να κάνουμε με δέκα διαφορετικά σύνολα γονιδίων.

Τα δέκα σύνολα γονιδίων στα οποία καταλήγουμε έχουν τομή τα δύο γονίδια στα οποία καταλήγουμε με τον Nearest Mean, 37-Nearest Neighbor και SVM για *C*=1. Κάθε σύνολο γονιδίων είναι υποσύνολο των μεγαλυτέρων του.

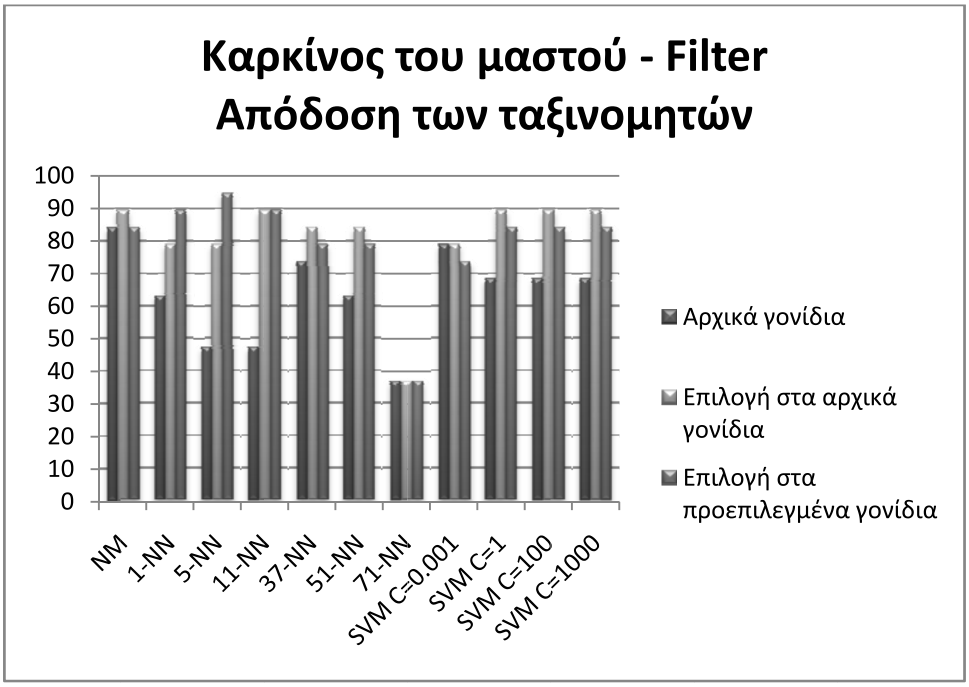
5.2.4 Filter Μέθοδος – Σύγκριση επιλογής στα αρχικά γονίδια και επιλογής στο προεπιλεγμένο σύνολο γονιδίων

Συγκρίνοντας τα αποτελέσματα που έχουμε με επιλογή στα αρχικά γονίδια με αυτά που έχουμε με επιλογή στο προεπιλεγμένο σύνολο γονιδίων, μπορούμε να βγάλουμε κάποιο συμπέρασμα για το αν η προεπεξεργασία των δειγμάτων εκπαίδευσης αποδίδει στην περίπτωση της filter μεθόδου για τον καρκίνο του μαστού.

Κατά την επιλογή στο αρχικό σύνολο γονιδίων τα γονίδια μειώνονται, εκτός από την περίπτωση του SVM για *C*=0.001 που έχει την καλύτερη απόδοση για τα αρχικά γονίδια. Παρατηρείται αύξηση στο ποσοστό επιτυχίας κατά 29.7% κατά μέσο όρο σε σχέση με αυτό που είχαμε για τα αρχικά γονίδια. Μάλιστα το ποσοστό επιτυχίας είναι κοντά στο 90% για όλους τους ταξινομητές, πέραν του SVM για *C*=0.001 και του *k*-Nearest Neighbor για *k*=1 και 5 όπου και πάλι όμως πλησιάζει το 80%.

Μετά την προεπεξεργασία, έχουμε μεγαλύτερη μείωση στα γονίδια εκτός και πάλι από την περίπτωση του SVM για *C*=0.001. Η απόδοση είναι στις περισσότερες των περιπτώσεων λιγότερη από αυτή που επιτυγχάνουμε με επιλογή στο αρχικό σύνολο γονιδίων. Εξαίρεση αποτελεί ο Nearest Neighbor για *k*=1 και 5, όπου παρατηρούμε πολύ καλύτερα αποτελέσματα και για *k*=11 όπου η απόδοση είναι σταθερή και για τον ίδιο αριθμό γονιδίων. Το ποσοστό επιτυχίας αυξάνεται κατά μέσο όρο 29.6% σε σχέση με το αρχικό και όλοι οι ταξινομητές, πέραν του SVM για *C*=0.001, έχουν καλή απόδοση. Για *k* πολύ μεγάλο ο *k*-Nearest Neighbor έχει πολύ χαμηλή απόδοση, για μικρό *k* είναι πολύ καλή, ενώ για *k*=37 και 51 είναι ικανοποιητική – κοντά στο 80%. Η απόδοση του SVM για *C*=1, 100 και 1000 είναι η ίδια στο test, όμως για *C*=1 αυτή επιτυγχάνεται για πολύ λιγότερα γονίδια και με χαμηλή απόδοση στο train set.

Στο Σχήμα 24 φαίνεται συγκριτικά η απόδοση των ταξινομητών στα αρχικά γονίδια με την απόδοση που επιτυγχάνεται κατά την επιλογή στο αρχικό σύνολο γονιδίων και με αυτή που επιτυγχάνεται κατά την επιλογή στο προεπιλεγμένο σύνολο γονιδίων.



Σχήμα 24. Καρκίνος του μαστού: Απόδοση των ταξινομητών για την filter μέθοδο

Παρατηρούμε ότι στην πλειοψηφία των περιπτώσεων η προεπεξεργασία των δειγμάτων εκπαίδευσης δεν οδηγεί σε καλύτερη απόδοση, αλλά οδηγεί σε λιγότερα γονίδια.

Προκειμένου να έχουμε μία γενικότερη άποψη για το τι τελικά επιτυγχάνεται με το στάδιο της προεπεξεργασίας, μελετάμε τα σύνολα γονιδίων στα οποία καταλήγουμε με και χωρίς την προεπεξεργασία (Πίνακας 12 και Πίνακας 10 αντίστοιχα). Όπως συμβαίνει και στα δεδομένα της λευχαιμίας για την filter μεθοδολογία, έτσι και εδώ τα γονίδια στα οποία καταλήγουμε με επιλογή στα προεπιλεγμένα γονίδια, είναι υποσύνολο αυτών που καταλήγουμε με επιλογή στα αρχικά γονίδια. Θα μπορούσαμε να καταλήξουμε στα ίδια γονίδια εάν από ένα σημείο και μετά στην πρώτη προσέγγιση μειώναμε κατά ένα γονίδιο τη φορά. Οπότε η διαδικασία της προεπεξεργασίας είναι μία χρονοβόρα διαδικασία που δεν οδηγεί σε καλύτερο αποτέλεσμα.

Όπως και με τα δεδομένα της λευχαιμίας, έτσι και εδώ υπολογίζουμε την ένωση των γονιδίων που προκύπτουν από κάθε ταξινομητή (Πίνακας 12). Δεν λαμβάνουμε υπόψη τα γονίδια που προκύπτουν από τον *k*-NN για *k*=11 (είναι πολλά σχετικά με τους υπόλοιπους ταξινομητές) και για *k*=71 λόγω χαμηλής απόδοσης. Ομοίως δεν λαμβάνουμε υπόψη τα γονίδια που προκύπτουν με τον SVM για *C*=0.001. Πραγματοποιείται γονιδιακή επιλογή σύμφωνα με το τυπικό σχήμα επιλογής – ταξινόμησης (Σχήμα 14) ελέγχοντας κάθε φορά την απόδοση με όλους τους ταξινομητές. Εντοπίζεται το ελάχιστο σύνολο γονιδίων για το οποίο κάθε ταξινομητής έχει την μέγιστη απόδοση και τα αποτελέσματα φαίνονται στον Πίνακα 13.

Παρατηρούμε ότι τα αποτελέσματα είναι αρκετά ικανοποιητικά για όλους τους ταξινομητές πλην του 71- Nearest Neighbor που σταθερά έχει χαμηλή απόδοση, του SVM για *C*=0.001 και του 51-Nearest Neighbor του οποίου όμως η απόδοση είναι περίπου 80% τόσο στο test όσο και στο train. Επίσης οι Nearest Mean και 37-Nearest Neighbor έχουν ικανοποιητική απόδοση στο test αλλά σχετικά χαμηλή στο train και μάλιστα ο τελευταίος για πολλά γονίδια.

Σχετικά με τα αποτελέσματα του Πίνακα 12, Οι ταξινομητές Nearest Neighbor για *k*=1, 5 και SVM για *C*=100, 1000 δεν παρουσιάζουν καμία διαφορά. Παρόλα αυτά, για τους υπόλοιπους ταξινομητές παρατηρείται βελτίωση στο αποτέλεσμα.

Σημαντική είναι η βελτίωση του 11-Nearest Neighbor όπου έχουμε μεγάλη μείωση στα γονίδια. για τον 37-Nearest Neighbor βελτιώνεται η απόδοση στο test μόνο που αυξάνονται πολύ τα γονίδια. Επίσης να παρατηρήσουμε ότι ο SVM για *C*=1 πλέον πλησιάζει την συμπεριφορά των τιμών 100 και 1000 της παραμέτρου *C*.

