Πολυτεχνείο Κρήτης

Τμήμα Ηλεκτρονικών Μηχανικών και Μηχανικών Υπολογιστών

Ανάπτυξη Νευρωνικού Μοντέλου Για Γονιδιακή Ανάλυση

Φοίβος Γύπας

Εξεταστική επιτροπή

Μιχάλης Ζερβάκης (Επιβλέπων)

Γεώργιος Σταυρακάκης

Γεώργιος Καρυστινός

Ιούνιος 2011

Περίληψη

Σκοπός της συγκεκριμένης διπλωματικής εργασίας είναι η μελέτη και υλοποίηση αλγορίθμων γονιδιακής ανάλυσης. Χρησιμοποιώντας ένα data set σχετιζόμενο με καρκίνο του μαστού και με τη χρήση τεσσάρων τεχνικών καταλήξαμε σε τέσσερις γονιδιακές υπογραφές διαφορετικού μεγέθους. Οι μέθοδοι στις οποίες εστιάσαμε είναι η RFE-LNW (Recursive Feature Elimination based on Linear Neuron Weights), SVMs (Support Vector Machines), LASSO (Least Absolute Shrinkage and Selection Operator), FSMLP (Feature Selection Multilayer Perceptron -με gate opening-) και NERFCM (Non-Euclidean Relation Fuzzy c-means). Πιο συγκεκριμένα ο τρόπος εφαρμογής των διάφορων αλγορίθμων που χρησιμοποιήθηκαν για να καταλήξουμε στις τέσσερις υπογραφές φαίνεται παρακάτω:

- Για την πρώτη γονιδιακή υπογραφή εφαρμόσαμε: RFE-LNW και SVM.
- Για τη δεύτερη γονιδιακή υπογραφή εφαρμόσαμε: LASSO και SVM.
- Για την τρίτη γονιδιακή υπογραφή εφαρμόσαμε: RFE-LNW, FSMLP και SVM.
- Για την τέταρτη γονιδιακή υπογραφή εφαρμόσαμε: RFE-LNW, FSMLP, NERFCM και SVM.

Τα γονίδια των υπογραφών στα οποία καταλήξαμε έχουν τόσο στατιστικό όσο και βιολογικό ενδιαφέρον.

Ευχαριστίες Πρώτα από όλα θα ήθελα να ευχαριστήσω τον καθηγητή μου κύριο Μιχάλη Ζερβάκη για τη βοήθεια και την υποστήριξη που μου προσέφερε σε όλη τη διάρκεια της εκπόνησης της διπλωματικής μου εργασίας. Θα ήθελα ακόμα να ευχαριστήσω τον κύριο Μιχάλη Μπλαζαντωνάκη για τη βοήθεια και το χρόνο που μου αφιέρωσε ιδιαίτερα στο ξεκίνημα της εργασίας, καθώς και την κυρία Αικατερίνη Σ. Μπέη για τη βοήθειά της στο βιολογικό κομμάτι της εργασίας. Ευχαριστώ όλους όσους με βοήθησαν στη διάρκεια της φοίτησής μου. Ευχαριστώ όλους τους φίλους μου στα Χανιά για τα χρόνια τα οποία περάσαμε παρέα. Και φυσικά ευχαριστώ πάνω από όλα την οικογένειά μου για τη συνεχή υποστήριξή της σε όλη τη διάρκεια της φοίτησης μου και της ζωής μου.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

Καμία ασθένεια στο παρελθόν δεν έχει προκαλέσει τόσες δυσκολίες στην κατανόηση και αντιμετώπισή της, από τους γιατρούς και τους βιολόγους, όσο ο καρκίνος. Η κατανόηση των μηχανισμών της ασθένειας είναι πολύ σημαντική υπόθεση, καθώς έτσι επιτυγχάνεται καλύτερη πρόγνωση, έγκαιρη διάγνωση και θεραπεία. Οι ασθενείς με καρκίνο οι οποίοι βρίσκονται στο ίδιο στάδιο της ασθένειας μπορεί να έχουν εντελώς διαφορετικές αποκρίσεις σε κάποια θεραπεία αλλά και στο συνολικό αποτέλεσμα. Οι κυριότεροι παράγοντες που προβλέπουν μεταστάσεις, όπως η κατάσταση των λεμφαδένων και ο ιστολογικός βαθμός, αποτυγχάνουν να ταξινομήσουν ακριβώς τους όγκους σύμφωνα με την κλινική συμπεριφορά τους. Η χημειοθεραπεία ή η ορμονοθεραπεία μειώνει τον κίνδυνο μετάστασης κατά ένα τρίτο, παρόλα αυτά το 70-80% των ασθενών που υπόκεινται σε αυτή τη θεραπεία θα επιβίωναν και χωρίς αυτήν. Οι γιατροί έχουν καταλήξει λοιπόν στο συμπέρασμα, ότι είναι πολύ σημαντική η ανάπτυξη της εξατομικευμένης θεραπείας. Για την επίτευξη του παραπάνω στόχου είναι σημαντική η ανάπτυξη νέων τεχνικών από διαφορετικούς τόμεις, όπως του bioengineering και bioinformatics.

Ο κύριος σκοπός του bioengineering είναι η εφαρμογή της γνώσης που προέρχεται από τη μηχανική στους τομείς της ιατρικής και της βιολογίας, ενώ του bioinformatics είναι η εφαρμογή της στατιστικής και τη επιστήμης των υπολογιστών στον τομέας της μοριακής βιολογίας (modecular biology). Έχουμε να κάνουμε λοιπόν με την ανάπτυξη και εξέλιξη βάσεων δεδομένων, αλγορίθμων, υπολογιστικών και στατιστικών τεχνικών και θεωρίας, για την επίλυση τυπικών και πρακτικών προβλημάτων που προκύπτουν από τη διαχείριση και την ανάλυση των βιολογικών δεδομένων.

1.1 Σκοπός

Σκοπός της συγκεκριμένης διπλωματικής εργασίας είναι η μελέτη και υλοποίηση αλγορίθμών γονιδιακής ανάλυσης. Χρησιμοποιώντας ένα data set [1] σχετιζόμενο με καρκίνο του μαστού και με τη χρήση τεσσάρων τεχνικών καταλήξαμε σε τέσσερις γονιδιακές υπογραφές διαφορετικού μεγέθους. Οι μέθοδοι στις οποίες εστιάσαμε είναι η RFE-LNW (Recursive Feature Elimination based on Linear Neuron Weights), SVMs (Support Vector Machines), LASSO (Least Absolute Shrinkage and Selection Operator), FSMLP (Feature Selection Multilayer Perceptron -με gate opening-) και NERFCM (Non-Euclidean Rela-

tion Fuzzy c-means). Πιο συγκεκριμένα ο τρόπος εφαρμογής των διάφορων αλγορίθμων που χρησιμοποιήθηκαν για να καταλήξουμε στις τέσσερις υπογραφές φαίνεται παρακάτω:

- Για την πρώτη γονιδιακή υπογραφή εφαρμόσαμε: RFE-LNW και SVM.
- Για τη δεύτερη γονιδιακή υπογραφή εφαρμόσαμε: LASSO και SVM.
- Για την τρίτη γονιδιακή υπογραφή εφαρμόσαμε: RFE-LNW, FSMLP και SVM.
- Για την τέταρτη γονιδιακή υπογραφή εφαρμόσαμε: RFE-LNW, FSMLP, NERFCM και SVM.

1.2 Μελέτη του γονιδιώματος

Όλα τα έμβυα όντα αποτελούνται από κύτταρα, μικρές μονάδες βιολογικής δραστηριότητας, που περιβάλλονται από μια ημιπερατή μεμβράνη και έχουν την αξιοσημείωτη ικανότητα να αυτοαναπαράγονται σε ένα περιβάλλον, από το οποίο απουσιάζουν άλλα ζωντανά συστήματα [2]. Στις απλούστερες μορφές ζωής, όπως είναι τα μονήρη κύτταρα, ένας καινούργιος οργανισμός προκύπτει από κάθε κυτταρική διαίρεση. Αντίθετα, ένα νέο άτομο στους πολυκύτταρους οργανισμούς δημιουργείται μετά από πολλές κυτταρικές διαιρέσεις [3]. Οι ανώτεροι οργανισμοί όπως ο άνθρωπος, αποτελούνται από ομάδες κυττάρων που αλληλοεπιδρούν μεταξύ τους εξασφαλίζοντας έτσι μια αρμονική συνεργασία και λειτουργία [4].

Το μόριο DNA (δεσοξυριβονουκλεϊκό οξύ), που βρίσκεται μέσα στα 23 ζεύγη χρωμοσωμάτων κάθε κυττάρου, περιέχει ολόκληρο το γενετικό αποτύπωμα των έμβιων όντων. Το σύνολο των μορίων DNA που υπάρχουν σε ένα κύτταρο αποτελούν το γενετικό υλικό του (γονιδίωμα). Εκτός από τα ώριμα ερυθρά αιμοσφαίρια, όλα τα ανθρώπινα κύτταρα περιέχουν ένα πλήρες γονιδίωμα [2], [3].

Τα γονίδια, οι βασικές φυσικές και λειτουργικές μονάδες της κληρονομικότητας, αποτελώντας μόνο ένα κλάσμα του συνόλου του γονιδιώματος, είναι τα τμήματα του DNA που περιέχουν κρίσιμες πληροφορίες για τη σύνθεση των πρωτεϊνών σε ένα συγκεκριμένο τύπο κυττάρων.

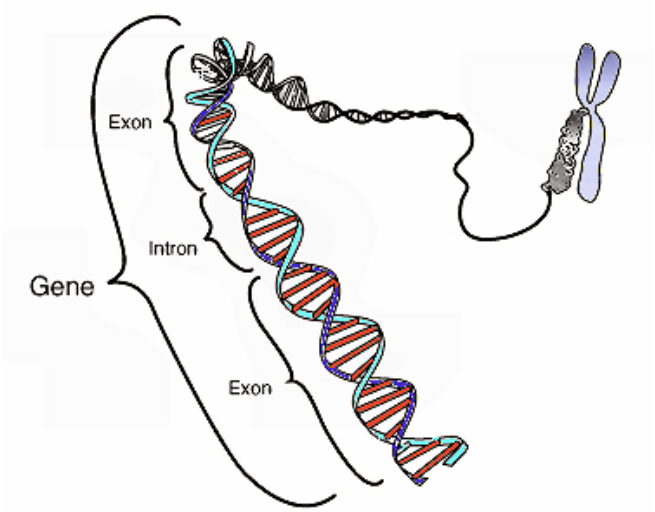
Σήμερα, πιθανολογείται ότι ένα σύνολο 24,500 γονιδίων κωδικεύουν πρωτεΐνες, αριθμός που συρρικνώνεται σε 20,500 γονίδια σύμφωνα με νεότερες μελέτες [5].

Το υπόλοιπο γονιδίωμα αποτελείται από μη κωδικεύουσες περιοχές, των οποίων οι λειτουργίες μπορεί να περιλαμβάνουν την δομική χρωμοσωμική ακεραιότητα και τη ρύθμιση της παραγωγής των πρωτεϊνών.

Η ανακάλυψη ότι το DNA περιέχει τον κώδικα για τη ζωή, έδωσε ώθηση σε μια παγκόσμια προσπάθεια προκειμένου να κατανοηθεί πώς οι αλληλουχίες του γονιδιώματος πολλών οργανισμών σχετίζονται με την υγεία τους. Η μελέτη του ανθρώπινου γονιδιώματος οδήγησε στη γονιδιωματική επανάσταση, δεδομένου ότι η γνωστοποίηση του πρώτου σχεδίου της αλληλουχίας του γονιδιώματος πριν από μια δεκαετία [6] είχε τεράστιες επιπτώσεις στην ανθρώπινη έρευνα για τον καρκίνο.

1.2.1 Γονίδια

Το γονίδιο (εικόνα 1.1) είναι η βασική φυσική μονάδα κληρονομικότητας στους ζωντανούς οργανισμούς, η οποία μεταβιβάζει πληροφορίες από το ένα κύτταρο στο άλλο και κατά επέκταση από τη μια γενιά στην άλλη. Τα γονίδια συνθέτουν το γονιδίωμα ενός οργανισμού, που αποτελείται από το DNA και το RNA, και κατευθύνουν τη φυσική ανάπτυξη και συμπεριφορά του οργανισμού.



Σχήμα 1.1: Απεικόνιση ενός γονιδίου σε σχέση με τη διπλή έλικα του DNA και ένα χρωμόσωμα (δεξιά). Εικόνα: wikipedia.org

Τα περισσότερα γονίδια κωδικοποιούν πρωτεΐνες, οι οποίες είναι βιολογικά μακρομόρια που αποτελούνται από γραμμικές αλυσίδες των αμινοξέων και επηρεάζουν τις περισσότερες από τις χημικές αντιδράσεις που πραγματοποιούνται από τα κύτταρα. Μερικά γονίδια δεν κωδικοποιούν πρωτεΐνες, αλλά τα μόρια του RNA διαδραματίζουν βασικούς ρόλους στην βιοσύνθεση πρωτεϊνών και στον έλεγχο της γονιδιακής έκφρασης. Τα μόρια που προκύπτουν από την γονιδιακή έκφραση, είτε RNA είτε πρωτεΐνη, είναι γνωστά ως γονιδιακά προϊόντα. Τέλος τα περισσότερα γονίδια περιέχουν κάποιες περιοχές που δεν κωδικοποιούν γονιδιακά προϊόντα, αλλά συχνά ρυθμίζουν τη γονιδιακή έκφραση.

1.2.2 Γονιδιακή υπογραφή (Gene Signature)

Η γονιδιακή υπογραφή (gene signature) είναι μια ομάδα από γονίδια σε ένα πυρήνα, των οποίων η από κοινού έκφραση χαρακτηρίζει μοναδικά κάποια πάθηση. Έτσι μπορούμε να χρησιμοποιήσουμε τη gene signature έτσι ώστε να ξεχωρίσουμε ένα group από ασθενείς.

1.2.3 Μικροσυστοιχίες γονιδίων (DNA-Microarray)

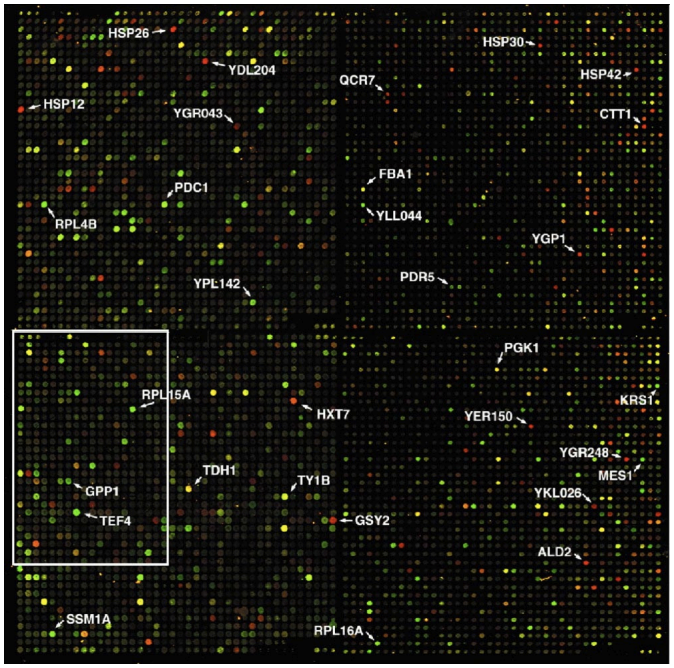
Οι μικροσυστοιχίες γονιδίων αποτελούν μια από τις πιο καινούριες βιοτεχνολογικές μεθόδους για την ανάλυση της έκφρασης των γονιδίων μέσω της διαδικασίας του υβριδισμού.

Εφαρμόζονται ευρέως σε βιομηχανικές, κλινικές, τοξικολογικές και φαρμακευτικές έρευνες καθώς το βασικό πλεονέκτημά τους είναι ότι μετράνε την έκφραση χιλιάδων γονιδίων ταυτόχρονα. Οι DNA μικροσυστοιχίες αποτελούμενες από χιλιάδες γονιδιακές ακολουθίες τυπωμένες σε ένα συμπυκνωμένο πίνακα πάνω σε μια γυάλινη επιφάνεια παρέχουν ένα πρακτικό και οικονομικό εργαλείο για τη μελέτη γονιδιακής έκφρασης σε πολύ μεγάλη κλίμακα.

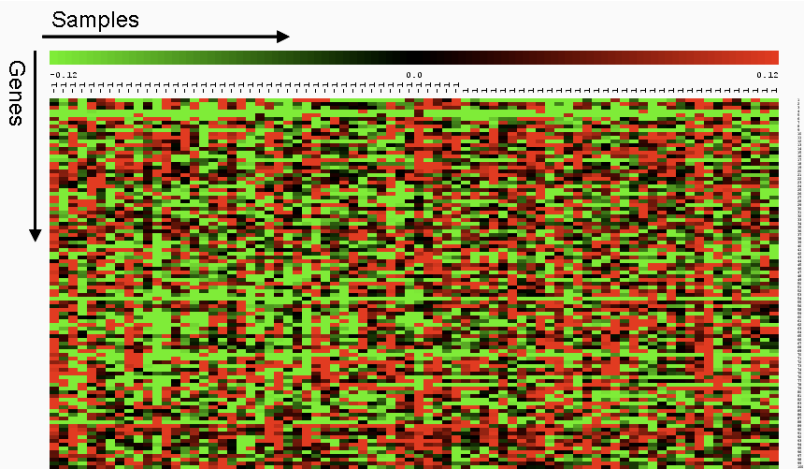
Ένα γονίδιο εκφράζεται όταν αναδεύει μόρια m-RNA. Ένα μυικό κύτταρο περιμένουμε να δημιουργεί m-RNA για διάφορες μυικές πρωτεΐνες όπως η ακτίνη και η μυοσίνη, αλλά όχι για τη χρωστική μελανίνη ή για την ορμόνη ινσουλίνη για παράδειγμα. Μετρώντας τώρα το ποσό του m-RNA που παράγεται από κάθε γονίδιο που εκφράζετα σε ένα δείγμα ιστού, μπορούμε να δημιουργήσουμε ένα expression profile των βιολογικών διεργασιών τα οποία ενεργοποιούνται στο συγκεκριμένο κύτταρο. Στη συνέχεα συγκρίνοντας τα expression profiles μεταξύ διαφόρων ιστών μπορεί να αποκαλύψουμε τους βιολογικούς μηχανισμούς που διαφοροποιούν τα κύτταρα μεταξύ τους. Η τεχνολογία των DNA μικροσυστοιχιών βοηθάει σε αυτή την κατεύθυνση δείχνοντας τις διαφορές των γονιδιακών εκφράσεων ανάμεσα σε δύο τύπους κυττάρων.

Η επιφάνεια της DNA μικροσυστοιχίας διασπάται σε χιλιάδες σημεία, όπου το κάθε σημείο περιλαμβάνει πολλαπλά αντίγραφα μιας μοναδικής DNA ακολουθίας η οποία αντιστοιχεί σε ένα γονίδιο. Με τη χρήση αυτής της τεχνολογίας είναι δυνατή η μέτρηση των διαφορών ανάμεσα σε υγιή και καρκινικά κύτταρα. Ο καρκίνος είναι ουσιαστικά μια ασθένεια των γονιδίων. Πολλά γονίδια ελέγχουν τον τρόπο με τον οποίο τα κύτταρα αναπτύσσονται και πεθαίνουν. Όταν αυτα τα γονίδια σταματάνε να δουλεύουν με φυσιολογικό τρόπο τότε τα κύτταρα μεγαλώνουν με μη φυσιολογικό τρόπο και οδηγούν στο σχημάτισμό όγκου και τελικά καρκίνου. Με σκοπό λοιπόν την πρόγνωση, διάγνωση, κατανόηση και θεραπεία του καρκίνου είναι σημαντική η αναγνώριση αυτων των γονιδίων. Αυτά τα πειράματα βοηθούν τους ερευνητές στη επίτευξη των παραπάνω στόχων και τους επιτρέπει να βρουν τις διαφορές στο επίπεδο έκφρασης μεταξύ καρκινικών και υγιών κυττάρων.

Κάτι τέτοιο επιτυγχάνεται με τη μέτρηση των τύπων m-RNA τα οποία βρίσκονται σε δύο είδη κυττάρου. Για να γίνει κάτι τέτοιο υγιή και καρκινικά δείγματα ιστού συλλέγονται και απομονώνονται από τον ασθενή. Το m-RNA εξάγεται και από τα δύο δείγματα ιστού με τη διάλυσή τους σε ένα μείγμα διαφόρων οργανικών διαλυτών. Τα δείγματα χαακτηρίζονται με διαφορετικό χρώμα, κόκκινο για τα καρκινικούς ιστούς και πράσινο για τα υγιείς ιστούς. Με τη χρήση μιας βιολογικής διαδικασίας κατά την οποία συμπληρωματικά μόρια m-RNA απομονώνονται και υποβαθμίζονται σε c-RNA. Για την ολοκλήρωση του πειράματος έχουμε την υβριδοποίηση. Μέσα από μια τέτοια διαδικασία δύο συμπληρωματικά σκέλη DNA αναμειγνύωντα μεταξύ τους, και σύντομα βρίσκεται μια βάση αντιστοιχίας και συνδυάζεται το ένα με το άλλο. Τα υγιή δείγματα ιστού αποτυπώνονται πάνω στην επιφάνεια του microarray, ενώ στη συνέχεα οι ακολουθίες c-DNA υβριδοποιούν με τις αντίστοιχες αλληλουχίες γονιδίων τη συστοιχία. Η ίδια διαδικασία επαναλαμβάνεται σε διαφορετικό chip για το καρκινικό δείγμα ιστού. Στο τέλος ακτίνα laser εκπέμπεται και στα δύο microarrays κάτι το οποίο κάνει τις υβριδίζουσες περιοχές να λάμψουν.



Σχήμα 1.2: Αποτέλεσμα του DNA micrarray πειράματος ύστερα από το συνδυασμό των δύο εικόνων. Εικόνα: [1]



Σχήμα 1.3: Αποτέλεσμα του DNA micrarray πειράματος με τη χρήση λογαριθμικού (log) μετασχηματισμού. Εικόνα: [1]

Μετά την ολοκλήρωση του πειράματος έχουμε ένα πίνακα με κόκκινα σημεία και ένα πίνακα με πράσινα σημεία. Τοποθετώντας την μία εικόνα πάνω στην άλλη (εικόνα 1.2), μπορούμε να παρατηρήσουμε πράσινα σημεία το οποίο δείχνει ότι τα συγκεκριμένα γονίδια έχουν εκφραστεί περισσότερο στον υγιή ιστό, ακόμη μπορούμε να παρατηρήσουμε κόκκινα σημεία τα οποία δείχνουν το ακριβώς αντίθετο, ενώ μπορούμε παρατηρήσουμε και κίτρινα σημεία το οποίο δείχνει ότι τα συγκεκριμένα γονίδια εκφράστηκαν με τον ίδιο τρόπο και για τις δύο περιπτώσεις, και τέλος μαύρα σημεία για τα οποία τα γονίδια δεν

έχουν εκφραστεί καθόλου. Με τη χρηση λογάριθμου (log) των εντάσεων ανάμεσα στις δύο κύριες κατηγορίες, καταλήγουμε σε μία αριθμητική αναπαράσταση για το επίπεδο έκφρασης κάθε γονιδίου. Αρνητικές τιμές δείχνουν μεγαλύτερη έκφραση σε υγιή ιστό, ενώ θετικές τιμές δείχνουν καρκινικό ιστό. Τιμές κοντά στο μηδέν δείχνουν παρόμοια έκφραση και για τους δύο ιστούς. Κάτι τέτοιο μπορεί να φανεί στο σχήμα 1.3.

1.3 Feature Selection

Στο machine learning και στη στατιστική feature selection είναι μια τεχνική για την επιλογή ενός υποσυνόλου σχετικών χαρακτηριστικών για το χτίσιμο εύρωστων μοντέλων μάθησης. Όταν κάτι τέτοιο εφαρμόζεται στον τομέα της βιολογίας καλείται και διακριτή γονιδιακή επιλογή, και το οποίο αναγνωρίζει ισχυρά γονίδια βασιζόμενο σε πειράματα πάνω σε DNA microarrays. Με την αφαίρεση των μη σχετικών και περιττών χαρακτηριστικών από τα δεδομένα, επιτυγχάνεται βελτίωση στην ταχύτητα των μοντέλων μάθησης καθώς:

- Ελαττώνεται η επίδραση των πολλών διαστάσεων του προβλήματος.
- Ενισχύεται η ικανότητα γενίκευσης.
- Το μοντέλο γίνεται ευκολότερα ερμηνεύσιμο.

Ακόμη το feature selection βοηθάει στην καλύτερη κατανόηση των δεδομένων με το να ξέρουμε ποιά είναι τα σημαντικότερα χαρακτηριστικά και πώς είναι συσχετισμένα το ένα με το άλλο.

Οι μέθοδοι για feature selection χωρίζονται σε δύο μεγάλες κατηγορίες [7]. Στις filter και wrapper μεθόδους. Κάποια σημαντικά χαρακτηριστικά φαίνονται παρακάτω:

- Οι filter μέθοδοι δίνουν έμφαση και επικεντρώνονται στα εγγενή χαρακτηριστικά των δεδομένων, παραμελώντας τις αλληλεπιδράσεις μεταξύ των γονιδίων [8]. Τα γονίδια ταξινομούνται σύμφωνα με τις τιμές που λαμβάνουν διάφορα στοχαστικά μέτρα όπως Fisher's ratio, T-statistics, χ^2 statistic, information gain, Pearson correlation και πολλά άλλα. Τα γονίδια με τις υψηλότερες τιμές είναι αυτά τα οποία παρουσιάζουν και την μεγαλύτερη επίδοση ταξινόμησης, και κατά συνέπεια είναι αυτά τα οποία συγκροτούν κάποια γονιδιακή υπογραφή.
- Οι wrapper μέθοδοι από την άλλη χρησιμοποιούν ένα ταξινομητή για να εκχωρήσουν τιμές ώστε να βαθμολογήσουν τα γονίδια. Αυτοί οι ταξινομητές συνήθως δημιουργούν διανύσματα βαρών τα οποία χρησιμοποιούνται σαν scores για τα γονίδια. Τα γονίδια λοιπόν τα οποία εμφανίζουν το μικρότερο ranking περιορίζονται και η διαδικασία συνεχίζει με επαναληπτικό τρόπο.
- Στις filter μεθόδους το κριτήριο για το feature ranking παραμένει σταθερό στη διαδικασία της γονιδιακής επιλογής, αντίθετα με τις wrapper μεθόδους όπου γίνεται επανεκτίμηση κα δυναμική ανανέωση του κριτηρίου από επανάληψη σε επανάληψη και έτσι ένα μη σημαντικό γονίδιο μπορεί να γίνει σημαντικό σε μια επανάληψη και αντίστροφα.

- Μια ακόμη σημαντική διαφορά μεταξύ των μεθόδων είναι ότι οι filter μέθοδοι επικεντρώνονται σε εγγενή χαρακτηριστικά και παραμελούν τις αλληλεπιδράσεις μεταξύ γονιδίων, ενώ οι wrapper λειτουργούν με τον ακριβώς αντίθετο τρόπο.

Σύμφωνα με όλα τα παραπάνω γίνεται εύκολα κατανοητή η χρησιμότητα του feature selection και ειδικότερα όταν χρησιμοποιείται για γονιδιακή ανάλυση.

1.4 Data Set

Ο καρκίνος του μαστού, η κύρια αιτία θανάτου από καρκίνο στις γυναίκες, χαρακτηρίζεται από μοριακή και κλινική εταιρογένεια. Οι καρκίνοι του μαστού, της μήτρας, του ενδομητρίου και των ωοθηκών αποτελούν το 45% των γυνακείων κακοηθειών και σχετίζονται με περίπου 880000 θανάτους ετησίως [9]. Η κατανόηση της μοριακής και βιολογικής βάσης αυτής της συχνής ασθένειας, καθώς και η κατανόηση των παραγόντων που συμβάλλουν στην εμφάνιση καρκίνου του μαστού είναι πολύ σημαντική υπόθεση και για αυτό το λόγο έχει αυξηθεί η προσπάθεια προς αυτή την κατεύθυνση.

Το Data Set το οποίο χρησιμοποιήσαμε στην παρούσα διπλωματική εργασία προέρχεται από πειράματα που έγιναν πάνω σε ασθενείς με καρκίνο του μαστού [1]. Αποτελείται από ένα Train και από ένα Test set. Το πρώτο αποτελείται από 24481 γονίδια και 78 δείγματα. Τα 44 εξ αυτών θεωρούνται αρνητικά και αντιστοιχούν σε ασθενείς οι οποίοι παρέμειναν απαλλαγμένοι από την ασθένεια για ένα διάστημα τουλάχιστον πέντε ετών. Αντίθετα τα υπόλοιπα 34 δείγματα θεωρούνται θετικά και αντιστοιχούν σε ασθενείς οι οποίοι άνεπτυξαν καποια υποτροπή μέσα σε διάστημα πέντε ετών. 293 από τα αρχικά γονίδια είχαν ελλιπείς πληροφορίες για όλους τους 78 ασθενείς και για αυτό το λόγο αφαιρέθηκαν. Το Test set αποτελείται από 19 ασθενείς σε αρχικό στάδιο (με αρνητικό λεμφαδενικό καρκίνο του μαστού). 7 από αυτούς παρέμειναν απαλλαγμένοι από κάποια μετάσταση για ένα διάστημα πέντε ετών, ενώ οι υπόλοιποι 12 ανέπτυξαν κάποια μετάσταση σε διάστημα πέντε ετών.