Ταξινομητής	**Αριθμός Γονιδίων**	**Train set**			**Test set**		
		SR %	**SE %**	**SP %**	**SR %**	**SE %**	**SP %**
Nearest Mean	11	76.92	73.53	79.54	84.21	91.67	71.43
1-Nearest Neighbor	10	100	100	100	89.47	91.67	85.71
5-Nearest Neighbor	11	87.18	91.18	84.09	94.74	100	85.71
11-Nearest Neighbor	7	83.33	79.41	86.36	89.47	91.67	85.71
37-Nearest Neighbor	153	78.20	76.47	79.54	84.21	83.33	85.71
51-Nearest Neighbor	9	80.77	73.53	86.36	78.95	83.33	71.43
71-Nearest Neighbor	1	56.41	0	100	36.84	0	100
SVM C=0.001	179	100	100	100	68.42	66.67	71.43
SVM C=1	179	100	100	100	84.21	83.33	85.71
SVM C=100	183	100	100	100	84.21	75	100
SVM C=1000	183	100	100	100	84.21	75	100

Πίνακας 13. Καρκίνος του μαστού, filter μέθοδος: Επιλογή γονιδίων με αφετηρία την ένωση συνόλων γονιδίων από τους διαφορετικούς ταξινομητές

Διαφορά λοιπόν σε σχέση με τα αποτελέσματα του Πίνακα 12 παρουσιάζουν οι ταξινομητές που κατά το στάδιο της προεπεξεργασίας καταλήγουν σε πολύ μικρά σύνολα γονιδίων (NM, 37-NN, 51-NN και SVM για *C*=1) και οι ταξινομητές που καταλήγουν σε πολύ μεγάλα σύνολα γονιδίων (11-NN, SVM για *C*=0.001). Αυτοί που κατέληξαν σε πολύ μικρά σύνολα γονιδίων παρουσιάζουν βελτίωση, άρα συμπεραίνουμε ότι το στάδιο της προεπεξεργασίας για αυτούς δεν είχε καλό αποτέλεσμα. Το ίδιο συμβαίνει και για τον 1-NN.

Οι ταξινομητές που δεν παρουσιάζουν καμία διαφορά στο αποτέλεσμα σε σχέση με τον Πίνακα 12, κατά την προεπεξεργασία καταλήγουν σε γονίδια που είναι υπερσύνολο των 183 (ένωση γονιδίων). Επειδή πρόκειται για filter μέθοδο επιλογής όπου το Fisher's Criterion παραμένει ίδιο για τα γονίδια, η ταξινόμηση των γονιδίων γίνεται με τον ίδιο τρόπο οπότε τελικά με επιλογή γονιδίων στην ένωση καταλήγουμε στα ίδια γονίδια με τον Πίνακα 12.

5.2.5 Wrapper μέθοδος – Τυπική διαδικασία επιλογής/ταξινόμησης στο αρχικό σύνολο γονιδίων

Εφαρμόζεται το τυπικό σχήμα επιλογής/ταξινόμησης (Σχήμα 14) που περιγράφηκε στο προηγούμενο κεφάλαιο για wrapper μεθοδολογία επιλογής γονιδίων. Σε κάθε στάδιο επιλογής γονιδίων γίνεται εκπαίδευση και εκτίμηση της απόδοσης των ταξινομητών. Μετά την διαδικασία επιλογής, για κάθε έναν από τους ταξινομητές εντοπίζεται το ελάχιστο σύνολο γονιδίων για το οποίο επιτυγχάνεται η μέγιστη απόδοση. Τα αποτελέσματα φαίνονται στον Πίνακα 14.

Ο Nearest Mean έχει την ίδια απόδοση με πριν στο test set, με λίγο μεγαλύτερο ποσοστό επιτυχίας στο train set.

Ο Nearest Neighbor για *k*=71 συνεχίζει να έχει το ίδιο χαμηλό ποσοστό επιτυχίας σε κάθε σύνολο γονιδίων, στο οποίο καταλήγουμε σε κάθε στάδιο επιλογής γονιδίων. Για *k*=1 έχουμε μικρή αύξηση του ποσοστού επιτυχίας στο test, όχι όμως τόση ώστε να ξεπεραστεί το φαινόμενο του over-fitting. Για *k*=5,11 παρατηρούμε ότι υπάρχει εξαιρετική βελτίωση στην απόδοση τόσο στο train όσο και στο test set. Παρόλα αυτά η απόδοση στο test εξακολουθεί να είναι χαμηλή. Ο 37-Nearest Neighbor πλέον πλησιάζει το 80% στο test με καλή απόδοση στο train και για λίγα γονίδια. Για *k*=51 έχουμε εξαιρετική απόδοση στο test όχι όμως τόσο ικανοποιητική στο train.

Ο SVM για *C*=0.001 έχει την ίδια συμπεριφορά με πριν, με λίγο μεγαλύτερο ποσοστό επιτυχίας στο train set και σταθερό στο test set. Σταθερά χαμηλή είναι η απόδοση του SVM για *C*=1, 100 και 1000 μόνο που εδώ την πετυχαίνουμε στα 8192 γονίδια και όχι στα 24188.

Στις περιπτώσεις όπου η απόδοση μπορεί να θεωρηθεί καλή (NM,51-NN), τα γονίδια ναι μεν μειώνονται, όμως δε μπορούμε να θεωρήσουμε ότι έχουμε το βέλτιστο δυνατό αποτέλεσμα αφού και πάλι είναι πολλά.

Ταξινομητής	Αριθμός Γονιδίων	Train set			Test set		
		SR %	SE %	SP %	SR %	SE %	SP %
Nearest Mean	8192	91.02	94.12	88.64	84.21	83.33	85.71
1-Nearest Neighbor	64	100	100	100	73.68	83.33	57.14
5-Nearest Neighbor	16	87.18	76.47	95.45	68.42	66.67	71.43
11-Nearest Neighbor	16	88.46	82.35	93.18	73.68	66.67	85.71
37-Nearest Neighbor	256	85.90	82.35	88.64	78.95	91.67	57.14
51-Nearest Neighbor	1024	79.49	67.65	88.64	84.21	83.33	85.71
71-Nearest Neighbor	1	56.41	0	100	36.84	0	100
SVM C=0.001	8192	96.15	97.06	95.45	78.95	83.33	71.43
SVM C=1	8192	100	100	100	68.42	91.67	28.57
SVM C=100	8192	100	100	100	68.42	91.67	28.57
SVM C=1000	8192	100	100	100	68.42	91.67	28.57

Πίνακας 14. Καρκίνος του μαστού, wrapper μέθοδος: Τυπική διαδικασία επιλογής/ταξινόμησης στα αρχικά γονίδια – Μέγιστο ποσοστό επιτυχίας για κάθε ταξινομητή

Όσον αφορά τις τιμές των παραμέτρων *k* και *C* μπορούμε να παρατηρήσουμε τα εξής:

- Ο *k*-NN έχει καλύτερη απόδοση για ενδιάμεσες τιμές της παραμέτρου *k* (37 και 51).
- Ο SVM έχει καλύτερη απόδοση για την τιμή 0.001 της παραμέτρου *C* ενώ η συμπεριφορά του είναι η ίδια για τις τιμές 1, 100 και 1000 της παραμέτρου.

Συμπεραίνουμε λοιπόν ότι παρόλο που το ποσοστό επιτυχίας αυξήθηκε δραματικά σε ορισμένες περιπτώσεις, τα αποτελέσματα δεν είναι ικανοποιητικά εκτός από την περίπτωση του Nearest Mean που έχει την καλύτερη απόδοση για πολλά όμως γονίδια.

5.2.6 Wrapper Μέθοδος – Επιλογή γονιδίων με προεπεξεργασία δειγμάτων

5.2.6.1 Προεπεξεργασία – Επιλογή αρχικού συνόλου γονιδίων

Γίνεται η προεπεξεργασία των δειγμάτων όπως αυτή περιγράφηκε στα Σχήματα 19 και 20. Εφαρμόζεται LOOCV στο train, γίνεται τυπική επιλογή γονιδίων με wrapper μέθοδο για κάθε ένα από τα 78 ζεύγη train και test και υπολογίζονται τα ποσοστά επιτυχίας σε κάθε βήμα επιλογής.

Ταξινομητής	Επανάληψη Επιλογής γονιδίων	Μέγεθος συνόλου γονιδίων	Μέγιστο μέσο ποσοστό επιτυχίας	Μέγεθος προεπιλεγμένου συνόλου
Nearest Mean	6^{η}	512	67.95	1320
1-Nearest Neighbor	6^{η}	512	66.67	1320
5-Nearest Neighbor	11^{η}	16	62.82	315
11-Nearest Neighbor	12^{η}	8	65.38	204
37-Nearest Neighbor	11^{η}	16	64.10	315
51-Nearest Neighbor	6^{η}	512	62.82	1320
71-Nearest Neighbor	15^{η}	1	56.41	33
SVM C=0.001	2^{η}	8192	62.82	13262
SVM C=1	6^{η}	512	66.66	1320
SVM C=100	6^{η}	512	66.66	1320
SVM C=1000	6^{η}	512	66.66	1320

Πίνακας 15. Καρκίνος του μαστού, wrapper μέθοδος: Μέγιστο μέσο ποσοστό επιτυχίας και μέγεθος προεπιλεγμένου συνόλου για κάθε ταξινομητή μετά το LOOCV

Στην δεύτερη και τρίτη στήλη του Πίνακα 15 παρουσιάζεται για κάθε ταξινομητή η επανάληψη της επιλογής γονιδίων, για την οποία έχουμε μέγιστο μέσο ποσοστό επιτυχίας ανά τα 78 διαφορετικά τρεξίματα και το μέγεθος καθενός από τα 78 διαφορετικά σύνολα γονιδίων που προκύπτουν σε αυτό το βήμα αντίστοιχα. Στην τέταρτη στήλη φαίνεται το μέγιστο μέσο ποσοστό επιτυχίας και τελικά στην τελευταία στήλη φαίνεται το μέγεθος του συνόλου, που προκύπτει από την ένωση των 78 διαφορετικών συνόλων γονιδίων. Τελικά, για κάθε ταξινομητή προκύπτει ένα σύνολο γονιδίων, υποσύνολο των 24188.