1.5 Cross Validation

Το cross-validation [10] είναι μια στατιστική ανάλυση σύμφωνα με την οποία κάνουμε μια αξιολόγηση του πόσο καλά έχουμε εκπαιδεύσει το μοντέλο μας. Χρησιμοποιείτα κυρίως σε περιπτώσεις όπου στόχος μας είναι η πρόβλεψη, και χρειάζεται να γίνει εκτίμηση του πόσο καλά έγινε η εκπαίδευση.

Υπάρχουν τρεις τεχνικές για cross validation.

- Η μέθοδος **holdout** είναι ο πιο απλός τρόπος για cross validation. Το data set χωρίζεται σε δύο set. Το Train και το Test. Η εκτίμηση γίνεται με βάση μια συνάρτηση η οποία έχει επιλεχθεί με βάση το Train set. Στη συνέχεια με τη χρήση της συγκεκριμένης συνάρτησης προσπαθούμε να προβλέψουμε την έξοδο για τα δεδομένα του test set. Το πλεονέκτημα της συγκεκριμένης μεθόδου είναι ότι είναι αρκετά γρήγορη. Το μειονέκτημα είναι ότι μπορεί να έχει μεγάλη διακύμανση ανάλογα με το πώς έχει γίνει ο διαχωρισμός στα δύο set.

- Η μέθοδος **K-fold** cross validation είναι στην ουσία μια βελτιστοποίηση της holdout μεθόδου. Το data set χωρίζεται σε k υποσύνολα και η μέθοδος holdout επαναλαμβάνεται k φορές. Κάθε φορά ένα από τα k υποσύνολα χρησιμοποιείται σαν test set και τα υπόλοιπα $k-1$ σαν train set. Στη συνέχεια υπολογίζεται το μέσο σφάλμα. Το πρόβλημα της συγκεκριμένης μεθόδου είναι ότι είναι k φορές πιο αργή σε σχέση με τη holdout μέθοδο, αλλά από την άλλη επιτυγχάνεται υψηλή ακρίβεια.

- Η μέθοδος **Leave-one-out** είναι μια παραλλαγή του k-fold, με το k να ισούται με N, όπου N ο αριθμός των σημείων δεδομένων του set. Η συνάρτηση εκτίμησης εκπαιδεύεται πάνω σε όλα τα σημεία εκτός από ένα, ενώ η πρόβλεψη γίνεται για αυτό το σημείο. Το υπολογιστικό κόστος σε αυτή τη μέθοδο είναι αρκετά μεγάλο.

Στην παρούσα εργασία χρησιμοποιήσαμε κατά κύριο λόγο τη μέθοδο holdout καθώς από την αρχή είχαμε ένα dataset χωρισμένο σε train και test set. Μόνο σε ένα σημείο έγινε χρήση του k-fold cross validation, καθώς ήταν ένα ενδιάμεσο στάδιο και δε θέλαμε να κάνουμε validate με το test set.

1.6 Ανάπτυξη της εργασίας

Η λογική η οποία ακολουθήσαμε για την εκπόνηση της παρούσας διπλωματικής εργασίας είναι η εξής:
Έγινε ένας συνδυασμός διάφορων αλγορίθμων ώστε να καταλήξουμε σε κάποιες γονιδιακά υπογραφές τις οποίες συγκρίνουμε τόσο στατιστικά όσο και βιολογικά. Στο κεφάλαιο 2 γίνεται περιγραφή των αλγορίθμων που χρησιμοποιήθηκαν. Αρχικά γίνεται ανάλυση των νευρωνικών δικτύων και του παραλαγμένου νευρωνικού δικτύου που υλοποιήσαμε. Στη συνέχεια αναλύονται αλγόριθμοι για Recursive Feature Elimination, Least Absolute Shrinkage and Selection Operator), Fuzzy c-means αλλά γίνεται και η περιγραφή του genotaror μιας μετά-βάσης δεδομένων που συγκεντρώνει και βαθμολογεί συσχετίσεις μεταξύ γονιδίων και ασθενειών από πολλές αξιόπιστες βάσεις δεδομένων από το διαδίκτυο. Στο κεφάλαιο 3 γίνεται συνδυασμός των παραπάνω υπολογιστικών μεθόδων με τέσσερις διαφορετικούς τρόπους ώστε να καταλήξουμε σε τέσσερις γονιδιακές υπογραφές. Ακολουθεί η στατιστική και βιολογική (με χρήση του genotator) σύγκριση των υπογραφών. Στο κεφάλαιο 4 καταλήγουμε σε κάποια συμπεράσματα και γίνονται κάποιες προτάσεις για μελλοντική εργασία. Τέλος στο παράρτημα μπορούν να βρεθούν τα γονίδια της κάθε υπογραφής ανάλογα και η σχέση τους με το genotator.

Η κύρια συνεισφορά της εργασίας είναι ότι συγκρίνει μοντέρνους αλγορίθμους ως προς τη στατιστική τους συμπεριφορά, ενώ γίνεται και σύγκριση των αποτελεσμάτων (υπογραφών) με βάση την υπάχουσα βιβλιογραφία. Τα συμπεράσματα στα οποία καταλήγουμε είναι σημαντικά και ποικίλουν από υπογραφή σε υπογραφή.

Κεφάλαιο 2

Υπολογιστικές Μέθοδοι

2.1 Νευρωνικά Δίκτυα

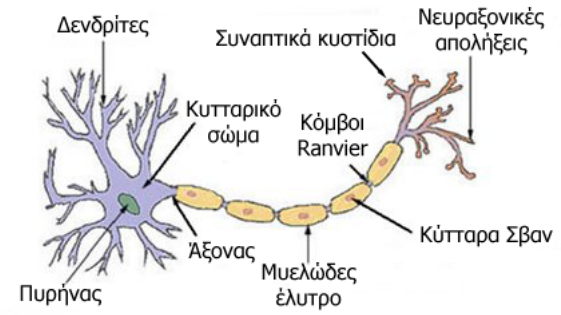
Η έρευνα σχετικά με τα τεχνητά νευρωνικά δίκτυα [11] είναι εμπνευσμένη από τη δομή και τη λειτουργία του εγκεφάλου. Βασικό δομικό στοιχείο του εγκεφάλου είναι οι νευρώνες, δηλαδή τα νευρικά κύτταρα τα οποία δημιουργούν ένα πυκνό δίκτυο επικοινωνίας μεταξύ τους. Κίνητρο για τη μελέτη του νευρώνα και των νευρωνικών δικτύων είναι η ελπίδα ανακάλυψης ενός νέου υπολογιστικού μοντέλου βασισμένου σε μια δικτυακή δομή παρόμοια με αυτή του εγκεφάλου. Αυτή η πλατφόρμα είναι γνωστή ως Connectionist Model με κύριο εκπρόσωπό της τα νευρωνικά δίκτυα.
Τα συνήθη τεχνητά νευρωνικά δίκτυα χρησιμοποιούν πολύ απλοποιημένα μοντέλα νευρώνων τέτοια ώστε να διατηρούν μόνο τα πολύ αδρά χαρακτηριστικά των λεπτομερών μοντέλων που χρησιμοποιούνται στη νευρολογία. Θα έλεγε κανείς ότι τα συνήθη τεχνητά νευρωνικά μοντέλα έχουν ελάχιστη σχέση με τα βιολογικά νευρωνικά συστήματα. Ωστόσο πιστεύεται ότι οι λεπτομέρειες δεν έχουν ιδιαίτερη σημασία στην κατανόηση της ευφυούς συμπεριφοράς των βιολογικών νευρωνικών συστημάτων. Ακόμη και αυτά τα απλά μοντέλα νευρώνων μπορούν να δημιουργήσουν ιδιατέρως ενδιαφέροντα δίκτυα αρκεί να πληρούν δύο βασικά χαρακτηριστικά:

- Οι νευρώνες να έχουν ρυθμιζόμενες παραμέτρους ώστε να διευκολύνεται η διαδικασία της μάθησης.
- Το δίκτυο να αποτελείται από μεγάλο πλήθος νευρώνων ώστε να επιτυγχάνεται παραλληλισμός της επεξεργασίας και κατανομή της πληροφορίας.

Η πρόκληση που αντιμετωπίζει η θεωρία των τεχνητών νευρωνικών δικτύων είναι η εύρεση κατάλληλων αλγορίθμων εκπαίδευσης των δικτύων και ανάκλησης της πληροφορίας που αυτά περιέχουν έτσι ώστε να προορίζονται ευφυείς διαδικασίες όπως αυτές που αναφέρθηκαν παραπάνω. Για την επίτευξη αυτού του στόχου απαιτείται ο ορισμός του κατάλληλου περιβάλλοντος εκπαίδευσης.

2.1.1 Ο Νευρώνας

Το νευρωνικό κύτταρο ή νευρώνας (βλέπε σχήμα 2.1) είναι το βασικό δομικό στοιχείο του εγκεφάλου τόσο στον άνθρωπο όσο και στα ζώα. Ο νευρώνας είναι ένα μεγάλο σε μέγεθος



Σχήμα 2.1: Νευρώνας. Εικόνα: wikipedia.org

κύτταρο το οποίο, ανατομικά, αποτελείται από τα ακόλουθα κύρια τμήματα. Το σώμα, τους δενδρίτες τον άξονα και τις συνάψεις.

Λειτουργικά, τα τμήματα του νευρώνα παίζουν διαφορετικούς ρόλους:

- Οι δενδρίτες είναι οι πύλες εισόδου του νευρώνα. Δέχονται ηλεκτρικά σήματα από άλλους νευρώνες.
- Ο άξονας είναι η πύλη εξόδου του νευρώνα. Μοιάζει με μια μακρόστενη κλωστή που μερικές φορές έχει μήκος μερικά χιλιοστά και άλλες φορές ξεπερνάει το 1 μέτρο. Ο άξονας στέλνει σήματα προς άλλους νευρώνες υπό μορφή ηλεκτρικών παλμών σταθερού πλάτους αλλά μεταβλητής συχνότητας.
- Οι συνάψεις είναι τα σημεία ένωσης μεταξύ διακλαδώσεων του άξονα ενός νευρώνα και των δενδριτών από άλλους νευρώνες. Είναι κύστες με ηλεκτροχημικό υλικό, κυρίως ιόντα Νατρίου και Καλίου. Το υλικό αυτό μεταδίδει την ηλεκτρική δραστηριότητα του άξονα-αποστολέα στους δενδρίτες-παραλήπτες. Το πλάτος της σύναψης, η απόστασή της από τον δενδρίτη και η πυκνότητα του ηλεκτροχημικού υλικούεπηρεάζουν την ευκολία με την οποία η ηλεκτρική δραστηριότητα μεταδίδεται από τον άξονα στον δενδρίτη. Το ποσοστό της ηλεκτρικής δραστηριότητας που μεταδίδεται τελικά στον δενδρίτη λέγεται συναπτικό βάρος. Οι συνάψεις χωρίζονται σε ενισχυτικές (exitatory) και σε ανασταλτικές (inhibitory) ανάλογα με το φορτίο που εκλύεται από τη σύναψη ερεθίζει το νευρώνα προς το να παράγει παλμούς με μεγαλύτερη συχνότητα ή αντίθετα αν τον καταστέλλει εμποδίζοντας τον να παράγει παλμούς.

Πως λειτουργεί ο Νευρώνας

Λειτουργία του βιολογικού νευρώνα

Στους βιολογικούς νευρώνες, φορείς πληροφορίας είναι ηλεκτρικοί παλμοί που ταξιδεύουν στον άξονα κάθε νευρώνα και μέσω των συνάψεων διαδίδονται στους δενδρίτες των παραληπτών νευρώνων. Κάθε νευρώνας Α συλλέγει όλο το ηλεκτρικό φορτίο που δέχεται από κάθε σύναψη στους δενδρίτες του ζυγίζοντας το εισερχόμενο φορτίο με το αντίστοιχο συναπτικό βάρος. Έτσι, όσο πιο ισχυρή είναι η συναπτική ζεύξη τόσο πιο έντοα συμμετέχει το συγκεκριμένο φορτίο εισόδου στο συνολικό άθροισμα. Αν το άθροισμα του

φορτίου ξεπερνάει κάποιο κατώφλι τότε ο άξονας του Α αρχίζει να παράγει παλμούς με μεγάλη συχνότητα οπότε λέμε ότι ο νευρώνας πυροβολεί (fires). Αν όμως το φορτίο δεν περνάει το συγκεκριμένο αυτό όριο τότε ο νευρώνας παράγει πολύ αραιά παλμούς σε τυχαίες στιγμές οπότε λέμε ότι ο νευρώνας είναι αδρανής. Κάθε παλμός έχει συγκεκριμένο χρονικό πλάτος t_p και μετά από κάθε παλμό ο νευρώνας χρειάζεται ένα ελάχιστο χρόνο ανάπαυσης t_r. Έτσι ο μέγιστος ρυθμός των παλμών δε ξεπερνάει το όριο

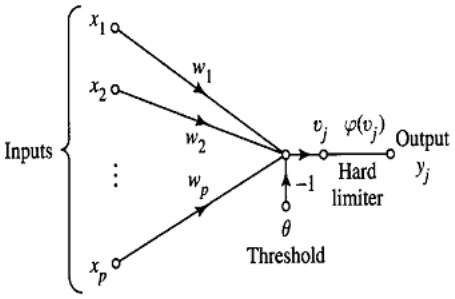
$$FiringFrequency < \frac{1}{(t_p + t_r)}. \tag{2.1}$$

Τελικά οι παλμοί που παράγονται ταξιδεύουν κατά μήκος του άξονα και τροφοδοτούν τους άλλους νευρώνες με τους οποίους συνδέεται ο Α.

Το μοντέλο McCulloth-Pitts

Τη δεκαετία του 1940 υπήρξε μια εντονότατη δραστηριότητα προς την κατεύθυνση της μελέτης των βιολογικών νευρωνικών δικτύων και της μαθηματικής μοντελοποίησης τους. Πρωτοπόροι στον τομέα αυτό οι McCulloch και Pitts που περιέγραψαν ένα απλό μοντέλο της δραστηριότητας του νευρώνα [12]. Η κατάσταση του νευρώνα περιγράφεται από ένα δυαδικό αριθμό y.

- $y = 0 \Rightarrow$ ο νευρώνας είναι αδρανής (δεν πυροβολεί)
- $y = 1 \Rightarrow$ ο νευρώνας πυροβολεί στη μέγιστη ισχύ.



Σχήμα 2.2: Το μοντέλο των McCulloch-Pitts για το νευρώνα. Εικόνα: wikipedia.org

Οι συνάψεις περιγράφονται από τα συναπτικά βάρη (synaptic weights) w_i που είναι πραγματικοί αριθμοί, θετικοί για τις ενισχυτικές συνάψεις και αρνητικοί για τις ανασταλτικές συνάψεις. Αν $x_1, x_2, ..., x_n$ είναι οι είσοδοι του νευρώνα τότε το άθροισμα u του φορτίου που δέχεται ο νευρώνας είναι απλά

$$u = \sum_{i=1}^{n} w_i \cdot x_i \tag{2.2}$$

Αν το άθροισμα u είναι μεγαλύτερο από το κατώφλι (threshold) j τότε ο νευρώνας πυροβολεί, διαφορετικά παραμένει αδρανής. Χρησιμοποιώντας μαθηματικά γράφουμε

$$y = f(u - \theta) \tag{2.3}$$

όπου $f(.)$ είναι η λεγόμενη βηματική συνάρτηση. Στο παραπάνω μοντέλο χρησιμοποιείται η βηματική συνάρτηση 0/1 (step function 0/1).

$$f(u) = \begin{cases} 0 & \text{αν u} \leq 0 \\ 1 & \text{αν u} > 0 \end{cases} \tag{2.4}$$

Σχηματικά το παραπάνω μαθηματικό μοντέλο παριστάνεται από ένα αθροιστή ακολουθούμενο από ένα μη-γραμμικό μετασχηματιστή f (βλέπε σχήμα 2.2).
Το κατώφλι j είναι ένας πραγματικός αριθμός (θετικός ή αρνητικός) όπως άλλωστε και τα συναπτικά βάρη $w_1, ..., w_n$. Κατά αυτή την έννοια το κατώφλι j μπορεί να θεωρηθεί ως ένα επιπλέον συναπτικό βάρος συνδεδεμένο με μια σταθερή είσοδο q_0 η οποία έχει πάντα την τιμή -1. Έτσι θα μπορούσαμε να γράψουμε

$$u = \sum_{i=1}^{n} w_i \cdot x_i - \theta = \sum_{i=0}^{n} w_i \cdot x_i \tag{2.5}$$

όπου $w_0 = \theta$ και $x_0 = -1$.

Άλλα διαδεδομένα μοντέλα

Υπάρχουν πολλές διαφορετικές μοντελοποιήσεις του νευρώνα που αποκλίνουν από το μοντέλο McCulloch-Pitts. Η πιο σημαντική διαφορά είναι στη μορφή της μη γραμμικής συνάρτησης $f()$ που χρησιμοποιείται την έξοδο. Η συνάρτηση αυτή (που καλείται και συνάρτηση ενεργοποίησης του νευρώνα (neuron activation function) μπορεί να πάρει εναλλακτικά τις παρακάτω μορφές:

- Βηματική συνάρτηση -1/1 (step function -1/1)

$$f(u) = \begin{cases} -1 & \text{, αν u} \leq 0 \\ 1 & \text{, αν u} > 0 \end{cases} \tag{2.6}$$

- Σιγμοειδής (sigmoid)

$$f(u) = \frac{1}{1 + e^{-u}} \tag{2.7}$$

  Τη σιγμοειδή συνάρτηση τη χρησιμοπουούμε σε πολύ μεγάλο βαθμό στην παρούσα διπλωματική εργασία.

- Υπερβολική εφαπτόμενη(hyperbolic targent)

$$f(u) = tanh(u) = \frac{(1 - e^{-u})}{(1 + e^{-u})} \tag{2.8}$$

- Συνάρτηση κατωφλίου (threshold function)

$$f(u) = \begin{cases} 0 & \text{, αν u} \leq 0 \\ u & \text{, αν } 0 < u < 1 \\ 1 & \text{, αν u} \geq 1 \end{cases} \tag{2.9}$$

- Συνάρτηση ράμπας (ramp function)

$$f(u) = \begin{cases} 0 & \text{, αν u} \leq 0 \\ u & \text{, αν u} > 0 \end{cases} \tag{2.10}$$

- Γραμμική (linear)

$$f(u) = u \tag{2.11}$$

2.1.2 Εκπαίδευση νευρωνικών δικτύων

Ο ανθρώπινος εγκέφαλος αποτελείται από 100 δισεκατομμύρια νευρώνες και σε κάθε νευρώνα αντιστοιχούν κατά μέσο όρο περίπου 1000 συνάψεις(δηλαδή έχουμε ένα σύνολο 100 τρισεκατομμύριων συνάψεων).
Η τρομερή πολυπλοκότητα του εγκεφάλου τον καθιστά ικανό να εκτελεί με επιτυχία διάφορες λειτουργίες που συλλογικά οδηγούν σε αυτό που αποκαλούμε νοημοσύνη. Τέτοιες λειτουργίες είναι:

- Η αναγνώριση εικόνων
- Η μνήμη
- Η αναγνώριση φωνής, η κατανόηση και η παραγωγή της γλώσσας
- Η αυτόνομη πλοήγηση στο χώρο
- Η λήψη αποφάσεων
- Η κατάστρωση στρατηγικής και η επιλογή της καλύτερης με βάση διαφορά κόστους
- Η λογική, η ανάπτυξη επιχειρημάτων, η συνεπαγωγή
- Η μάθηση και η αυτοπροσαρμογή σε νέο περιβάλλον και σε νέες καταστάσεις

Το τελευταίο αντικείμενο, δηλαδή η μάθηση είναι ίσως ένα από τα πιο σημαντικά χαρακτηριστικά του εγκεφάλου και γενικά των βιολογικών νευρωνικών δικτύων. Ο λόγος που η μάθηση θεωρείται κλειδί της νοημοσύνης είναι το γεγονός ότι οι περισσότερες από τις υπόλοιπες λειτουργίες που περιγράψαμε παραπάνω μαθαίνονται κατά τη διάρκεια του βίου. Παραδείγματα, κανένα παιδί δε γεννήθηκε γνωρίζοντας κάποια γλώσσα, ούτε η αναγνώριση των οικείων προσώπων του και περιβάλλοντος δεν υπάρχει εκ γεννετής αλλά μαθένεται.

Βασική αρχή της Τεχνητής Νοημοσύνης είναι η ύπαρξη ενός υλικού στρώματος πάνω στο οποίο εκτελούνται όλες οι παραπάνω λειτουργίες. Στον άνθρωπο και στα ζώα το υλικό αυτό είναι οι νευρώνες και η δομή του υλικού είναι ένα πυκνό δίκτυο μεταξύ των νευρώνων με εκατοντάδες έως χιλιάδες συνάψεις ανά νευρώνα. Το αντικείμενο μελέτης της τεχνητής μοημοσύνης είναι διπλό:

- Η ανάπτυξη ενός υλικού ο οποίο μπορεί να υποστηρίξει τις παραπάνω επιθυμητές λειτουργίες, άσχετα αν αυτό το υλικό μιμείται τους νευρώνες ή όχι. Στις μέρες μας το υλικό αυτο είναι οι ηλεκτρονικοί υπολογιστές.

- Η ανάπτυξη αλγορίθμων που μιμούνται αυτές τις λειτουργίες, δηλαδή κάνουν αναγνώριση προσώπων και περιβάλλοντος, επιτυγχάνουν αυτόματη πλοήγηση ενός ρομπότ σε περιβάλλον με φυσικά εμπόδια, αναπτύσσουν βέλτιστες στρατηγικές για ένα πρόβλημα, εκτελούν συλλογισμούς και καταλήγουν σε λογικά συμπεράσματα, έχουν μνήμη και αυτοπροσαρμόζονται σε νέες καταστάσεις και σε αγνώριστα περιβάλλοντα και μαθαίνουν από την εμπειρία τους. Όλοι αυτοί οι στόχοι είναι βέβαια πολλοί και ιδιαίτερα απαιτητικοί αλλά αποτελούν το κύριο αντικείμενο μελέτης της Τεχνητής Νοημοσύνης.

Τα τεχνητά νευρωνικά δίκτυα είναι μοντέλα που μιμούνται τη λειτουργία των βιολογικών νευρώνων και τη δομή των βιολογικών νευρωνικών δικτύων. Το αντικείμενο των τεχνητών νευρωνικών δικτύων είναι η ανάπτυξη και η μελέτη αλγορίθμων που μιμούνται την αρχιτεκτονική και το πρότυπο των βιολογικών νευρωνικών δικτύων.

2.1.3 Ταξινόμηση Νευρωνικών Αλγορίθμων

Τα τεχνητά νευρωνικά δίκτυα χωρίζονται στα ακόλουθα δίκτυα

- Εκπαιδευόμενα βάρη
  - { Με επίβλεψη
    - * Perceptron
    - * ADALINE
    - * Back-propagation
    - * Αναδρομικά δίκτυα Back-propagation
    - * Δίκτυα RBF
    - * Μοντέλα SVM
    - * Στοχαστικές μηχανές
  - { Χωρίς επίβλεψη
    - * Συσχετιστικά μοντέλα (Κανόνας του Hebb)
      - · Δίκτυα PCA
      - · Δίκτυα ICA
    - * Ανταγωνιστικά μοντέλα
      - · Δίκτυο Kohonen(SOM)

- Learning VQ
- ART

- Σταθερά βάρη
  - Δίκτυο Hopfield
  - Συσχετιστικές μνήμες
  - Brain State in a Box

2.1.4 Το δίκτυο Perceptron

Το πιο απλό νευρωνικό δίκτυο που μπορεί να σχεδιαστεί και να μελετηθεί είναι ένα δίκτυο που αποτελείται από ένα μόνο νευρώνα. Οι μόνες συνδέσεις που υπάρχουν είναι αυτές μεταξύ των εισόδων $x_1, x_2, ..., x_n$ και του νευρώνα (βλέπε σχήμα 2.3). Στο μοντέλο perceptron ο ένας και μοναδικός νευρώνας υλοποιεί την παρακάτω συνάρτηση μεταφοράς:

$$u = \sum_{i=1}^{n} w_i \cdot x_i - \theta$$

$$y = f(u)$$

η συνάρτηση μεταφοράς απεικονίζει το διάνυσμα εισόδου $x = [x_1, x_2, ..., x_n]^T$ στην έξοδο y. Σε συντομία μπορεί κανείς να γράψει:

$$y = f(\sum_{i=1}^{n} w_i \cdot x_i - \theta)$$

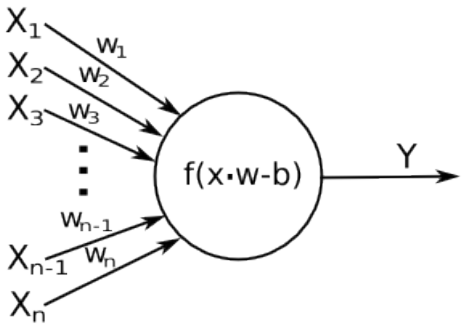
Οι παράμετροι $w_1, w_2, ..., w_n$ είναι τα συναπτικά βάρη του νευρώνα ενώ η παράμετρος θ λέγεται κατώφλι ενεργοποίησης. Ο όρος αυτός εξηγείται από το γεγονός ότι η διέγερση u του νευρώνα η οποία προκύπτει από το περιβάλλον (δηλαδή από τις εισόδους) είναι θετική αν το άθροισμα $\sum_{i=1}^{n} w_i \cdot x_i$ ξεπεράσει το όριο θ:

$$u > 0, \text{αν} \sum_{i=1}^{n} w_i \cdot x_i > \theta$$

$$u = 0, \text{αν} \sum_{i=1}^{n} w_i \cdot x_i = \theta$$

$$u < 0, \text{αν} \sum_{i=1}^{n} w_i \cdot x_i < \theta$$

Με το σκεπτικό αυτό λέμε ότι το θ είναι το κατώφλι πάνω από το οποίο ενεργοποιείται ο νευρώνας.



Σχήμα 2.3: Perceptron νευρωνικό δίκτυο. Εικόνα: wikipedia.org

Η συνάρτηση ενεργοποίησης $f(.)$ τροφοδοτείται από τη διέγερση u και δίνει την έξοδο y του νευρώνα. Η συνάρτηση ενεργοποίησης είναι μη γραμμική και ειδικά στο Perceptron χρησιμοποιείται είτε η βηματική 0/1 είτε η βηματική -1/1. Μικρή σημασία έχει αυτό βέβαια. Οι παράμετροι που ουσιαστικά ρυθμίζουν τη συμπεριφορά του νευρώνα είναι είναι το διάνυσμα των συναπτικών βαρών $w = [w_1, w_2, ..., w_n]^T$ και το κατώφλι θ. Το μοντέλο μπορεί να απλοποιηθεί αν το κατώφλι θ θεωρηθεί ως ένα επιπλέον συναπτικό βάρος(έστω w_0) το οποίο πολλαπλασιάζεται με μια σταθερή είσοδο $x_0 = -1$ οπότε μπορούμε να γράψουμε:

$$u = \sum_{i=1}^{n} w_i \cdot x_i + w_0 \cdot x_0 \tag{2.12}$$

ή πιο απλά, αν ο δείκτης του αθροίσματος i πηγαίνει από 0 έως n:

$$u = \sum_{i=0}^{n} w_i \cdot x_i \tag{2.13}$$

Κατά τον τρόπο αυτό εισάγουμε μια επιπλέον είσοδο x_0 με σταθερή τιμή -1 και άρα είναι σαν να αυξάνουμε τη διάσταση του διανύσματος εισόδου κατά 1. Έτσι η διάσταση του επαυξημένου διανύσματος εισόδου

$$x = [x_0, x_1, ..., x_n]$$

είναι $n + 1$ αντί για n.

Η εκπαίδευση του δικτύου Perceptron

Το ζητούμενο σε ένα νευρωνικό δίκτυο όπως το Perceptron είναι η αυτόματη εκμάθηση των παραμέτρων του συστήματος ώστε να επιτυγχάνεται ο επιθυμητός στόχος. Το δίκτυο εκπαιδεύεται με επίβλεψη, δηλαδή υπάρχει ένας ¨δάσκαλος¨ που μας δίνει την τιμή στόχου $d^{(p)}$ για κάθε πρότυπο εκπαίδευσης p. Το δίκτυο μαθαίνει προσαρμόζοντας τις παραμέτρους $w_0, w_1, ..., w_n$ λαμβάνοντας υπ' όψη του τα επαυξημένα πρότυπα εκπαίδευσης $x^{(1)}, ..., x^{(n)}$ και τους στόχους $d^{(1)}, ..., d^{(p)}$ των προτύπων αυτών χρησιμοποιώντας κάποιον επαναληπτικό αλγόριθμο.