Παρατηρούμε ότι το μέσο ποσοστό επιτυχίας δεν είναι καλό για κανέναν ταξινομητή. Επίσης έξι ταξινομητές καταλήγουν στο ίδιο σύνολο των 1320 γονιδίων, δύο σε σύνολο των 315 γονιδίων, ενώ μόνο δύο ταξινομητές καταλήγουν σε μικρό σύνολο γονιδίων – ο 4-Nearest Neighbor καταλήγει σε 64 γονίδια και ο 71-Nearest Neighbor σε 33.

5.2.6.2 Τελική επιλογή γονιδίων με αφετηρία το προεπιλεγμένο σύνολο γονιδίων

Για κάθε έναν ταξινομητή καταλήγουμε σε ένα προεπιλεγμένο σύνολο γονιδίων (Πίνακας 15). Κάνοντας επιλογή ξεκινώντας από αυτά τα γονίδια και χρησιμοποιώντας τα αρχικά train και independent test sets όπως περιγράφηκε στο Σχήμα 18, καταλήγουμε στα αποτελέσματα του Πίνακα 16.

Ο Nearest Mean έχει σταθερά καλή απόδοση και για πολύ λίγα γονίδια, σε αντίθεση με την επιλογή στο αρχικό σύνολο γονιδίων, όπου τα γονίδια ήταν πολλά. Παρατηρείται λίγο χαμηλότερη απόδοση στο train set.

Ομοίως, πολύ καλή απόδοση και για λίγα γονίδια έχει ο Nearest Neighbor για k=1, 11 και 37. Για k=1 δεν έχουμε πλέον το φαινόμενο του over-fitting. Για k=5 και 51 η απόδοση πλησιάζει το 80% στο test και είναι καλή στο train. Για τον 71-Nearest Neighbor η απόδοση παραμένει εξαιρετικά χαμηλή τόσο στο train όσο και στο test set.

Το 80% πλησιάζει και η απόδοση του SVM, μόνο που στην περίπτωση της τιμής 0.001 της παραμέτρου *C* αυτό συμβαίνει για πολλά γονίδια.

Όσον αφορά τις τιμές των παραμέτρων *k* και *C* μπορούμε να παρατηρήσουμε τα εξής:

- Η καλύτερη απόδοση του *k*-NN παρατηρείται για πολύ μικρό *k* (*k*=1), και για ενδιάμεσες τιμές του (11 και 37).
- Ο SVM έχει καλύτερη απόδοση για την τιμή 100 της παραμέτρου *C* (λίγο καλύτερη απόδοση στο train). Αμέσως καλύτερη είναι η απόδοση των τιμών 1 και 1000 (πάλι κρίνοντας ως προς την απόδοση στο train παρόλο που για *C*=1000 έχουμε λίγο λιγότερα γονίδια), και τελευταία ακολουθεί η απόδοση για *C*=0.001 (ίδια απόδοση με τις υπόλοιπες τιμές, μόνο που επιτυγχάνεται για πολύ περισσότερα γονίδια). Για *C*=1 και 100 καταλήγουμε στα ίδια γονίδια.

Ταξινομητής	**Αριθμός Γονιδίων**	**Train set**			**Test set**		
		SR %	**SE %**	**SP %**	**SR %**	**SE %**	**SP %**
Nearest Mean	8	88.33	85.29	81.82	84.21	91.67	71.43
1-Nearest Neighbor	6	100	100	100	84.21	91.67	71.43
5-Nearest Neighbor	83	94.87	97.06	93.18	78.95	83.33	71.43
11-Nearest Neighbor	13	79.49	70.59	86.36	84.21	75	100
37-Nearest Neighbor	86	85.90	97.06	77.27	84.21	100	57.14
51-Nearest Neighbor	128	83.33	76.47	88.63	78.95	83.33	71.42
71-Nearest Neighbor	1	56.41	0	100	36.84	0	100
SVM C=0.001	8192	96.15	97.06	95.45	78.95	83.33	71.43
SVM C=1	10	94.87	94.12	95.45	78.95	75	85.71
SVM C=100	10	98.72	97.06	100	78.95	75	85.71
SVM C=1000	8	92.31	88.23	95.45	78.95	91.67	57.14

Πίνακας 16. Καρκίνος του μαστού, wrapper μέθοδος: Επιλογή γονιδίων με αφετηρία το προεπιλεγμένο σύνολο γονιδίων– Μέγιστο ποσοστό επιτυχίας για κάθε ταξινομητή

Συμπεραίνουμε ότι σε γενικές γραμμές έχουμε πολύ καλά αποτελέσματα – κοντά στο 80% για το test – και για λίγα γονίδια.

Καταλήγουμε στο συμπέρασμα ότι η προεπεξεργασία των δειγμάτων εκπαίδευσης οδηγεί σε καλύτερο αποτέλεσμα από αυτό που πετυχαίνουμε με επιλογή στο αρχικό σύνολο γονιδίων.

Κατά την διαδικασία αυτή καταλήγουμε για κάθε ταξινομητή σε διαφορετικό σύνολο γονιδίων για το οποίο ο συγκεκριμένος έχει τη βέλτιστη απόδοση. Όπως έγινε και με την filter μέθοδο έτσι και εδώ μας ενδιαφέρει να δούμε εάν υπάρχουν κοινά γονίδια μεταξύ αυτών των διαφορετικών συνόλων γονιδίων.

Λόγω του ότι ο ταξινομητής Nearest-Neighbor για *k*=71 δε δίνει ικανοποιητικά αποτελέσματα, δεν τον λαμβάνουμε υπόψη στα συμπεράσματά μας όσον αφορά τα γονίδια στα οποία καταλήγουμε. Οπότε έχουμε να κάνουμε με δέκα διαφορετικά σύνολα γονιδίων.

Ανάμεσα στα δέκα διαφορετικά σύνολα γονιδίων στα οποία καταλήγουμε με τον κάθε ταξινομητή δεν υπάρχει τομή, δεν υπάρχει δηλαδή έστω ένα κοινό γονίδιο. Ένα γονίδιο απαντάται σε εννέα των δέκα περιπτώσεων και ένα σε οκτώ. Ακολουθούν γονίδια με μικρότερη συχνότητα.

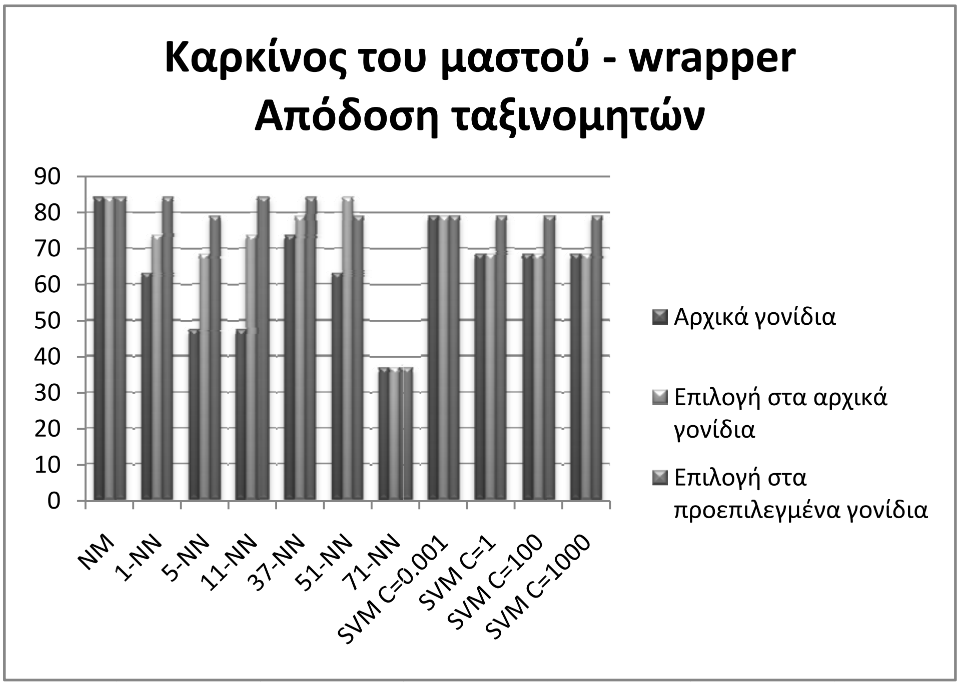
5.2.7 Wrapper Μέθοδος – Σύγκριση επιλογής στα αρχικά γονίδια και επιλογής στο προεπιλεγμένο σύνολο γονιδίων

Συγκρίνοντας τα αποτελέσματα που έχουμε με επιλογή στα αρχικά γονίδια με αυτά που έχουμε με επιλογή στο προεπιλεγμένο σύνολο γονιδίων, μπορούμε να βγάλουμε κάποιο συμπέρασμα για το αν η προεπεξεργασία των δειγμάτων εκπαίδευσης αποδίδει στην περίπτωση της wrapper μεθόδου.

Κατά την επιλογή στα αρχικά γονίδια έχουμε αύξηση στο ποσοστό επιτυχίας όχι όμως αρκετή για να χαρακτηρίσουμε το αποτέλεσμα καλό. Για τον *k*-Nearest Neighbor η καλύτερη απόδοση παρατηρείται για *k*=51 και ακολουθεί η απόδοση για *k*=37. Για πολύ μεγάλο *k* δεν έχουμε ικανοποιητικό αποτέλεσμα. Η απόδοση του SVM για *C*=1,100 και 1000 είναι λιγότερη από αυτή για *C*=0.001 και μάλιστα για τις τρεις πρώτες τιμές τις παραμέτρου η συμπεριφορά του SVM είναι η ίδια. Καλή απόδοση έχει μόνο ο Nearest mean και ο *k*-Nearest Neighbor για *k*=51 για πολλά όμως γονίδια – 8192 και 1024 αντίστοιχα.