Ο κλασσικός κανόνας εκπαίδευσης Perceptron είναι γνωστός και ως κανόνας σταθερής αύξησης. Ο κανόνας είναι επαναληπτικός: τα πρότυπα παρουσιάζονται στο δίκτυο με κυκλική σειρά και όταν τελειώσουν επαναλαμβάνονται από την αρχή. Ένας πλήρης κύκλος χρήσης όλων των προτύπων ονομάζεται εποχή (epoch).
Ο κανόνας τροποποιεί το επαυξημένο διάνυσμα των συναπτικών βαρών w μόνο όταν υπάρχει σφάλμα ταξινόμησης, δηλαδή όταν ο στόχος $d^{(p)}$ για το πρότυπο p, διαφέρει από την έξοδο του δικτύου

$$y = f(w(k-1)^T \cdot x^{(p)}) \tag{2.14}$$

όπου $w(k-1)$ είναι το επαυξημένο διάνυσμα των συναπτικών βαρών μετά από την επανάληψη $k-1$. Όταν υπάρχει σφάλμα τότε η διόρθωση των βαρών γίνεται προσθέτοντας ή αφαιρώντας ένα ποσοστό του προτύπου $x^{(p)}$. Συγκεκριμένα, αν κατά την επανάληψη k, εισάγεται το πρότυπο p τότε ο κανόνας της διόρθωσης είναι:

$$w(k) = w(k-1) + \beta \cdot (d^{(p)} - y) \cdot x^{(p)} \tag{2.15}$$

όπου $w(k)$ είναι το επαυξημένο διάνυσμα των συναπτικών βαρών μετά την επανάληψη k. Αν θυμηθούμε ότι τα επαυξημένα διανύσματα βαρών και εισόδου ορίζονται ως $w^T = [w_0, w^T]$ και $x^T = [1, x^T]$ (δεχόμενου ότι $x_0 = 1$), τότε ο παραπάνω κανόνας εκπαίδευσης γράφεται ενναλακτικά ως:

$$w(k) = w(k-1) + \beta \cdot (d^{(p)} - y) \cdot x^{(p)} \tag{2.16}$$

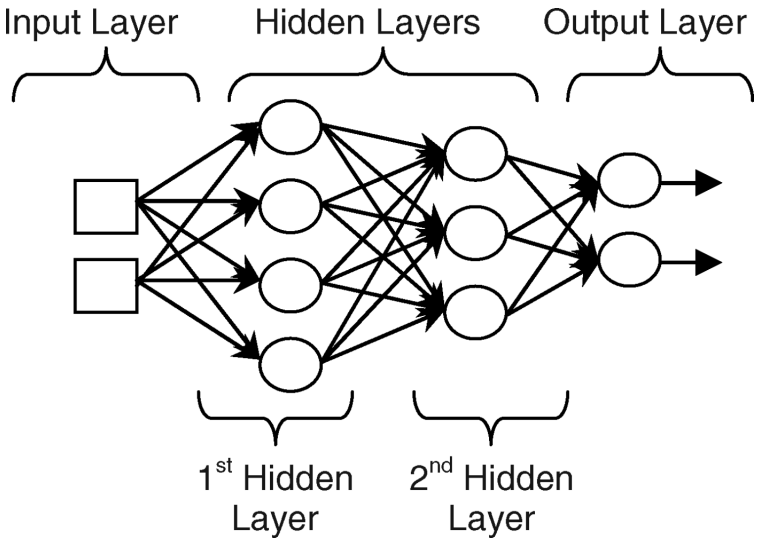
$$w_0(k) = w_0(k-1) + \beta \cdot (d^{(p)} - y) \tag{2.17}$$

Η παράμετρος β βυθμίζει το μέγεθος της διόρθωσης και καλείται βήμα εκπαίδευσης ή ρυθμός εκπαίδευσης. Το β είναι μικρός θετικός αριθμός. Αποδεικνύεται ότι η εκπαίδευση του w γίνεται με τρόπο τέτοιο ώστε το πρότυπο που ταξινομήθηκε τώρα εσφαλμένα, την επόμενη φορά είτε θα ταξινομηθεί σωστά είτε θα πλησιάζει περισσότερο στο να ταξινομηθεί σωστά.
Συμπεραίνουμε λοιπόν ότι το μοντέλο Perceptron εφοδιασμένο με τον παραπάνω αλγόριθμο εκπαίδευσης συγκλίνει σε μία λύση η οποία ταξινομεί σωστά όλα τα πρότυπα αρκεί να υπάρχει μια τέτοια λύση, να είναι δηλαδή το πρόβλημα γραμμικά διαχωρίσιμο. Σε μη γραμμικά διαχωρίσιμα προβλήματα ο αλγόριθμος Perceptron δεν συγκλίνει ποτέ.

2.1.5 Το δίκτυο MLP (Multi-Layer Perceptron)

Όπως είδαμε οι δυνατότητες αναπαράστασης διαχωριστικών επιφανειών είναι περιορισμένες στο δίκτυο Perceptron καθώς με ένα μόνο νευρώνα το δίκτυο μπορεί να αναπαραστήσει μόνο επίπεδες επιφάνειες. Ο περιορισμός αυτός αίρειται με τη χρήση περισσότερων νευρώνων.

Τέτοια δίκτυα όπως αυτό του σχήματος 2.4 καλούνται δίκτυα Perceptron πολλών στρωμάτων (Multi-Layer Perceptron MLP). Το χαρακτηριστικό των δικτύων είναι οι νευρώνες του οποιαδήποτε στρώματος τροφοδοτούν αποκλειστικά τους νευρώνες του επόμενου στρώματος και τροφοδοτούνται αποκλειστικά από τους νευρώνες του προηγούμενου στρώματος.

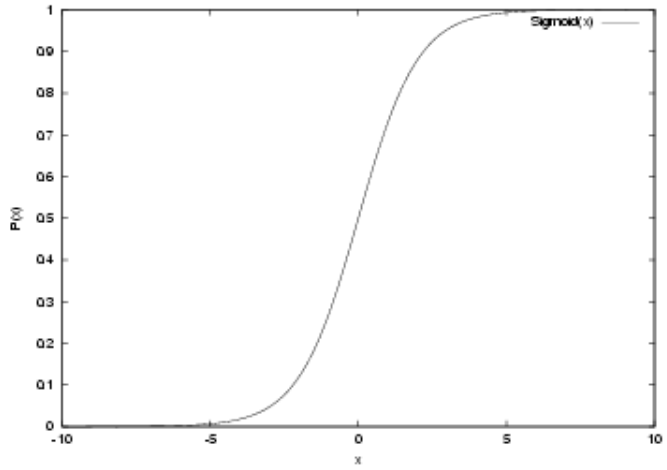


Σχήμα 2.4: Multilayer Perceptron.

Στα δίκτυα πολλών στρωμάτων MLP δε συνιθίζεται οι χρήση της βηματικής συνάρτησης 0/1 ή -1/1 αλλά προτιμάται συνήθως η χρήση της βηματικής συνάρτησης. Ο σημαντικότερος λόγος για αυτή την επιλογή είναι ότι οι μέθοδοι βελτιστοποίησης χρησιμοποιούν παραγώγους, ενώ η βηματική συνάρτηση δεν είναι παραγωγίσιμη. Αυτή η τεχνική δυσκολία ξεπερνιέται με τη χρήση της σιγμοειδούς συνάρτησης:

$$f(u) = \frac{1}{1 + e^{-u}} \tag{2.18}$$

Η συνάρτηση αυτή είναι παραγωγίσιμη (βλέπε εικόνα 2.5) και πρακτικά μοιάζει πολύ με τη βηματική 0/1.



Σχήμα 2.5: Σιγμοειδής συνάρτηση

Η χρήση μαλακών συναρτήσεων κατωφλίωσης όπως η σιγμοειδής δημιουργεί ομαλές επιφάνειες χωρίς απότομες μεταβολές στην τιμή y της εξόδου του δικτύου.

Feed-Forward

Χαρακτηριστικό σε ένα δίκτυο Perceptron πολλών στρωμάτων είναι ότι κάθε στρώμα (l) τροφοδοτείται μόνο από τους νευρώνες του προηγούμενου στρώματος ($l-1$). Έστω λοιπόν το στρώμα εισόδου το οποίο ονομάζουμε μηδενικό. L το πλήθος των στρωμάτων του δικτύου εκτός από το στρώμα εισόδου. $N(0), N(1), ..., N(L)$ το πλήθος των νευρώνων των στρωμάτων $0, 1, ..., L$. Άρα $N(0)$ το πλήθος των εισόδων και $N(L)$ το πλήθος των εξόδων. $a_i(l)$ είναι οι ενεργοποιήσεις των νευρώνων του στρώματος l, $w_{ij}(l)$ το συναπτικό βάρος που συνδέει το νευρώνα $a_j(l-1)$ του στρώματος $l-1$ με το νευρώνα $a_i(l)$ του στρώματος l. $w_{i0}(l)$ είναι το κατώφλι του νευρώνα $a_i(l)$ του στρώματος l. Τέλος $x_i = a_i(0)$ είναι οι είσοδοι του δικτύου και $y_i = a_i(L)$ οι έξοδοι του δικτύου.

Το feedforward στάδιο είναι η διαδικασία υπολογισμού των τιμών όλων των νευρώνων του δικτύου με δεδομένες τιμές των εισόδων $x_1, ..., x_{N(0)}$. Έτσι οι ενεργοποιήσεις των νευρώνων για οποιοδήποτε στρώμα l δίνονται από την σχέση:

$$a_i(l) = f\left(\sum_{j=1}^{N\cdot(l-1)} w_{ij}(l)\cdot a_j(l-1) + w_{i0}(l)\right) \tag{2.19}$$

Αυτό που κάνουμε λοιπόν σε αυτό το στάδιο είναι: Μας δίνονται οι τιμές x_i των εισόδων του δικτύου. Με βάση τις εισόδους υπολογίζουμε πρώτα τις ενεργοποιήσεις του στρώματος 1, στη συνέχεια του στρώματος 2 και σενεχίζουμε με αυτό τον τρόπο μέχρι να φτάσουμε στο y.

Εκπαίδευση δικτύων MLP με Back-Propagation

Η εκπαίδευση ενός MLP είναι η διαδικασία ρύθμισης των συναπτικών βαρών του έτσι ώστε να ικανοποιείται κάποιο κριτήριο καταλληλότητας. Αυτός είναι και ο στόχος της εκπαίδευσης σε οποιοδήποτε νευρωνικό δίκτυο. Αν έχουμε λοιπόν ένα νευρωνικό δίκτυο (MLP) με κατάλληλο μέγεθος μπορούμε να το εκπαιδεύσουμε να μάθει οποιαδήποτε συνάρτηση εμείς επιθυμούμε με οποιαδήποτε ποιότητα προσέγγισης εμείς επιθυμούμε. Αυτό δικαιολογεί τη δημοτικότητα των αλγορίθμων εκπαίδευσης MLP με κύριο εκπρόσωπο τον αλγόριθμο Back-Propagation.

Ο αλγόριθμος Back-Propagation

Το μοντέλο ανήκει στην κατηγορία των δικτύων που εκπαιδεύονται με επίβλεψη. Βασικό χαρακτηριστικό της μεθόδου είναι η ύπαρξη στόχων. Γενικά θεωρούμε σε ένα δίκτυο MLP με L στρώματα, $n = N(0)$ εισόδους και $m = N(L)$ εξόδους. Όπως και σε προηγούμενη ενότητα έτσι και εδώ ονομάζουμε τις εισόδους $x_1, x_2, ..., x_n$ και εξόδους $y_1, y_2, ..., y_n$. Πιο συγκεκριμένα για μια σειρά από P ζευγάρια διανυσμάτων εισόδων - στόχων έχουμε:

- $x^{(p)} = [x_1(p), ..., x_n(p)]^T$, το p-οστό διάνυσμα εισόδου
- $y^{(p)} = [y_1(p), ..., y_n(p)]^T$, το p-οστό διάνυσμα εξόδου
- $d^{(p)} = [d_1(p), ..., d_n(p)]^T$, το p-οστό διάνυσμα στόχων

Τα δεδομένα που απαιτούνται για να εκπαιδεύσουμε το νευρωνικό μας δίκτυο είναι τα P ζευγάρια διανυσμάτων εισόδων - στόχων. Θα ήταν ιδανικό να πετύχουμε τέλεια ταύτιση των εξόδων y με τις επιθυμητές εξόδους d. Κάτι τέτοιο δυστυχώς είναι σχεδόν πάντα αδύνατο οπότε επιθυμούμε τη βέλτιστη προσέγγιση της επιθυμητής κατάστασης χρησιμοποιώντας ένα κριτήριο κόστους. Ένα πολύ κλασσικό κριτήριο κόστους που χρησιμοποιείται ευρύτατα είναι το μέσο τετραγωνικό σφάλμα

$$J = \frac{1}{P} \cdot \sum_{p=1}^{P} ||d^{(p)} - y^{(p)}||^{(2)} = \frac{1}{P} \cdot \sum_{p=1}^{P} \sum_{i=1}^{m} \left[d_i^{(p)} - y_i^{(p)} \right]^2 \tag{2.20}$$

Έχει το πλεονέκτημα ότι η ελαχιστοποίησή του σημαίνει την ελαχιστοποίηση της τετραγωνικής απόστασης μεταξύ των διανυσμάτων $y^{(i)}$, $d^{(i)}$ και επιπλέον παραγωγίζεται εύκολα οπότε μπορεί να χρησιμοποιηθεί πολύ εύκολα σε μεθόδους gradient descent και να δώσει συμπαγείς μαθηματικές λύσεις.

Gradient Descent

Τα συναπτικά βάρη w_{ij} είναι οι παράμετροι οι οποίοι πρέπει να διορθωθούν ώστε το σφάλμα J να ελαχιστοποιηθεί, καθώς τόσο οι είσοδοι $x^{(p)}$ όσο και οι στόχοι $d^{(p)}$ είναι δεδομένοι και σταθεροί. Σύμφωνα με τη μέθοδο gradient descent η μεταβολή το w_{ij} ως προς το χρόνο t γίνεται με τη χρήση της παραγώγου του J ως προς w_{ij}.

$$\frac{dw_{ij}}{dt} = -\frac{\partial J}{\partial w_{ij}} \tag{2.21}$$

και η οποία ονομάζεται κλίση του J ως προς w_{ij}. Επειδή θέλουμε να προσθέσουμε και την έννοια του χρόνου στο συναπτικό βάρος w_{ij}, θα αναφερόμαστε στο τελευταίο ως $w_{ij}(l,k)$ που συνδέει το νευρώνα j του στρώματος $l-1$ με το νευρώνα i του στρώματος l κατά τη χρονική στιγμή k. Οπότε η μέθοδος gradient descent μετατρέπεται στη διακριτή εξίσωση:

$$w_{ij}(l,k+1) - w_{ij}(l,k) = -\beta \frac{\partial J}{\partial w_{ij}(l,k)} \tag{2.22}$$

Όπως και στο απλό Perceptron έτσι και εδώ εμφανίζεται μία παράμετρος βήματος εκπαίδευσης β η οποία είναι ένας μικρός θετικός αριθμός.