Κατά την επιλογή στο προεπιλεγμένο σύνολο γονιδίων καταλήγουμε πλέον σε λίγα γονίδια εκτός από τον SVM για *C*=0.001. Η αύξηση στο ποσοστό επιτυχίας είναι μεγαλύτερη από την πρώτη προσέγγιση. Καλή απόδοση έχουν πλέον οι Nearest Mean, *k*-Nearest Neighbor για *k*=1, 11 και 37 και οι υπόλοιποι έχουν ποσοστό επιτυχίας κοντά στο 80% για λίγα γονίδια. Για μεγάλο *k* δεν έχουμε καλή απόδοση.

Στο Σχήμα 25 φαίνεται συγκριτικά η απόδοση των ταξινομητών στα αρχικά γονίδια με την απόδοση που επιτυγχάνεται κατά την επιλογή στο αρχικό σύνολο γονιδίων και με αυτή που επιτυγχάνεται κατά την επιλογή στο προεπιλεγμένο σύνολο γονιδίων.



Σχήμα 25. Καρκίνος του μαστού: Απόδοση των ταξινομητών για την wrapper μέθοδο.

Παρατηρούμε ότι στην πλειοψηφία τους οι ταξινομητές έχουν καλύτερη απόδοση μετά την προεπεξεργασία των δειγμάτων εκπαίδευσης.

Προκειμένου να έχουμε μία γενικότερη άποψη για το τι τελικά επιτυγχάνεται με το στάδιο της προεπεξεργασίας, μελετάμε τα σύνολα γονιδίων στα οποία καταλήγουμε με και χωρίς την προεπεξεργασία (Πίνακας 16 και Πίνακας 14). Σε αντίθεση με τα δεδομένα της λευχαιμίας, μετά την προεπεξεργασία υπάρχουν κοινά γονίδια με αυτά που έχουμε χωρίς την προεπεξεργασία. Τα γονίδια στα οποία καταλήγουμε με τους Nearest Mean, 51-Nearest Neighbor και SVM για *C*=1, 100 και 1000 είναι υποσύνολο αυτών στα οποία καταλήγουμε χωρίς την προεπεξεργασία. Σε όλες τις υπόλοιπες περιπτώσεις ταξινομητών υπάρχουν επίσης κοινά γονίδια αλλά δεν αποτελούν υποσύνολο αυτών χωρίς την προεπεξεργασία. Πιο συγκεκριμένα για τον *k*-NN για *k*=1 έχουμε τρία κοινά γονίδια, για *k*=5 πέντε κοινά, για *k*=11 ένα κοινό γονίδιο, για *k*=37 εβδομήντα ένα κοινά γονίδια και για τον SVM για *C*=0.001 8073 κοινά γονίδια.

Παρόλο που σε ορισμένες περιπτώσεις τα γονίδια μετά την προεπεξεργασία είναι υποσύνολο αυτών χωρίς την προεπεξεργασία, δεν παρατηρείται το φαινόμενο που παρατηρείται στην filter μέθοδο επιλογής γονιδίων. Δηλαδή με το στάδιο της προεπεξεργασίας καταλήγουμε σε διαφορετικά γονίδια από εκείνα στα οποία καταλήγουμε χωρίς την προεπεξεργασία. Για παράδειγμα τα 8 γονίδια στα οποία καταλήγουμε με τον Nearest Mean είναι μεν υποσύνολο των 8192 που καταλήγουμε με τον ίδιο ταξινομητή χωρίς προεπεξεργασία, δεν είναι όμως τα ίδια με τα 8 γονίδια στα οποία καταλήγουμε χωρίς την προεπεξεργασία. Λόγω του ότι με την wrapper μεθοδολογία σε κάθε επανάληψη υπολογίζονται εκ νέου τα βάρη των γονιδίων και δεν είναι στατικά όπως συμβαίνει με την filter, το στάδιο της προεπεξεργασίας βοηθάει πραγματικά στον εντοπισμό αποδοτικότερων γονιδίων.

Επομένως το στάδιο της προεπεξεργασίας ευνοεί την απόδοση των ταξινομητών στην περίπτωση της wrapper μεθοδολογίας για επιλογή γονιδίων. Προχωρώντας ένα βήμα παραπέρα όπως κάναμε και στην περίπτωση της λευχαιμίας, συγκεντρώνουμε τα σύνολα των γονιδίων που προκύπτουν για κάθε ταξινομητή μετά την επιλογή στο προεπιλεγμένο σύνολο γονιδίων (Πίνακας 16). Δεν λαμβάνονται υπόψη τα γονίδια που προκύπτουν από τον Nearest Neighbor για *k* = 71 (λόγω του ότι η απόδοσή του είναι εξαιρετικά χαμηλή) καθώς και τα γονίδια που προκύπτουν από τον SVM για *C* = 0.001 (λόγω του ότι είναι πάρα πολλά). Υπολογίζεται η ένωση των υπόλοιπων συνόλων γονιδίων η οποία αποτελείται από 166 γονίδια. Τελικά γίνεται επιλογή γονιδίων σε αυτό το σύνολο σύμφωνα με την τυπική επιλογή γονιδίων του Σχήματος 14 για wrapper μεθοδολογία, μειώνοντας κατά ένα γονίδιο τη φορά και ελέγχοντας την απόδοση με όλους τους ταξινομητές. Τα αποτελέσματα που προκύπτουν φαίνονται στον Πίνακα 17.

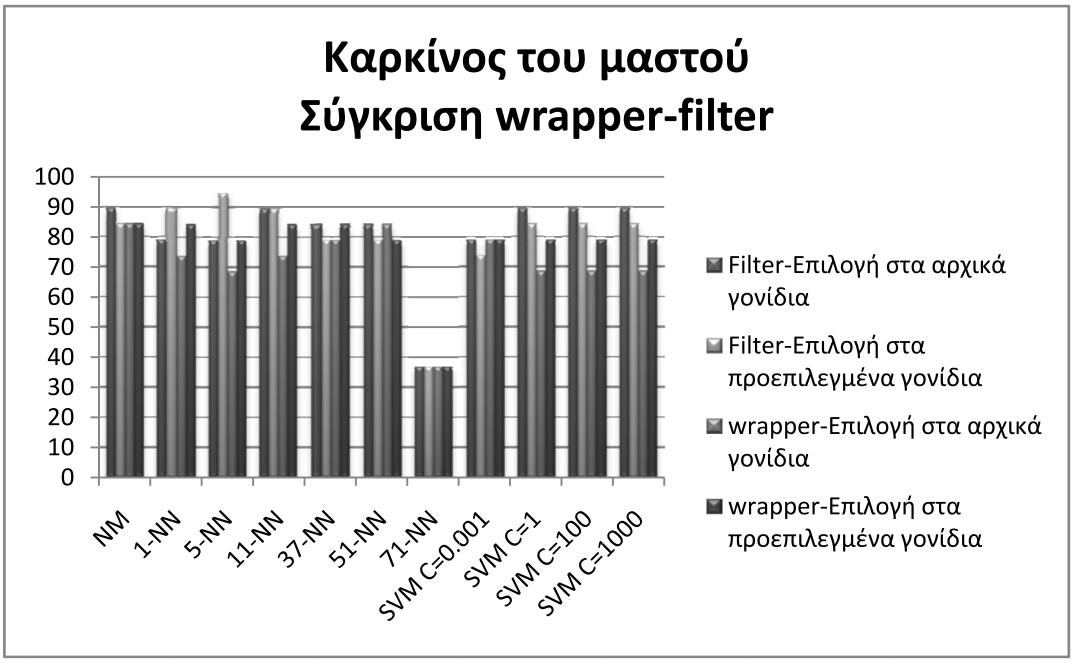
Ταξινομητής	Αριθμός Γονιδίων	Train set			Test set		
		SR %	SE %	SP %	SR %	SE %	SP %
Nearest Mean	19	87.18	94.12	81.82	89.47	91.67	85.71
1-Nearest Neighbor	48	100	100	100	84.21	91.67	71.43
5-Nearest Neighbor	26	93.59	91.18	95.45	84.21	100	57.14
11-Nearest Neighbor	25	84.61	94.12	77.27	89.47	100	71.43
37-Nearest Neighbor	24	80.77	88.23	75	89.47	100	71.43
51-Nearest Neighbor	20	82.05	88.23	77.27	84.21	83.33	85.71
71-Nearest Neighbor	1	56.41	0	100	36.84	0	100
SVM C=0.001	1	56.41	0	100	36.84	0	100
SVM C=1	22	100	100	100	78.95	83.33	71.43
SVM C=100	27	100	100	100	84.21	91.67	71.43
SVM C=1000	27	100	100	100	84.21	91.67	71.43

Πίνακας 17. Καρκίνος του μαστού, wrapper μέθοδος: Επιλογή γονιδίων με αφετηρία την ένωση συνόλων γονιδίων από τους διαφορετικούς ταξινομητές

Παρατηρούμε ότι η πλειοψηφία των ταξινομητών έχει απόδοση μεγαλύτερη από 84%, κάτι το οποίο είναι πολύ ικανοποιητικό για τα δεδομένα του καρκίνου του μαστού. Εξαίρεση αποτελούν ο SVM για C = 1, όπου όμως και πάλι η απόδοση πλησιάζει το 80%, καθώς και οι ταξινομητές που δεν συμπεριλήφθηκαν στην ένωση των γονιδίων, δηλαδή οι 71-NN και SVM για C = 0.001 των οποίων η απόδοση είναι εξαιρετικά χαμηλή.

5.2.8 Σύγκριση filter και wrapper μεθόδων επιλογής

Στο Σχήμα 26 φαίνεται συνολικά η απόδοση των ταξινομητών για τα δεδομένα του καρκίνου του μαστού, τόσο με επιλογή (filter και wrapper) στα αρχικά γονίδια, όσο και με επιλογή (filter και wrapper) στο προεπιλεγμένο σύνολο γονιδίων.



Σχήμα 26. Σύγκριση απόδοσης ταξινομητών με wrapper και filter για τα δεδομένα της Λευχαιμίας

Όσον αφορά την επιλογή στα αρχικά γονίδια, η filter μέθοδος είναι καλύτερη από την wrapper. Μοναδική εξαίρεση αποτελούν ο SVM για *C*=0.001 και ο *k*-Nearest Neighbor για *k*=51, όπου η wrapper είναι καλύτερη της filter, όχι ως προς την απόδοση αλλά ως προς το πλήθος των γονιδίων στα οποία αυτή επιτυγχάνεται.

Όσον αφορά την επιλογή στα προεπιλεγμένα γονίδια δεν είναι ξεκάθαρο από το σχήμα ποια είναι καλύτερη μέθοδος επιλογής για όλους τους ταξινομητές. Έτσι παρατηρούμε τα εξής:

- Για τους NM και 51-NN, καλύτερη είναι η filter – ίδια απόδοση με τη wrapper για λιγότερα όμως γονίδια.

- Για τους 1-NN, 5-NN, και SVM για *C*=1, 100 και 1000, καλύτερη είναι επίσης η filter, ως προς την απόδοση.
- Για τον 11-NN είναι καλύτερη η wrapper ως προς τα γονίδια.
- Για τους 37-NN και SVM για *C*=0.001 καλύτερη είναι η wrapper ως προς την απόδοση.

Όσον αφορά την τιμή της παραμέτρου *C* , για 0.001 έχουμε χειρότερη απόδοση από ότι για τις υπόλοιπες τρεις τιμές της παραμέτρου. Εξαίρεση αποτελεί η επιλογή στα αρχικά γονίδια με χρήση της wrapper μεθόδου.

Όσον αφορά την τιμή της παραμέτρου *k* δε μπορούμε να βγάλουμε κάποιο συγκεκριμένο συμπέρασμα, σε κάθε περίπτωση η συμπεριφορά του ταξινομητή *k*-NN είναι διαφορετική.

Γενικά η προεπεξεργασία ωφελεί την wrapper μέθοδο – εκτός από τον 51-NN – ενώ δεν ωφελεί την filter – εκτός από τους 1-NN και 5-NN.

Επίσης να παρατηρήσουμε ότι ανάμεσα στα είκοσι σύνολα γονιδίων που καταλήγουμε με τις δύο μεθόδους (wrapper και filter) δεν υπάρχει τομή και το πιο συχνό γονίδιο απαντάται σε δεκαοκτώ των είκοσι διαφορετικών περιπτώσεων.

5.3 Σύγκριση αλγορίθμων επιλογής γονιδίων και πως η προεπεξεργασία επηρεάζει το αποτέλεσμα

Έχοντας παρουσιάσει αναλυτικά τα αποτελέσματα για κάθε μία παθολογία, συνεχίζουμε στο σχολιασμό των αποτελεσμάτων με σκοπό να συγκριθούν οι δύο μεθοδολογίες που ακολουθήθηκαν (filter και wrapper), καθώς και να εξεταστεί κατά πόσο το στάδιο της προεπεξεργασίας βοηθάει στο τελικό αποτέλεσμα.

Για το σκοπό ατό εξετάζονται και πάλι οι δύο παθολογίες ξεχωριστά. Τα αποτελέσματα παρουσιάζονται στις επόμενες δύο υποενότητες.

5.3.1 Σύγκριση των αποτελεσμάτων της λευχαιμίας

Προκειμένου να γίνει σύγκριση των αποτελεσμάτων που έχουμε χωρίς και με την προεπεξεργασία, συγκεντρώνονται οι μέγιστες αποδόσεις των ταξινομητών σε κάθε περίπτωση. Ο κάθε ταξινομητής έχει μέγιστο ποσοστό επιτυχίας για μια συγκεκριμένη τιμή της παραμέτρου του, οπότε λαμβάνουμε υπόψη μόνο το αποτέλεσμα για εκείνη την τιμή της παραμέτρου του. Για παράδειγμα στην wrapper μεθοδολογία ο *k*-Nearest Neighbor έχει καλύτερη απόδοση για *k*=1 χωρίς την προεπεξεργασία, οπότε λαμβάνεται υπόψη η απόδοσή του μόνο για αυτή την τιμή της παραμέτρου. Αντίθετα μετά την προεπεξεργασία ο ίδιος ταξινομητής έχει καλύτερη απόδοση για *k*=5, οπότε υπολογίζεται μόνο η απόδοσή του για αυτή τη συγκεκριμένη τιμή του *k*. Το ίδιο ισχύει και για τις διάφορες τιμές της παραμέτρου *C* του SVM. Τελικά έχουμε τρία αποτελέσματα, ένα για κάθε ταξινομητή. Εξετάζουμε κάθε μεθοδολογία επιλογής γονιδίων ξεχωριστά.

Ταξινομητής	**Χωρίς προεπεξεργασία**			**Με προεπεξεργασία**			**Κοινά γονίδια**
	Αριθμός γονιδίων	**SR στο train**	**SR στο test**	**Αριθμός γονιδίων**	**SR στο train**	**SR στο test**	
Nearest Mean	64	100%	88.23%	6	100%	88.23%	9.375%
***k*-Nearest Neighbor**	4	100%	88.23%	3	100%	88.23%	75%
SVM	128	100%	91.18%	5	100%	88.23%	3.91%
Κοινά γονίδια	3.125%			50%			

Πίνακας 18. Λευχαιμία, Filter μέθοδος: Σύγκριση πριν και μετά την προεπεξεργασία

Ταξινομητής	**Χωρίς προεπεξεργασία**			**Με προεπεξεργασία**			**Κοινά γονίδια**
	Αριθμός γονιδίων	**SR στο train**	**SR στο test**	**Αριθμός γονιδίων**	**SR στο train**	**SR στο test**	
Nearest Mean	4	100%	85.29%	2	97.37%	94.12%	0%
***k*-Nearest Neighbor**	4	100%	82.35%	1	97.37%	91.18%	25%
SVM	4	100%	82.35%	4	100%	97.06%	0%
Κοινά γονίδια	100%			0%			

Πίνακας 19. Λευχαιμία, Wrapper μέθοδος: Σύγκριση πριν και μετά την προεπεξεργασία

Στους Πίνακες 18 και 19 παρουσιάζονται τα αποτελέσματα για την filter και την wrapper μεθοδολογία αντίστοιχα. Φαίνεται ο αριθμός των γονιδίων για τα οποία επιτυγχάνεται η μέγιστη απόδοση, ενώ στην τελευταία στήλη φαίνεται το ποσοστό των γονιδίων που είναι κοινά πριν και μετά το στάδιο της προεπεξεργασίας. Τέλος στην τελευταία σειρά παρουσιάζεται το ποσοστό των γονιδίων που είναι κοινά στους τρεις ταξινομητές τόσο χωρίς όσο και με την προεπεξεργασία.

Όσον αφορά την συνεισφορά του σταδίου της προεπεξεργασίας έχουμε να παρατηρήσουμε τα εξής:

- Η προεπεξεργασία βοηθάει σημαντικά την wrapper μεθοδολογία επιλογής γονιδίων τόσο με αύξηση της απόδοσης των ταξινομητών όσο και με μείωση των γονιδίων.
- Στη filter μεθοδολογία έχουμε μείωση των γονιδίων μετά την προεπεξεργασία για την ίδια απόδοση που είχαμε χωρίς την προεπεξεργασία, εκτός του SVM όπου έχουμε μείωση στην απόδοση.
- Χωρίς την προεπεξεργασία στην filter παρατηρείται μικρή επικάλυψη ανάμεσα στα γονίδια για κάθε ταξινομητή, ενώ μετά την προεπεξεργασία παρατηρείται μεγαλύτερη επικάλυψη (κάθε σύνολο γονιδίων είναι υποσύνολο των μεγαλυτέρων του).
- Χωρίς την προεπεξεργασία στην wrapper οι ταξινομητές έχουν μέγιστη απόδοση για τα ίδια γονίδια, ενώ μετά την προεπεξεργασία για διαφορετικά γονίδια (υπάρχει μόνο ένα κοινό γονίδιο ανάμεσα στους Nearest Mean και SVM).
- Στην filter μεθοδολογία τα γονίδια μετά την προεπεξεργασία είναι υποσύνολο αυτών πριν την προεπεξεργασία για κάθε ταξινομητή ενώ στην wrapper υπάρχει μόνο ένα κοινό γονίδιο στην περίπτωση του *k*-Nearest Neighbor.

Προκειμένου να γίνει σύγκριση των δύο μεθόδων επιλογής γονιδίων, για κάθε μία συγκεντρώνονται οι μέγιστες αποδόσεις των ταξινομητών μετά την προεπεξεργασία. Και εδώ λαμβάνονται υπόψη μόνο οι παράμετροι που δίνουν το καλύτερο αποτέλεσμα για κάθε ταξινομητή. Στον Πίνακα 20 παρουσιάζονται τα αποτελέσματα μετά την προεπεξεργασία για κάθε ταξινομητή και για κάθε μεθοδολογία. Φαίνεται για πόσα γονίδια επιτυγχάνεται κάθε φορά η εκάστοτε απόδοση του κάθε ταξινομητή, καθώς και το ποσοστό των γονιδίων που είναι κοινά για το συγκεκριμένο ταξινομητή στις δύο μεθόδους επιλογής γονιδίων (τελευταία στήλη του πίνακα). Επιπλέον παρουσιάζεται το ποσοστό των γονιδίων που είναι κοινά στους τρεις ταξινομητές για μια συγκεκριμένη μεθοδολογία επιλογής γονιδίων (τελευταία σειρά του πίνακα).