Η έξοδος του νευρώνα i του στρώματος l δίνεται από τη σχέση:

$$\alpha_i^{(k)}(l) = f(u_i^{(k)}(l)) \tag{2.23}$$

όπου

$$u_i^{(k)}(l) = \sum_{\xi=1}^{N(l-1)} w_{i\xi}(l,k)\alpha_\xi^{(k)}(l-1) + w_{i0}(l,k) \tag{2.24}$$

είναι η δικτυακή διέγερση του νευρώνα (net-input) και είναι το άθροισμα των διεγέρσεων των νευρώνων του προηγούμενου στρώματος με τα συναπτικά βάρη $w_{i\xi}(l,k)$. Για διευκόλυνσή μας θα ονομάσουμε

$$\delta_i^{(k)}(l) = -\frac{\partial J}{\partial u_i^{(k)}(l)} \tag{2.25}$$

την παράγωγο του κόστους ως προς τη δικτυακή διέγερση του νευρώνα. Συνδυάζοντας λοιπόν τις παραπάνω γνώσεις μπορούμε να καταλήξουμε στο συμπέρασμα ότι

$$\frac{\partial J}{\partial w_{ij}(l,k)} = -\delta_i^{(k)}(l)\alpha_j^{(k)}(l-1) \tag{2.26}$$

για $j = 0, 1, 2, ..., N(l-1)$ και $l = 1, 2, ..., L$.

Τώρα θα υπολογίσουμε τα σφάλματα $\delta_i^{(k)}$ για όλους τους νευρώνες όλων των στρωμάτων.

Για το στρώμα εξόδου η παράγωγος του κόστους J ως προς τη διέγερση $u_i^{(k)}(L)$ είναι:

$$\delta_i^{(k)}(L) = (d_i^{(k)} - y_i^{(k)})f'(u_i^{(k)}(L)) \tag{2.27}$$

όπου f' η παράγωγος της συνάρτησης ενεργοποίησης. Στην περίπτωση που η συνάρτηση ενεργοποίησης είναι η σιγμοειδής, δηλαδή $f(u) = \frac{1}{1+e^{-u}}$, η παράγωγός της είναι $f'(u) = f(u)(1 - f(u))$, ενώ αν είναι η γραμμική, δηλαδή $f(u) = u$, η παράγωγός της είναι $f'(u) = 1$.

Για οποιοδήποτε άλλο στρώμα $l = 1, ..., L-1$, εκτός δηλαδή από αυτό της εξόδου έχουμε:

$$\delta_i^{(k)}(l) = \sum_{\mu=1}^{N(l-1)} \delta_\mu^{(k)}(l+1)w_{\mu i}^{(k)}(l+1)f'(u_i^{(k)}(l)) \tag{2.28}$$

Για την παράγωγο f' ισχύουν τα ίδια όπως παραπάνω.

Γενικά λοιπόν ο αλγόριθμος Back-Propagation φαίνεται στον πίνακα 2.1.

Για την εξίσωση ενημέρωσης βαρών λοιπόν χρησιμοποιούμε την εξίσωση:

$$w_{ij}(l,k+1) = w_{ij}(l,k) + \beta\delta_i^{(k)}(l)\alpha_j^{(k)}(l-1), j = 0, 1, ..., N(l), l = 1, ..., L \tag{2.29}$$

Ενώ οι τύποι υπολογισμού του σφάλματος δ είναι:
Για το τελευταίο στρώμα

$$\delta_i^{(k)}(L) = f'(u_i^{(k)}(L))(d_i^{(k)} - y_i^{(k)}) \tag{2.30}$$

Για τα στρώματα $l = 1, ..., L-1$

$$\delta_i^{(k)}(l) = f'(u_i^{(k)}(l)) \sum_{\mu=1}^{N(l+1)} w_{\mu i}(l+1)\delta_\mu(l+1) \tag{2.31}$$

2.1.6 Το FSMLP νευρωνικό

Η ακριβής πρόβλεψη της κατηγορίας στην οποία ανήκουν δεδομένα κάποιας γονιδιακής έκφρασης είναι πολύ σημαντική καθώς μπορεί να βοηθήσει στην καλύτερη θεραπεία των ασθενών. Δυστυχώς οι υψηλές διαστάσεις των microarray data σε συνδυασμό με το θόρυβο ο οποίος εμπεριέχεται, κάνει αυτή την πρόβλεψη αρκετά δύσκολη. Με τη χρήση ενός παραλλαγμένου γραμμικού feature selection multilayer perceptron νευρωνικού δικτύου είναι δυνατόν να κάνουμε μια καλή πρόβλεψη για το πρόβλημά μας.

Table 2.1: Back-Propagation algorithm

{Είδοδοι: P ζεύγη διανυσμάτων εισόδων-στόχων$(x^{(p)}, d^{(p)})$ και βάρη $w_{ij}(l)$}
$n = 1$·
repeat
  {Για κάθε πρότυπο p}
  for $p = 1 : P$ **do**
    {Feed Forward}
    {Υπολογισμός των εξόδων του στρώματος 1}
    {...}
    {Υπολογισμός των εξόδων του στρώματος L}
    {Back Propagation}
    {Υπολογισμός σφάλματος $\delta_i(L)$ του στρώματος L}
    $\delta_i^{(k)}(L) = f'(u_i^{(k)}(L))(d_i^{(k)} - y_i^{(k)})$
    {...}
    {Υπολογισμός σφάλματος $\delta_i(1)$ του στρώματος 1}
    $\delta_i^{(k)}(1) = f'(u_i^{(k)}(1)) \sum_{\mu=1}^{2N} w_{\mu i}(l+1)\delta_\mu(l+1)$
    {Ενημέρωση βαρών}
    $w_{ij}(l, k+1) = w_{ij}(l, k) + \beta\delta_i^{(k)}(l)\alpha_j^{(k)}(l-1)$
  end for
  $n = n + 1$;
until Το συνολικό σφάλμα J σε μια εποχή να είναι μικρότερο από κάποιο κατώφλι

FSMLP με gate opening

Σε ένα multilayer perceptron δίκτυο η επίδραση κάποιων γονιδίων μπορεί να εξαλειφθεί με το να μην τους επιτρέπεται η είσοδος στο δίκτυο. Κάτι τέτοιο επιτυγχάνεται με τον εξοπλισμό των κόμβων εισόδου με μια πύλη, και το άνοιγμα ή το κλείσιμο αυτής της πύλης.

Όταν κάποιο γονίδιο θεωρείται σημαντικό τότε οι πύλες οι οποίες σχετίζονται με αυτό πρέπει να είναι τελειώς ανοιχτές, ενώ για μερικώς σημαντικά γονίδια οι πύλες πρέπει να είναι μερικώς ανοιχτές. Οι Pal και Chintalapudi [13] πρότειναν ένα τέτοιο μηχανισμό ώστε τα μη σημαντικά γονίδια να αναγνωριστούν αρχικά και να ελαττωθούν στη συνέχεια. Για να μοντελοποιήσουμε τις πύλες συνδέουμε μια συνάρτηση πύλης (gate function) σε κάθε κόμβο του στρώματος εισόδου (input layer). Κάθε κόμβος υπολογίζει το γινόμενο της εισόδου (τιμή του γονιδίου) με τη συνάρτηση πύλης. Το αποτέλεσμα περνάει στα επόμενα επίπεδα του δικτύου. Η συνάρτηση πύλης θα πρέπει να παράγει υψηλές τιμές για τα γονίδια τα οποία είναι σημαντικά,ενώ για τα περιττά οι τιμές πρέπει να είναι κοντά στο μηδέν. Οι συναρτήσεις πύλης εξασθενούν (attenuate) τα γονίδια πριν διαδοθούν στο δίκτυο και έτσι καλούμε αυτές τις συναρτήσεις ως attenuating functions. Σαν attenuation function, μπορούμε να χρησιμοποιήσουμε οποιαδήποτε συνάρτηση $F_i : \mathbb{R} \to [0, 1]$, η οποία ικανοποιεί τα ακόλουθα κριτήρια:

- Έχει μία συντονίσιμη (tunable) παράμετρο και είναι παραγωγίσιμη με βάση αυτή την παράμετρο.

- Είναι μονότονη με βάση την παραπάνω παράμετρο.

Η σιγμοειδής συνάρτηση $F(m) = \frac{1}{1+e^{-m}}$, στην οποία έχουμε αναφερθεί εκτενέσταρα παραπάνω, είναι μια τέτοια συνάρτηση η οποία ικανοποιεί τα κριτήρια μας.

Όπως προαναφέραμε η φιλοσοφία εκπαίδευσης είναι να κρατάμε τις πύλες κλειστές στην αρχή της εκπαίδευσης και να τις ανοίγουμε σταδιακά στη διάρκεια της εκπαίδευσης. Έστω λοιπόν F_i η συνάρτηση attenuation η οποία σχετίζεται με το $i^{\text{στό}}$ γονίδιο εισόδου. Η F_i έχει ένα argument m_i. $F_i'(m_i)$ είναι η παράγωγος της συνάρτησης στο m_i. Ακόμα μ είναι ο βαθμός εκπαίδευσης (learning rate) της παραμέτρου εξασθένησης, ενώ ν ο βαθμός εκπαίδευσης των συνδεδεμένων βαρών. Το x_i είναι η $i^{\text{στη}}$ είσοδος, ενώ x_i' είναι η τιμή της εξόδου του πρώτου επιπέδου για το i γονίδιο, δηλαδή $x_i' = x_i F(m_i)$. Το w_{ij}^0 είναι το βάρος το οποίο συνδέει το $j^{\text{στο}}$ κόμβο του πρώτου κρυφού επιπέδου με τον $i^{\text{στο}}$ κόμβο του επιπέδου εισόδου. Τέλος το δ_j^1 είναι το σφάλμα του $j^{\text{στου}}$ κόμβου του πρώτου κρυφού επιπέδου.

Πέρα από το w_{ij}^0 η ανανέωση των βαρών σε όλα τα άλλα στάδια είναι παρόμοια με αυτή ενός κανονικού MLP νευρωνικού δικτύου εκπαιδευμένο με backpropagation. Υποθέτοντας ότι το πρώτο κρυφό επίπεδο έχει q κόμβους οι κανόνες ανανέωσης για w_{ij}^0 και m_i θα είναι αντίστοιχα:

$$w_{ij,new}^0 = w_{ij,old}^0 - \nu x_i \delta_j^1 F(m_i) \tag{2.32}$$

$$m_{i,new}^0 = m_{i,old}^0 - \mu x_i \sum_{j=1}^{q} w_{ij}^0 \delta_j^1 F'(m_i) \tag{2.33}$$

Οι παράμετροι των πυλών για όλα τα γονίδια αρχικοποιούνται με τέτοιο τρόπο έτσι ώστε όταν αρχίσει η εκπαίδευση το $F(m)$ να είναι μηδέν για όλες τις πύλες, κανένα δηλαδή γονίδιο δεν μπαίνει μέσα στο δίκτυο. Καθώς η διαδικασία μάθησης προχωράει, οι πύλες των γονιδίων που ελαττώνουν το σφάλμα ταχύτερα, ανοίγουν και ταχύτερα. Η εκπαίδευση της συνάρτησης της πύλης με άλλα βάρη του δικτύου. Στο τέλος της εκπαίδευσης επιλέγουμε τα σημαντικά γονίδια με βάση τις τιμές του gate opening. Η εκπαίδευση σταματάει είτε όταν το σφάλμα έχει φτάσει στο μηδέν (ή κοντά σε αυτό) είτε ύστερα από ένα συγκεκριμένο αριθμό επαναλήψεων. Διαφορετικές αρχικοποιήσεις του δικτύου μπορεί να οδηγήσουν σε υπογραφή με διαφορετικά γονίδια κάθε φορά. Καταλήγουμε συνήθως σε ένα μικρό αριθμό γονιδίων των οποίων οι πύλες είναι ανοικτές αρκετά.

2.2 Support Vector Machines

Σε αυτή την ενότητα γίνεται μελέτη των μηχανών διανυσμάτων υποστήριξης (Support Vector Machines, SVM). Ουσιαστικά το SVM [14] είναι μια δυαδική μηχανή με δυνατότητα μάθησης η οποία επιδεικνύει ορισμένες εξαιρετικά κομψές ιδιότητες. Η κεντρική ιδέα

πίσω από τα SVM συνοψίζεται ως εξής: Δοθέντος ενός δείγματος εκπαίδευσης, η μηχανή διανυσμάτων υποστήριξης κατασκευάζει ένα υπερεπίπεδο ως επιφάνεια απόφασης με τρόπο ώστε το περιθώριο διαχωρισμού μεταξύ θετικών και αρνητικών παραδειγμάτων να μεγιστοποιείται. Στην παρούσα εργασία η χρήση των SVM γίνεται με σκοπό την ταξινόμηση. Γενικά πάντως ο τομέας ο οποίος ωφελείται περισσότερο από τη χρήση των SVM είναι αυτός της ταξινόμησης προτύπων.

Όπως είπαμε και πιο πάνω το SVM προσπαθεί να βρει το καλύτερο υπερεπίπεδο το οποίο να διαχωρίζει τις δύο κλάσεις ενδιαφέροντος, θετικές (+1) και αρνητικές (-1). Αυτό επιτυγχάνεται με την μεγιστοποίηση της απόστασης $\frac{2}{||w||}$ μεταξύ δύο παράλληλων γραμμών $w \cdot x + b = 1$ και $w \cdot x + b = -1$ οι οποίες ορίζουν το περιθώριο του διαχωρισμού όπως φαίνεται και στο σχήμα 2.6. Το τελικό διαχωριστικό υπερεπίπεδο περνάει από το μέσο του περιθωρίου με την εξίσωση να είναι $w \cdot x + b = 0$. Επομένως η συνάρτηση απόφασης, είναι μια συνάρτηση της μορφής:

$$f(x) = sgn((w \cdot x) + b) \tag{2.34}$$

όπου το w αντιπροσωπεύει την κατεύθυνση του διανύσματος του υπερεπιπέδου. Η τιμή του προσήμου το οποίο επιστρέφεται από την εξίσωση (2.35) υποδηλώνει την προβλεπόμενη κλάση με την οποία συσχετίζεται το παράδειγμα x, ενώ το $|f(x)|$ υποδηλώνει το βαθμό εμπιστοσύνης στο αποτέλεσμα της επιλογής. Το πρόβλμα SVM μπορεί λοιπόν ισοδύναμα να διατυπωθεί ως:

$$\begin{aligned} minimize\ \frac{1}{2}||w||^2 + c\sum_{j=1}^{n}\xi_j^2, \\ \text{δεδομένου ότι } y_j((w \cdot x_j) + b) \geq 1 - \xi_j, \xi_j \geq 0, \\ j = 1, ..., n \end{aligned} \tag{2.35}$$

Σύμφωνα με τη θεωρία της δυϊκότητας (duality theory) το πρόβλημα μπορεί να μετατραπεί στο ακόλουθο πρόβλημα μεγιστοποίησης, όπου το λ αντιπροσωπεύει το διάνυσμα των πολλαπλασιαστών του Lagrange και το y_i αντιπροσωπεύει το label (είτε +1 είτε -1) του i ου δείγματος.

$$\begin{aligned} maximize\ (\lambda \in \mathbb{R}) \quad & \sum_{j=1}^{n}\lambda_j - \frac{1}{2}\sum_{i,j=1}^{n}\lambda_i\lambda_j y_i y_j (x_i \cdot xj) \\ & \text{δεδομένου ότι} \quad \sum_{j=1}^{n}\lambda_i y_i = 0 \\ & \text{και} \quad 0 \leq \lambda_j \leq C,\ j = 1, ..., n \end{aligned} \tag{2.36}$$

Μέσω της λύσης του παραπάνω προβλήματος, καταλήγουμε στην ακόλουθη έκφραση για την κατεύθυνση του διανύσματος w.

$$w = \sum_{j=1}^{n}\lambda_j y_j x_j \tag{2.37}$$

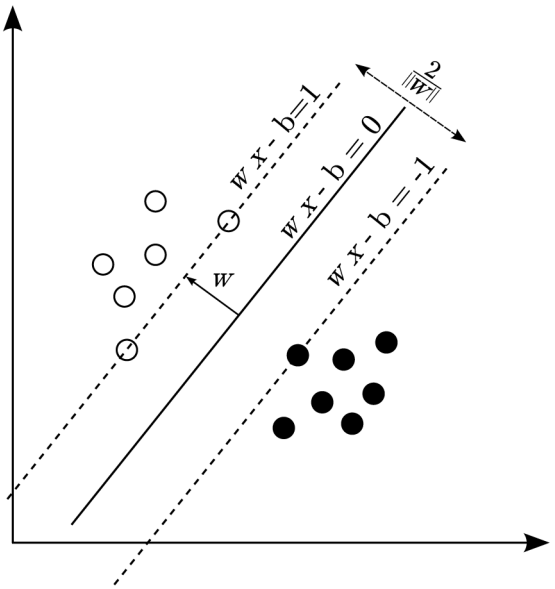


Figure 2.6: Απεικόνιση της δυαδικής ταξινόησης, που δείχνει το διαχωριστικό περιθώριο μσταξύ δύο κλάσεων. Τα σημεία πάνω στις διακεκομμένες γραμμές δείχνουν τα support vectors. Εικόνα: wikipedia.org

το οποίο είναι στην ουσία μία επέκταση για τα δείγματα εκπαίδευσης με μη μηδενικό λ_j όπως για παράδειγμα τα support vectors. Αποδεικνύεται ότι τα support vectors βρίσκονται στα όρια των κλάσεων και μπορούν να χτησιμοποιηθούν για τον υπολογισμό αντικαθιστώντας ένα από τα support vectors στην ακόλουθη εξίσωση:

$$y_j((w \cdot x_j) + b) = 1 \tag{2.38}$$

Ένα σημαντικό στοιχείο το οποίο κάνει τη χρήση των SVMs ιδιαίτερα ελκυστική είναι ότι επιτρέπουν τη χρήση kernels έτσι ώστε η εξίσωση (2.36) να μπορεί να αντικατασταθεί με μία συνάρτηση πυρηνα (kernel function) με την ακόλουθη μορφή:

$$\begin{aligned} maximize\ (\lambda \in \mathbb{R}) \quad & \sum_{j=1}^{n} \lambda_j - \frac{1}{2} \sum_{i,j=1}^{n} \lambda_i \lambda_j y_i y_j k(x_i, xj) \\ & \text{δεδομένου ότι} \quad \sum_{j=1}^{n} \lambda_i y_i = 0 \\ & \text{και} \quad 0 \leq \lambda_j \leq C,\ j = 1, ..., n \end{aligned} \tag{2.39}$$

Πέρα από γραμμικά kernels και άλλα είδη από kernel μπορούν να χρησιμοποιηθούν όπως πολυώνυμα οποιουδήποτε βαθμού καθώς και radial basis συναρτήσεις (RBF) σε κάποια από τις ακόλουθες μορφές:

$$\begin{aligned} k(x, y) &= (1 + (x \cdot y))^d \\ k(x, y) &= exp(-\gamma\, ||x - y||^2) \end{aligned} \tag{2.40}$$

2.3 Recursive Feature Elimination

2.3.1 Διαφορικώς εκφρασμένα γονίδια

Βασική ιδέα πίσω από την ανάπτυξη της μεθόδου που χρησιμοποιήσαμε είναι η αναγνώριση και επιλογή των διαφορικώς εκφρασμένων γονιδίων (differentially expressed genes). Η ιδέα αυτή δεν είναι νέα στο πεδίο του marker selection. Πολλές είναι οι εργασίες στις οποίες έχουν χρησιμοποιηθεί συντελεστές Fisher. Ο συντελεστής Fisher δίνεται από την ακόλουθη εξίσωση:

$$f_1(g_i) = \frac{(\mu_+(g_i) - \mu_-(g_i))^2}{\sigma_+(g_i)^2 + \sigma_-(g_i)^2} \tag{2.41}$$

Ενώ μια παραλαγή της παραπάνω εξίσωσης φαίνεται παρακάτω:

$$f_2(g_i) = \frac{\sum_{j=1}^{n} |g_{ij} - c(g_i)|}{\sigma_+(g_i) + \sigma_-(g_i)} \tag{2.42}$$

όπου $\mu_+(g_i)$, $\mu_-(g_i)$, $\sigma_+(g_i)$ και $\sigma_-(g_i)$ είναι τα μέσα και οι τυπικές αποκλίσεις των εκφράσεων του γονιδίου g_i στην θετική και αρνητική κλάση αντίστοιχα. Το n Είναι ο αριθμός των δειγμάτων ενώ το c ορίζεται ως:

$$c(g_i) = \frac{\mu_+(g_i) + \mu_-(g_i)}{2} \tag{2.43}$$

Μια τρίτη παραλλαγή χαμηλότερου υπολογιστικού κόστους είναι η ακόλουθη:

$$f_3(g_i) = \frac{|\mu_+(g_i) + \mu_-(g_i)|}{\sigma_+(g_i) + \sigma_-(g_i)} \tag{2.44}$$

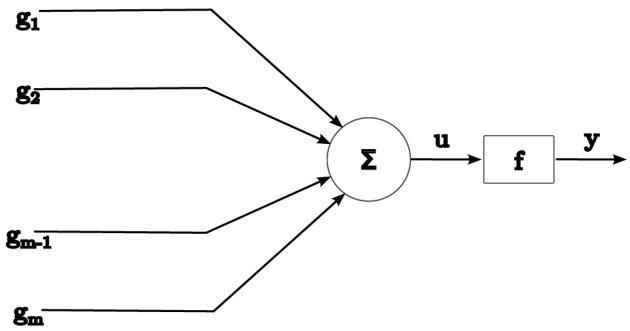
Είναι εύκολο να διαπιστώσει κανείς ότι οι (2.41), (2.42), (2.44) στην ουσία εκφράζουν την ίδια ιδέα. Η χρήση λοιπόν των παραπάνω συναρτήσεων γίνεται για την ανάθεση βαρών σε ένα set δοθέντων γονιδίων. Είναι προφανές ότι τα γονίδια τα οποία διαφοροποιούν περισσότερο τις εκφράσεις τους στις δύο καταστάσεις, δηλαδή στο αν ανήκουν στη μία ή στην άλλη κλάση, λαμβάνουν και υψηλότερες τιμές τα βάρη αυτών, από ότι τα γονίδια τα οποία κάνουν μικρότερη διαφοροποίηση. Τέλος γονίδια τα οποία εκφράζονται με ακριβώς τον ίδιο τρόπο μεταξύ των δύο καταστάσεων τότε αυτά λαμβάνουν τη χαμηλότερη τιμή στα βάρη τους, δηλαδή μηδενίζονται.

2.3.2 Η μέθοδος RFE-LNW

Οι περισσότερες marker selection τεχνικές οι οποίες εφαρμόζονται στον τομέα του DNA microarray, λόγω πολλών διαστάσεων (dimensionality) των δεδομένων χρησιμοποιούν γραμμικά εργαλεία για την εκτίμηση του προβλήματος. Μία τέτοια μέθοδος είναι και η RFE-SVM [15] όπου ένας γραμμικός πυρήνας χρησιμοποιείται για την εκτίμηση του

διανύσματος βαρών του διαχωριστικού υπερεπιπέδου, η απόλυτη τιμή του οποίου χρησιμοποιείται για το ranking των γονιδίων. Η χρήση ενός γραμμικού νευρώνα μπορεί να επιτύχει το ίδιο αποτέλεσμα καθώς μπορεί να προσεγγίσει οποιαδήποτε γραμμική συνάρτηση. Για αυτό συνιστάται η χρήση ενός γραμμικού νευρώνα (Linear Neuron) για να προσεγγίσουμε το διαχωριστικό υπερεπίπεδο μεταξύ των δύο κλάσεων.

Μπορούμε λοιπόν εύκολα να ενσωματώσουμε μία νέα διαδικασία μάθησης με εφαρμογή στο πρόβλημά μας. Κάτι τέτοιο επιτυγχάνεται με τη χρήση ενός νευρωνικού δικτύου με ένα μόλις νευρώνα m εισόδων όπως φαίνεται και στο σχήμα 2.7 όπου m είναι ο αριθμός των γονιδίων. Όταν έχουμε λοιπόν δύο πιθανές εξόδους, 0 για τη μία κλάση και 1 για την άλλη, μπορούμε να χρησιμοποιήσουμε ένα γραμμικό νευρώνα για την προσέγγιση του διαχωριστικού υπερεπιπέδου μεταξύ των δύο κλάσεων του ενδιαφέροντός μας.



Σχήμα 2.7: Ένας νευρώνας.

Με τη χρήση μάλιστα της σιγμοειδούς συνάρτησης $f(u)$ επιτυγχάνουμε

$$y = \frac{1}{1+e^u} = f(u) \tag{2.45}$$

$$u = \sum_{i=1}^{n} w_i g_i \tag{2.46}$$

$$f'(u) = y(1-y) \tag{2.47}$$

όπου $f'(u)$ είναι πάντα μεγαλύτερο ή ίσο του μηδενός καθώς το y κυμαίνεται μεταξύ 0 και 1.

2.3.3 Εκπαίδευση του RFE-LNW

Η ενότητα αυτή έχει να κάνει με την εκπαίδευση του RFE-LNW (Recursive Feature Elimination based on Linear Neuron Weights). Η συνάρτηση λάθους ενός νευρώνα δίνεται από την ακόλουθη συνάρτηση:

$$E = \frac{1}{2}\sum_{j=1}^{n}(d_j - y_j)^2 \tag{2.48}$$

όπου n ο αριθμός των δειγμάτων, d_j η επιθυμητή έξοδος του νευρώνα του δείγματος j, ενώ y_j η πραγματική έξοδος του συγκεκριμένου νευρώνα για το δείγμα j. Για την ελαχιστοποίηση της εξίσωσης (2.48) χρησιμοποιείται η gradient descent μέθοδος και με αυτό τον τρόπο ανανεώνουμε το βάρος w_i το οποίο σχετίζεται με το γονίδιο g_i. Αυτό φαίνεται παρακάτω:

$$w_i(t+1) = w_i(t) - (\mu\frac{\partial E}{\partial w_i}) = w_i(t) - \mu\sum_{j=1}^{n}(\frac{\partial E}{\partial y_i}\frac{\partial y_i}{\partial u}\frac{\partial u}{\partial w_i}) \tag{2.49}$$

Μέσω της χρήσης της μετρικής του Fisher (2.42) στην εξίσωση ανανέωση βαρών έχουμε:

$$\begin{aligned} w_i(t+1) &= w_i(t) - \frac{\mu}{2}\sum_{j=1}^{n}\left[(\frac{\partial E}{\partial y_i}\frac{\partial y_i}{\partial u}\frac{\partial u}{\partial w_i})\frac{|g_{ij} - c(g_i)|}{\sigma_+(g_i) + \sigma_-(g_i)}\right] \\ &= w_i(t) - \frac{\mu}{2}\sum_{j=1}^{n}(-2(d_j - y_j)y_j(1-y_j)g_{ij}f_2(g_i)) \\ &= w_i(t) + \mu\sum_{j=1}^{n}(d_j - y_j)y_j(1-y_j)g_{ij}f_2(g_i)) \\ &= w_i(t) + \mu\sum_{j=1}^{n}(d_j - y_j)f'(u_j)g_{ij}f_2(g_i)) \end{aligned}$$

Επομένως έχουμε:

$$w_i(t+1) = w_i(t) + \mu\sum_{j=1}^{n} e_j f'(u_j)g_{ij}\frac{|g_{ij} - c(g_i)|}{\sigma_+(g_i) + \sigma_-(g_i)} \tag{2.50}$$

όπου το t αντιπροσωπεύει την τωρινή επανάληψη, μ είναι ο βαθμός εκπαίδευσης (learning rate) ενώ το e_j είναι:

$$e_j = (d_j - y_j) \tag{2.51}$$

Αν δουλέψουμε με πρόσημα τότε οι εξισώσεις (2.50) και (2.51) μετατρέπονται σε:

- Στην περίπτωση όπου $f_2(g_i) = 1$ τότε έχουμε:

$$w_i(t+1) = w_i(t) + \mu \sum_{j=1}^{n} e_j f'(u_j) g_{ij} \tag{2.52}$$

$$w_i(t+1) = w_i(t) + \mu \sum_{j=1}^{n} sign(e_j f'(u_j)) sign(g_{ij}) \tag{2.53}$$

- ενώ γενικά έχουμε:

$$w_i(t+1) = w_i(t) + \mu \sum_{j=1}^{n} sign(e_j f'(u_j)) sign(g_{ij}) f_2(g_i) \tag{2.54}$$

$$w_i(t+1) = w_i(t) + \mu \sum_{j=1}^{n} |d_j - y_j| sign(e_j f'(u_j)) sign(g_{ij}) f_2(g_i) \tag{2.55}$$