Ταξινομητής	**Filter**			**Wrapper**			**Κοινά γονίδια**
	Αριθμός γονιδίων	**SR στο train**	**SR στο test**	**Αριθμός γονιδίων**	**SR στο train**	**SR στο test**	
Nearest Mean	6	100%	88.23%	2	97.37%	94.12%	0%
***k*-Nearest Neighbor**	3	100%	88.23%	1	97.37%	91.18%	33.33%
SVM	5	100%	88.23%	4	100%	97.06%	0%
Κοινά γονίδια	50%			0%			

Πίνακας 20. Λευχαιμία: Σύγκριση filter και wrapper μεθόδων (μετά την προεπεξεργασία)

Όσον αφορά την σύγκριση των δύο μεθοδολογιών έχουμε να παρατηρήσουμε τα εξής:

- Με την wrapper μεθοδολογία επιτυγχάνονται καλύτερες αποδόσεις στο test set απ' ότι με την filter. Στο train set οι NM και *k*-NN αδυνατούν να ταξινομήσουν μόνο ένα δείγμα στα 34.
- Στη filter υπάρχουν κοινά γονίδια ανάμεσα στους ταξινομητές ενώ στην wrapper δεν υπάρχουν (μόνο ένα κοινό γονίδιο ανάμεσα στον Nearest Mean και στον SVM).
- Ανάμεσα στις δύο μεθοδολογίες δεν υπάρχουν κοινά γονίδια μεταξύ των ταξινομητών παρά μόνο στην περίπτωση του *k*-Nearest Neighbor όπου υπάρχει ένα κοινό γονίδιο.

5.3.2 Σύγκριση των αποτελεσμάτων του καρκίνου του μαστού

Προκειμένου να γίνει σύγκριση των αποτελεσμάτων που έχουμε χωρίς και με την προεπεξεργασία, συγκεντρώνονται οι μέγιστες αποδόσεις των ταξινομητών σε κάθε περίπτωση. Λαμβάνονται υπόψη μόνο οι παράμετροι που δίνουν το καλύτερο αποτέλεσμα για κάθε ταξινομητή. Εξετάζεται κάθε μεθοδολογία επιλογής γονιδίων ξεχωριστά.

Ταξινομητής	Χωρίς προεπεξεργασία			Με προεπεξεργασία			Κοινά γονίδια
	Αριθμός γονιδίων	SR στο train	SR στο test	Αριθμός γονιδίων	SR στο train	SR στο test	
Nearest Mean	1024	82.05%	89.47%	2	74.36%	84.21%	0.19%
***k*-Nearest Neighbor**	512	84.61%	89.47%	11	87.18%	94.73%	2.15%
SVM	512	100%	89.47%	183	100%	84.21%	35.74%
Κοινά γονίδια	50%			1.09%			

Πίνακας 21. Καρκίνος του μαστού, filter μέθοδος: Σύγκριση πριν και μετά την προεπεξεργασία

Ταξινομητής	Χωρίς προεπεξεργασία			Με προεπεξεργασία			Κοινά γονίδια
	Αριθμός γονιδίων	SR στο train	SR στο test	Αριθμός γονιδίων	SR στο train	SR στο test	
Nearest Mean	8192	91.02%	84.21%	8	83.33%	84.21%	0.098%
***k*-Nearest Neighbor**	1024	79.49%	84.21%	6	100%	84.21%	0.58%
SVM	8192	96.15%	78.95%	10	98.71%	78.95%	0.12%
Κοινά γονίδια	12.5%			12.5%			

Πίνακας 22. Καρκίνος του μαστού, wrapper μέθοδος: Σύγκριση πριν και μετά την προεπεξεργασία

Στους πίνακες 21 και 22 παρουσιάζονται τα αποτελέσματα για την filter και την wrapper μεθοδολογία αντίστοιχα. Φαίνεται ο αριθμός των γονιδίων για τα οποία επιτυγχάνεται η μέγιστη απόδοση, ενώ στην τελευταία στήλη φαίνεται το ποσοστό των γονιδίων που είναι κοινά πριν και μετά το στάδιο της προεπεξεργασίας. Τέλος στην τελευταία σειρά παρουσιάζεται το ποσοστό των γονιδίων που είναι κοινά στους τρεις ταξινομητές τόσο χωρίς όσο και με την προεπεξεργασία.

Όσον αφορά τη συνεισφορά του σταδίου της προεπεξεργασίας έχουμε να παρατηρήσουμε τα εξής:

- Η προεπεξεργασία βοηθάει στην περίπτωση που έχουμε wrapper μεθοδολογία με μείωση των γονιδίων, ενώ η απόδοση των ταξινομητών παραμένει σταθερή στο test και αυξάνεται στο train εκτός από την περίπτωση του NM όπου μειώνεται, αλλά όχι σημαντικά.
- Στην filter μεθοδολογία η προεπεξεργασία βελτιώνει την απόδοση μόνο στην περίπτωση του *k*-Nearest Neighbor.
- Χωρίς την προεπεξεργασία για την filter μεθοδολογία παρατηρείται πολύ μεγαλύτερη επικάλυψη ανάμεσα στα γονίδια των διαφορετικών ταξινομητών απ' ότι μετά την προεπεξεργασία.
- Στην wrapper μεθοδολογία η επικάλυψη των γονιδίων μεταξύ των ταξινομητών παραμένει ίδια και μετά την προεπεξεργασία.
- Μετά την προεπεξεργασία τα γονίδια που προκύπτουν για κάθε ταξινομητή είναι υποσύνολο αυτών που προκύπτουν χωρίς την προεπεξεργασία, τόσο για filter όσο και για wrapper μεθοδολογία επιλογής γονιδίων.

Ταξινομητής	**Filter**			**Wrapper**			**Κοινά γονίδια**
	Αριθμός γονιδίων	**SR στο train**	**SR στο test**	**Αριθμός γονιδίων**	**SR στο train**	**SR στο test**	
Nearest Mean	2	74.36%	84.21%	8	83.33%	84.21%	11.11%
***k*-Nearest Neighbor**	11	87.18%	94.73%	6	100%	84.21%	6.25%
SVM	183	100%	84.21%	10	98.71%	78.95%	1.05%
Κοινά γονίδια	1.09%			12.5%			

Πίνακας 23. Καρκίνος του μαστού: Σύγκριση filter και wrapper μεθόδων (μετά την προεπεξεργασία)

Προκειμένου να γίνει σύγκριση των δύο μεθόδων επιλογής γονιδίων, για κάθε μία συγκεντρώνονται οι μέγιστες αποδόσεις των ταξινομητών μετά την προεπεξεργασία. Υπολογίζονται μόνο οι παράμετροι που δίνουν το καλύτερο αποτέλεσμα για κάθε ταξινομητή. Στον Πίνακα 23 παρουσιάζονται τα αποτελέσματα μετά την προεπεξεργασία για κάθε ταξινομητή και για κάθε μεθοδολογία. Φαίνεται για πόσα γονίδια επιτυγχάνεται κάθε φορά η εκάστοτε απόδοση του κάθε ταξινομητή, καθώς και το ποσοστό των γονιδίων που είναι κοινά για το συγκεκριμένο ταξινομητή στις δύο μεθόδους επιλογής γονιδίων (τελευταία στήλη του πίνακα). Επιπλέον παρουσιάζεται το ποσοστό των γονιδίων που είναι κοινά στους τρεις ταξινομητές για μια συγκεκριμένη μεθοδολογία επιλογής γονιδίων (τελευταία σειρά του πίνακα).

Όσον αφορά λοιπόν τη σύγκριση των δύο μεθοδολογιών επιλογής γονιδίων έχουμε να παρατηρήσουμε τα εξής:

- Με την filter επιτυγχάνεται καλύτερη απόδοση των ταξινομητών απ' ότι με την wrapper, με μόνη εξαίρεση τον NM όπου έχουμε χαμηλή απόδοση στο train.
- Τόσο στην filter όσο και στην wrapper υπάρχουν κοινά γονίδια ανάμεσα στα γονίδια που προκύπτουν για κάθε ταξινομητή.
- Ανάμεσα στις δύο μεθοδολογίες υπάρχουν κοινά γονίδια για κάθε ταξινομητή (ένα κοινό στις περιπτώσεις του NM και *k*-NN, 2 κοινά στην περίπτωση του SVM). Μάλιστα ένα γονίδιο είναι κοινό συνολικά και στις τρεις περιπτώσεις ταξινομητών.

5.4 Συγκριτικά αποτελέσματα στις δύο παθολογίες

Όπως συζητήθηκε στο προηγούμενο κεφάλαιο κατά την περιγραφή των δεδομένων, τα δεδομένα του καρκίνου του μαστού είναι πιο δύσκολο να αντιμετωπιστούν από εκείνα της λευχαιμίας, λόγω του ότι παρουσιάζουν μεγάλη επικάλυψη μεταξύ των δύο κλάσεων. Είναι επομένως ενδιαφέρον να δούμε συγκριτικά τα αποτελέσματα για τις δύο παθολογίες.

Η απόδοση των ταξινομητών στα αρχικά γονίδια είναι καλύτερη για τα δεδομένα της λευχαιμίας από ότι για αυτά του καρκίνου του μαστού, γεγονός αναμενόμενο από τα συμπεράσματα που καταλήξαμε στο προηγούμενο κεφάλαιο για την φύση των δεδομένων. Μάλιστα στα δεδομένα του καρκίνου του μαστού σε κάποιους ταξινομητές εμφανίζεται το φαινόμενο του over-fitting.

Όσον αφορά την απόδοση των ταξινομητών μετά την προεπεξεργασία παρατηρούμε τα εξής:

- Και στις δύο παθολογίες ο ταξινομητής Nearest Mean έχει σταθερά καλή απόδοση, είτε πρόκειται για wrapper είτε για filter μεθοδολογία επιλογής γονιδίων και οδηγεί σε μικρά σύνολα γονιδίων.
- Ο *k*-Nearest Neighbor δεν έχει καλή απόδοση για πολύ μεγάλες τιμές – σχετικά με το πλήθος των δειγμάτων εκπαίδευσης.
- Για τη λευχαιμία στη filter μέθοδο όσο το k αυξάνεται η απόδοση μειώνεται ενώ στη wrapper μέθοδο έχουμε την καλύτερη απόδοση για k=5.
- Για τον καρκίνο του μαστού στη filter μέθοδο έχουμε καλύτερη απόδοση για μικρά *k* (μικρότερα ή ίσα του 11) με μέγιστη απόδοση για *k*=5. Στη wrapper μέθοδο έχουμε καλύτερη απόδοση για *k* =1 και για ενδιάμεσες τιμές του *k* (11 και 37).
- Και στις δύο παθολογίες ο SVM για *C* = 0.001 παρουσιάζει τη χαμηλότερη απόδοση από τις υπόλοιπες τιμές της παραμέτρου.
- Στη λευχαιμία ο SVM για τις τιμές 1, 100 και 1000 της παραμέτρου *C* έχει την ίδια συμπεριφορά είτε αφορά filter είτε wrapper επιλογή.
- Στον καρκίνο του μαστού στη filter μεθοδολογία ο SVM έχει καλύτερη απόδοση για *C*=100 και 1000 (ίδια συμπεριφορά για αυτές τις δύο τιμές) απ' ότι για *C*=1, ενώ στη wrapper μεθοδολογία η καλύτερη απόδοση παρατηρείται για C=1000 (ίδια συμπεριφορά για τις τιμές 1 και 100).

Με βάση τα συγκριτικά αποτελέσματα μεταξύ filter και wrapper μεθοδολογίας τα οποία σχολιάστηκαν στην ενότητα 5.3, συγκριτικά για τις δύο παθολογίες τα αποτελέσματα φαίνονται στον Πίνακα 24.

	Λευχαιμία	**Καρκίνος του μαστού**
Η προεπεξεργασία βοηθάει	• Τη wrapper με αύξηση στην απόδοση και μείωση των γονιδίων • Τη filter (εκτός SVM) με μείωση των γονιδίων	• Τη wrapper με μείωση των γονιδίων • Τη filter με αύξηση της απόδοσης και μείωση των γονιδίων μόνο για τον *k*-NN
Επικάλυψη γονιδίων για filter μέθοδο	Μεγαλύτερη μετά την προεπεξεργασία	Μεγαλύτερη χωρίς την προεπεξεργασία
Επικάλυψη γονιδίων για wrapper μέθοδο	Μέγιστη χωρίς την προεπεξεργασία, μηδενική μετά την προεπεξεργασία	Ίδια μικρή επικάλυψη πριν και μετά την προεπεξεργασία
Γονίδια μετά την προεπεξεργασία	• Filter: υποσύνολο αυτών πριν την προεπεξεργασία • Wrapper: μόνο ο *k*-NN έχει ένα κοινό γονίδιο με αυτά χωρίς την προεπεξεργασία	Υποσύνολο αυτών χωρίς την προεπεξεργασία τόσο για filter όσο και για wrapper
Σύγκριση αλγορίθμων ως προς την απόδοση των ταξινομητών	Μεγαλύτερη απόδοση για wrapper	Μεγαλύτερη απόδοση για filter
Σύγκριση αλγορίθμων ως προς τα γονίδια για κάθε ταξινομητή	Δεν υπάρχουν κοινά γονίδια παρά μόνο για τον *k*-NN (1 κοινό γονίδιο)	Υπάρχουν κοινά γονίδια (το πολύ δύο) για κάθε ταξινομητή

Πίνακας 24. Συγκριτικά αποτελέσματα για τις δύο παθολογίες

Όσον αφορά την wrapper μεθοδολογία ακολουθείται ένα επιπλέον βήμα τυπικής επιλογής/ταξινόμησης (Σχήμα 15) με αφετηρία την ένωση των συνόλων των γονιδίων που προκύπτουν από κάθε ταξινομητή μετά την επιλογή στα προεπιλεγμένα γονίδια (Πίνακες 7 για τη λευχαιμία και 16 για τον καρκίνο του μαστού). Το αποτέλεσμα είναι οι Πίνακες 8 και 17και οι παρατηρήσεις μας είναι οι εξής:

- Και στις δύο παθολογίες δεν υπολογίζεται στην ένωση των γονιδίων το σύνολο που προκύπτει από τον Nearest Neighbor για *k* πολύ μεγάλο (μικρή απόδοση) όπως επίσης και αυτό που προκύπτει από τον SVM για *C* = 0.001 (πολλά γονίδια)
- Παρατηρείται καλύτερη απόδοση για όλους τους ταξινομητές εκτός των δύο παραπάνω όπου η απόδοση είναι πολύ χαμηλή.

Όσον αφορά την filter μεθοδολογία αυτό το επιπλέον βήμα έγινε μόνο για τα δεδομένα του καρκίνου του μαστού, διότι στα δεδομένα της λευχαιμίας καταλήγουμε στα ίδια αποτελέσματα μετά την προεπεξεργασία (Πίνακας 4). Παρατηρούνται καλύτερα αποτελέσματα (Πίνακας 13) εκτός από τον SVM για *C*=0.001 και τον 71-Nearest Neighbor, ενώ για τους Nearest Mean, Nearest Neighbor για *k*=1, 5 και SVM για *C*=100, 1000 το αποτέλεσμα παραμένει ίδιο.

Στους πίνακες 25 και 26 παρουσιάζονται συγκεντρωτικά τα αποτελέσματα για αυτό το βήμα για την filter και την wrapper μέθοδο αντίστοιχα ώστε να συγκρίνουμε τις δύο παθολογίες. Όσον αφορά την filter μέθοδο για τα δεδομένα της λευχαιμίας όπου δεν έγινε αυτό το επιπλέον βήμα, υπολογίζουμε τα αποτελέσματα που προκύπτουν με επιλογή γονιδίων μετά την προεπεξεργασία. Λαμβάνονται υπόψη μόνο οι καλύτερες αποδόσεις των ταξινομητών ανάμεσα στις διαφορετικές παραμέτρους τους.

Ταξινομητής	**Λευχαιμία**			**Καρκίνος του μαστού**		
	Αριθμός γονιδίων	**SR στο train**	**SR στο test**	**Αριθμός γονιδίων**	**SR στο train**	**SR στο test**
Nearest Mean	6	100%	88.23%	11	76.92%	84.21%
***k*-Nearest Neighbor**	3	100%	88.23%	10	100%	89.47%
SVM	5	100%	88.23%	179	100%	84.21%
Κοινά γονίδια	50%			5.58%		

Πίνακας 25. Filter μέθοδος: Σύγκριση ανάμεσα στις δύο παθολογίες

Ταξινομητής	Λευχαιμία			Καρκίνος του μαστού		
	Αριθμός γονιδίων	**SR στο train**	**SR στο test**	**Αριθμός γονιδίων**	**SR στο train**	**SR στο test**
Nearest Mean	3	100%	94.12%	19	87.18%	89.47%
***k*-Nearest Neighbor**	4	100%	94.12%	26	93.59%	84.21%
SVM	4	100%	97.06%	27	100%	84.21%
Κοινά γονίδια	75%			70.37%		

Πίνακας 26. Wrapper μέθοδος: Σύγκριση ανάμεσα στις δύο παθολογίες

Παρατηρούμε και πάλι ότι καλύτερα αποτελέσματα έχουμε και για τις δύο μεθοδολογίες στην περίπτωση της λευχαιμίας, τόσο στο train όσο και στο test. Γενικά και με τις δύο μεθόδους επιτυγχάνονται ικανοποιητικά ποσοστά και για λίγα γονίδια, εκτός από τον SVM στην filter μέθοδο για τα δεδομένα του καρκίνου του μαστού όπου τα γονίδια είναι πολλά. Επίσης ο Nearest Mean παρουσιάζει χαμηλό ποσοστό επιτυχίας στο train για τον καρκίνο του μαστού.

Επιπλέον στα δεδομένα της λευχαιμίας η wrapper μέθοδος δίνει καλύτερα αποτελέσματα και με μεγαλύτερο ποσοστό επικάλυψης γονιδίων ανάμεσα στους διαφορετικούς ταξινομητές από την filter. Το ίδιο συμβαίνει και για τα δεδομένα του καρκίνου του μαστού, εκτός από την περίπτωση του *k*-Nearest Neighbor όπου έχουμε καλύτερο αποτέλεσμα για την filter μέθοδο.

Γενικά μπορούμε να πούμε ότι στα δεδομένα της λευχαιμίας επιτυγχάνουμε πολύ υψηλά ποσοστά επιτυχίας και για λίγα γονίδια και χωρίς την προεπεξεργασία, με την προεπεξεργασία όμως βελτιώνονται ακόμα περισσότερο. Αυτό συμβαίνει διότι από την αρχή τα δεδομένα της λευχαιμίας παρουσίαζαν ευκολία στον διαχωρισμό μεταξύ των κλάσεων. Στα δεδομένα του καρκίνου του μαστού όμως που εμφάνιζαν δυσκολία στο διαχωρισμό των κλάσεων αρχικά, χωρίς την προεπεξεργασία δεν έχουμε τόσο καλή απόδοση και μάλιστα αυτή επιτυγχάνεται για πολλά γονίδια. Η προεπεξεργασία των δειγμάτων εκπαίδευσης οδηγεί σε καλύτερη απόδοση και για πολύ λιγότερα γονίδια. Μάλιστα ακολουθώντας και το τελευταίο βήμα που αφορά στην επιλογή γονιδίων με αφετηρία την ένωση των συνόλων γονιδίων κάθε ταξινομητή και στις δύο παθολογίες έχουμε πολύ καλό αποτέλεσμα. Πλέον ακόμα και στα δεδομένα του καρκίνου του μαστού έχουμε πολύ ικανοποιητική απόδοση και οι κλάσεις διαχωρίζονται με μεγάλη επιτυχία.

6 Συμπεράσματα – Μελλοντική δουλειά

6.1 Συμπεράσματα

Σε αυτή τη διπλωματική εργασία αντιμετωπίσαμε δύο διαφορετικά είδη καρκίνου, την λευχαιμία και τον καρκίνο του μαστού. Σκοπός ήταν η εύρεση ενός βέλτιστου συνόλου γονιδίων μέσω προεπεξεργασίας των δειγμάτων εκπαίδευσης και μετέπειτα στατιστικής γονιδιακής επιλογής, ώστε να καταλήξουμε σε αποδοτικότερα μοντέλα ταξινόμησης.

Το συμπέρασμα στο οποίο καταλήγουμε είναι ότι η προτεινόμενη προεπεξεργασία αποδίδει στην περίπτωση που έχουμε να κάνουμε με wrapper μεθοδολογία επιλογής γονιδίων. Παρόλα αυτά στην περίπτωση της filter μεθοδολογίας, το στάδιο της προεπεξεργασίας είναι μια διαδικασία η οποία ναι μεν σε ορισμένες περιπτώσεις ταξινομητών οδηγεί σε καλύτερα αποτελέσματα, αλλά εξετάζοντας λίγο καλύτερα παρατηρούμε ότι στο ίδιο αποτέλεσμα θα καταλήγαμε χωρίς την προεπεξεργασία, εάν κατά το τυπικό σχήμα επιλογής/ταξινόμησης από ένα σημείο και μετά μειώναμε τα γονίδια κατά ένα. Τα συμπεράσματα αυτά ισχύουν και για τις δύο εξεταζόμενες παθολογίες.

Σταθερότερη και πολύ ικανοποιητική απόδοση είχε ο ταξινομητής Nearest Mean καθ' όλη τη διάρκεια της υλοποίησης. Η απόδοση του SVM για *C*=0.001 δεν ήταν ιδιαίτερα ικανοποιητική και στις δύο παθολογίες, ενώ επίσης ο *k*-Nearest Neighbor είχε ιδιαίτερα χαμηλή απόδοση για πολύ μεγάλες τιμές της παραμέτρου *k*. Διαφοροποίηση ανάμεσα στην συμπεριφορά του SVM για τις υπόλοιπες τιμές της παραμέτρου *C* (1, 100 και 1000) παρατηρείται μόνο για τα δεδομένα του καρκίνου του μαστού.

Για συγκεκριμένη μεθοδολογία επιλογής γονιδίων, για κάθε ταξινομητή καταλήγουμε σε διαφορετικό σύνολο γονιδίων. Παίρνοντας την ένωση των συνόλων αυτών και προχωρώντας σε γονιδιακή επιλογή, τα αποτελέσματα είναι πολύ ικανοποιητικά για όλους τους ταξινομητές, όσον αφορά τη wrapper μέθοδο επιλογής. Για τη filter μέθοδο μία τέτοια κίνηση στα δεδομένα της λευχαιμίας δεν συνεπάγεται διαφορετικό αποτέλεσμα από αυτό στο οποίο έχουμε ήδη καταλήξει, ενώ στα δεδομένα του καρκίνου του μαστού παρατηρείται κάποια μικρή βελτίωση σε μικρό ποσοστό ταξινομητών.

Μεγαλύτερο ενδιαφέρον από άποψη δυσκολίας διαχωρισμού των δεδομένων των δύο κλάσεων αποτέλεσε ο καρκίνος του μαστού. Όσον αφορά τα δεδομένα της λευχαιμίας ο διαχωρισμός των δεδομένων γινόταν με ευκολία και στα αρχικά γονίδια. Αυτό που τελικά καταφέραμε ήταν η εύρεση πολύ λίγων γονιδίων που οδηγούν σε ακόμα καλύτερη απόδοση των ταξινομητών. Το γεγονός ότι ακόμα και για τα δεδομένα του καρκίνου του μαστού, που αρχικά εμφάνιζαν μεγάλη δυσκολία διαχωρισμού, καταλήγουμε σε σύνολα γονιδίων που οδηγούν σε αποδοτικά μοντέλα ταξινόμησης είναι πολύ ικανοποιητικό. Μπορεί η απόδοση να μην είναι εξαιρετική όπως στα δεδομένα της λευχαιμίας (94% κατά μέσο όρο), είναι όμως πολύ ικανοποιητική (84% κατά μέσο όρο).

6.2 Μελλοντική δουλειά

Η μέχρι τώρα υλοποίηση έγινε με χρήση δύο μεθοδολογιών για επιλογή γονιδίων, μία filter και μία wrapper. Χρησιμοποιήθηκαν το Fisher's Criterion για την filter και η Fisher's Linear Discriminant Analysis για την wrapper. Μία μελλοντική εργασία αφορά στην χρήση της προτεινόμενης προεπεξεργασίας και για άλλα είδη wrapper μεθόδων (εφόσον στην filter μέθοδο δεν είχαμε βελτίωση με την προεπεξεργασία), όπως για παράδειγμα η χρήση των βαρών του ταξινομητή SVM.

Επίσης η συγκεκριμένη μεθοδολογία θα μπορούσε να εφαρμοστεί και σε υβριδικές μεθοδολογίες επιλογής γονιδίων κατά τις οποίες filter μεθοδολογίες συμπεριφέρονται ως wrapper, προκειμένου να ερευνηθεί κατά πόσο το στάδιο της προεπεξεργασίας συνεισφέρει θετικά στην εύρεση μιας βέλτιστης γονιδιακής υπογραφής. Επιπλέον 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.

Μία τελευταία πρόταση για μελλοντική δουλειά θα μπορούσε να είναι η υλοποίηση της μεθόδου σε hardware. Λόγω του ότι η LOOCV μεθοδολογία είναι αρκετά χρονοβόρα, θα ήταν καλό να εξεταστεί εάν μέσω υλοποίησης σε hardware θα είχαμε βελτίωση στον χρόνο εκτέλεσης.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[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] Wikipedia the free encyclopedia, Cancer, http://en.wikipedia.org/wiki/Cancer, accessed September 2008.

[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] Βικιπαίδεια, Κύτταρο, http://el.wikipedia.org/wiki/Κύτταρο, πρόσβαση Σεπτέμβριος 2008.

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

[11] Βικιπαίδεια, Γονιδιακή έκφραση, http://el.wikipedia.org/wiki/Γονιδιακή_έκφραση, πρόσβαση Σεπτέμβριος 2008.

[12] Βικιπαίδεια, Γονίδιο, http://el.wikipedia.org/wiki/Γονίδιο, πρόσβαση Σεπτέμβριος 2008.

[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] Wikipedia the free encyclopedeia, Gene expression profiling, http://en.wikipedia.org/wiki/Expression_profiling, accessed September 2008.

[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] Livepedia.gr, Λευχαιμία, http://www.live-pedia.gr/index.php/Λευχαιμία, accessed September 2008.

[20] R. T. Golub, K. D. Slonim, P. Tamayo, C. Huard, et al., Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, 286 (1999) 531-536.

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

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

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

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

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

[26] Wikipedia the free encyclopedia, Dimension reduction, http://en.wikipedia.org/wiki/Dimension_reduction, accessed September 2008.

[27] Wikipedia the free encyclopedia, Feature selection, http://en.wikipedia.org/wiki/Feature_selection, accessed September 2008.

[28] 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.

[29] Wikipedia the free encyclopedia, Cross-validation, http://en.wikipedia.org/wiki/Cross_validation, accessed September 2008.

[30] Cor J. Veenman, David M.J. Tax, A Weighted Nearest Mean Classifier for Sparse Subspaces, in Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), Vol 2 (2005) pp. 1171-1176.

[31] Wikipedia the free encyclopedia, Support vector machine, http://en.wikipedia.org/wiki/Support_vector_machine , accessed September 2008.

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

[33] Wikipedia the free encyclopedia, Principal components analysis, http://en.wikipedia.org/wiki/Principal_components_analysis, accessed September 2008.

[34] Wikipedia the free encyclopedia, Linear discriminant analysis, http://en.wikipedia.org/wiki/Fisher_linear_discriminant, accessed September 2008.

[35] 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.

[36] *k*-Nearest Neighbors , http://www.statsoft.com/textbook/stknn.html.

[37] Duda R., Hart P., Stork D., Pattern Classification (2nd Edition), *Wiley-Interscience* (2000).

[38] Jang B. Rampal, Microarrays: Volume2: Applications and Data Analysis, 2nd edition (Methods in Molecular Biology), *HUMANA PRESS* (2007).

[39] Steen Knudsen, Cancer Diagnostics with DNA Microarrays, *John Wiley & Sons* (2006).

[40] I. Guyon, An introduction to variable and feature selection, *The Journal of machine learning research*, MIT Press, Vol 3 (2003) pp 1157-1182.

[41] PH Chen, CJ Lin, B. Schölkopf, A tutorial on v-support vector machines: Research Articles, *Applied Stochastic Models in Business and Industry*, John Wiley and Sons Ltd, Vol 21: 2 (2005), pp 111-136.

[42] M. E. Blazadonakis and M. Zervakis, Wrapper Filtering Criteria via Linear Neuron and Kernel Approaches, *Computers in Biology and Medicine*, Pergamon Press, Vol 38: 8, (2008) pp 894-912.

[43] Ferdinand van der Heijden, Robert Duin, Dick de Ridder, David M. J. Tax, Classification, Parameter Estimation and State Estimation: An Engineering Approach using Matlab, *John Wiley & Sons Ltd* (2004).

[44] R.P.W. Duin, P. Juszczak, P. Paclik, E. Pekalska, D. de Ridder, D.M.J. Tax, S. Verzakov, PRTools4.1: A Matlab Toolbox for Pattern Recognition, Delft University of Technology (2007).