Η εξίσωση (2.52) είναι ο βασικός gradient descent αλγόριθμος εκπαίδευσης ο οποίος ελαχιστοποίει τη συνάρτηση σφάλματος (2.51). Και η εξίσωση (2.53) όμως συγκλίνει καθώς χρησιμοποιώντας το πρόσημο κάνουμε ελαχιστοποίηση. Αναμένουμε ότι η (2.53) θα συγκλίνει ταχύτερα από την (2.52) επειδή στην τελευταία τα e_j και $f'(u_j)$ λαμβάνουν πολύ μικρές τιμές κάτι το οποίο σημαίνει πολύ μικρές διαφοροποιήσεις στο w. Με τη χρήση λοιπόν μόνο του προσήμου στη (2.53) κάνουμε μεγαλύτερα βήματα και επιταχύνουμε τη διαδικασία, ειδικά όταν έχουμε πάρα πολλά δεδομένα (γονίδια). Η εξίσωση (2.54) διαφέρει από την (2.53) στο ότι λαμβάνει υπόψη και μια εκτίμηση του αθροίσματος της μετρικής του Fisher. Κάτι τέτοιο πολλές φορές καθυστερεί την σύγκλιση. Ένα πρόβλημα που εμφανίζεται σε αυτό το σημείο είναι ότι καθώς μειώνονται τα δεδομένα είναι όλο και πιο δύσκολο να επιλυθεί το πρόβλημα κάτι που σημαίνει καθυστέρηση στη σύγκιση. Για αυτο το λόγο χρησιμοποιείται η εξίσωση (2.55). Η εξίσωση αυτή χρησιμοποιεί ένα παραλλαγμένο learning rate $|d_j - y_j|$. Το καλύτερο είναι λοιπόν όταν είμαστε ακόμη μακριά από το στόχο μας να κάνουμε μεγάλα βήματα, ενώ όταν πλησιάζουμε σε αυτόν τα βήματα να γίνονται μικρότερα. Επομένως η καλύτερη λύση σε αυτό είναι η χρήση της (2.54) και (2.55) συνδυαστικά.

Οι εξισώσεις (2.52), (2.53), (2.54) και (2.55) ανανεώνουν τον βάρος $w(t+1)$ αφού όλα τα δείγματα παρουσιαστούν στο δίκτυο και αφού το άθροισμα των όρων έχει υπολογιστεί. Αυτό σύμφωνα με τη θεωρία των νευρωνικών δικτύων ονομάζεται batch training. Εναλλακτικά μπορούμε είναι δυνατή η ανανέωση των βαρών εξετάζοντας ένα δείγμα τη φορά και σεαυτή την περίπτωση οι όροι του αθροίσματος μπορούν να αγνοηθούν. Έτσι η ανανέωση μπορεί να γίνει με τις ακόλουθες εξισώσεις:

$$w_i(t+1) = w_i(t) + \mu \cdot e \cdot f'(u) \cdot g_i \tag{2.56}$$

$$w_i(t+1) = w_i(t) + \mu \cdot sign(e \cdot f'(u)) \cdot sign(g_i) \tag{2.57}$$

$$w_i(t+1) = w_i(t) + \mu \cdot sign(e \cdot f'(u)) \cdot sign(g_i) \cdot f_2(g_i) \tag{2.58}$$

$$w_i(t+1) = w_i(t) + |d-y| \cdot sign(e \cdot f'(u)) \cdot sign(g_i) \cdot f_2(g_i) \tag{2.59}$$

η ανανέωση λοιπόν το βαρών γίνεται στοιχείο προς στοιχείο και οδηγεί μάλιστα σε καλύτερα αποτελέσματα έναντι της batch μεθόδου.

2.3.4 Αλγοριθμική παρουσίαση του RFE-LNW

Στον πίνακα 2.2 φαίνεται η αλγοριθμική παρουσίαση του RFE-LNW. Ο συγκεκριμένος αλγόριθμος κατορθώνει να εφαρμόσει filter κριτήρια με ένα wrapper τρόπο, όπου τα βάρη επαναεκτιμούνται σε κάθε επανάληψη και ανανεώνονται σε κάθε βήμα. Με τη μείωση των γονιδίων επιτυγχάνεται και μείωση των διαστάσεων του προβλήματος (dimensionality). Έτσι όσο το πρόβλημα μειώνεται σιγά σιγά αφαιρούνται τα μη σημαντικά γονίδια με βάση το διάνυσμα w. Είναι σημαντικό κάτι τέτοιο να γίνεται βήμα προς βήμα ώστε τα μη σημαντικά γονίδια να μην επισκιάσουν τα σημαντικά.

2.4 LASSO

2.4.1 Εισαγωγή σε Linear Regression

Το Linear Regression (γραμμική παλινδρόμηση) έχει σαν σκοπό την μοντελοποίηση των μεταβλητών με τη χρήση γραμμικών συνδυασμών κάποιων παρατηρήσεων ή μετρήσεων. Οι παρατηρήσεις αυτές συνήθως καλούνται predictors και συμβολίζονται με x, ενώ οι έξοδοι καλούνται responses και συμβολίζονται με y. Το linear regression μας παρέχει μια εκτίμηση $\hat{y}$ της πραγματικής εξόδου.
Έστω λοιπόν το διάνυσμα εισόδου:

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix}$$

Και το linear regression μοντέλο έχει τη μορφή:

$$\hat{y} = f(x) = \beta_0 + \sum_{j=1}^{n} x_j \beta_j \tag{2.60}$$

Table 2.2: RFE-LNW algorithm

1. Έστω m ο αρχικός αριθμός των γονιδίων.
2. While (m>=0)
3. Ανανέωση του διανύσματος βαρών με τη χρήση τω εξισώσεων (2.58) και (2.59).
4. Ταξινόμηση (rank) των γονιδίων σύμφωνα με την απόλυτη τιμή του διανύσματος w.
5. Αφαίρεση των γονιδίων τα οποία έχουν την μικρότερη απόλυτη τιμή στο w ($m \leftarrow m - 1$). Περισσότερα από ένα γονίδια μπορούν να αφαιρεθούν σε κάθε επανάληψη.
6. Εκτίμηση της ακρίβειας της ταξινόμησης με τη χρήση ενός γραμμικού SVM classifier.
7. Τέλος του while
8. Κρατάμε τα λιγότερα γονίδια τα οποία επιτυγχάνουν το καλύτερο classification accuracy.

Υποθέτουμε ότι οι σχέσεις μεταξύ x και y είναι γραμμικές ή περίπου γραμμικές. Το β_i είναι αρχικά άγνωστο και υπολογίζεται μέσω μιας διαδικασίας εκπαίδευσης. Ορίζουμε ως X τον πίνακα εισόδου διαστάσεων $N \times p$, του οποίου κάθε γραμμή αντιστοιχεί σε ένα διάνυσμα εισόδου. Ορίζουμε το y ως το διάνυσμα το οποίο αντιστοιχεί στις εξόδους των γραμμών του X. Ο δημοφιλέστερος εκτιμητής για το β_i είναι αυτός των ελαχίστων τετραγώνων (least squares) ο οποίος ελαχιστοποιεί το τετραγωνικό αθροιστικό σφάλμα (residual sum of squares - RSS):

$$RSS(\beta) = \sum_{i=1}^{N}(y_i - f(x_i))^2 = \sum_{i=1}^{N}(y_i - \sum_{j=1}^{p} X_{ij}\beta_j))^2 \tag{2.61}$$

Ορίζουμε ως X το $N \times (p+1)$ πίνακα εκπαίδευσης, όπου κάθε γραμμή αντιστοιχεί σε ένα διάνυσμα δεδομένων και το y αντιστοιχεί στο $N \times 1$ διάνυσμα εξόδων. Οπότε μπορούμε να γράψουμε το RSS σαν:

$$RSS(\beta) = ||y - X\beta||^2 = (y - X\beta)^T(y - X\beta)$$
$$\frac{\partial RSS}{\partial \beta} = -2X^T(y - X\beta) \Rightarrow$$
$$\frac{\partial^2 RSS}{\partial \beta^2} = 2X^T X$$

Υποθέτουμε ότι το X είναι πλήρους βαθμού στηλών και ότι ο $(X^T X)^{-1}$ υπάρχει. Έτσι θέτοντας $\frac{\partial RSS}{\partial \beta} = 0$ καταλήγουμε στη μοναδική λύση:

$$\beta_{LS} = (X^T X)^{-1} X^T y \tag{2.62}$$

2.4.2 Μέθοδοι ελαχίστων τετραγώνων

Εκτός από το απλώς να ελαχιστοποιούμε τη συνάρτηση του τετραγωνικού αθροιστικού σφάλματος (residual sum of squares) υπάρχει και η δυνατότητα προσθήκης ενός penalty στη συνάρτηση κόστους το οποίο οδηγεί σε πιο ερμηνεύσιμα μοντέλα με τη μείωση ή απόρριψη κάποιων τιμών των predictors.

2.4.3 Lasso Regression

Το LASSO (Least Absolute Shrinkage and Selection Operator) είναι μια μέθοδος ελάττωσης και επιλογής για προβλήματα linear regression, η οποία παρουσιάστηκε από τον Tibshirani το 1996 [16]. Το lasso ελαχιστοποιεί το RSS, και επιπλέον επιβάλει έναν επιπλέον περιορισμό στην $\ell - 1$ νόρμα του υπό εκτίμηση διανύσματος. Με τη χρήση αυτού του περιορισμού, κάποιοι συντελεστές τείνουν να μηδενιστούν και με αυτό τον τρόπο έχουμε ένα πιο εννοιολογικά ερμηνεύσιμο μοντέλο.

Το μοντέλο του lasso είναι παρόμοιο με αυτό του linear regression. Υποθέτουμε ότι X ένας πίνακας μεγέθους $N \times p$ ο οποίος περιλαμβάνει τις predictors τιμές και y διάνυσμα το οποίο περιέχει τις responses. Ακόμη δηλώνουμε το:

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{bmatrix}$$

Έτσι η εκτίμηση του $(\hat{\alpha}, \hat{\beta})$ γίνεται μέσω της ακόλουθης φόρμουλας:

$$\min_{\beta} \{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j})^2 \} \tag{2.63}$$
$$\text{δεδομένου ότι} \sum_{j=1}^{p} |\beta_j| \leq t$$

Το άνω όριο t είναι μια ρυθμιζόμενη παράμετρος. Όταν το t είναι αρκετά μεγάλο, ο περιορισμός δεν έχει καμία επίδραση και σε αυτή την περίπτωση η λύση είναι απλά το γραμμικό regression των ελαχίστων τετραγώνων του y πάνω στο X. Για μικρότερες όμως τιμές του t οι λύσεις είναι μειωμένες εκδόσεις του προβλήματος της εκτίμησης των ελαχίστων τετραγώνων. Συχνά οι συντελεστές β_j μηδενίζονται κάτι και το οποίο πολύ συχνά επιθυμούμε. Η επιλογή του t είναι σαν να επιλέγουμε τον αριθμό των predictors για να χρησιμοποιήσουμε στο regression μοντέλο.

Μια εναλλακτική μέθοδος για να εκφράσουμε το lasso είναι:

$$\min_{\beta}\{\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{i,j})^2 + \lambda\sum_{j=1}^{p}|\beta_j|\} \tag{2.64}$$

με το $\lambda \geq 0$ να αντικαθιστά το t της εξίσωσης (2.63) υπό την έννοια ότι για κάθε t υπάρχει ένα αντίστοιχο λ που δίνει την ίδια λύση. Στην παρούσα εργασία βασιστήκαμε κατά κύριο λόγο στη χρήση της εξίσωσης (2.64)

ℓ_1 - Regularized Least Squares

Στο ℓ_1 - Regularized Least Squares για την επίλυση του Lasso regression έχουμε:

$$\min ||Ax - y||_2^2 + \lambda\,||x||_1 \tag{2.65}$$

με το $||x||_1 = \sum_i^n |x_i|$ να υποδηλώνει την ℓ_1 νόρμα του x ενώ $\lambda > 0$ η παράμετρος regularization.

2.4.4 Ridge Regression

Μια ενναλακτική μέθοδος για την ελαχιστοποίηση του RSS είναι αυτή του Ridge Regression. Όπως και παραπάνω έτσι και εδώ X είναι ο $N \times p$ πίνακας των τιμών των predictors και y οι αποκρίσεις. Η εκτίμηση για το Ridge Regression επιτυγχάνεται με τη λύση της ακόλουθης εξίσωσης:

$$\min_{\beta}\{\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{i,j})^2 + \lambda\sum_{j=1}^{p}\beta_j^2\} \tag{2.66}$$

ή εναλλακτικά:

$$\min_{\beta}\{\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{i,j})^2\} \tag{2.67}$$

$$\text{δεδομένου ότι} + \lambda\sum_{j=1}^{p}\beta_j^2 \leq t \tag{2.68}$$

Σε αυτό το σημείο παρουσιάζουμε το Ridge Regression για να καταδείξουμε την σημασία του ℓ_1 norm penalty που εισαγάγει το Lasso , στην εφαρμογή μας.

2.4.5 Σύγκριση Ridge και Lasso Regression

Με τη χρήση του Lasso αρκετά συχνά πολλοί από τους συντελεστές μηδενίζονται. Αυτό είναι κάτι το οποίο επιθυμούμε όπως θα φανεί και αργότερα. Θα δείξουμε λοιπόν γιατί το Lasso έναντι του Ridge Regression παράγει μηδενικούς ή σχεδόν μηδενικούς συντελεστές.

Το penalty για κάθε μία από τις δύο μεθόδους σχετίζεται με την p-norm στη φόρμα βελτιστοποίησης. Στην περίπτωση του Ridge Regression έχουμε τα τετράγωνα των συντελεστών, και έτσι η νόρμα η οποία προστίθεται είναι p=2. Στο Lasso Regression τώρα οι απόλυτες τιμές των συντελεστών προστίθενται, οπότε σε αυτή την περίπτωση p=1. Μελετώντας τις εξισώσεις των δύο μεθόδων παρατηρούμε ότι το μόνο το οποίο διαφέρει είναι το penalty. Οι διαφορές λοιπόν στους συντελεστές του regression οφείλονται στη διαφορά του penalty.

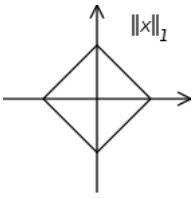
Το κριτήριο:

$$\sum_{i=1}^{N}(y_i - \sum_j \beta_j x_{i,j}) \tag{2.69}$$

ισοδυναμεί με την τετραγωνική εξίσωση:

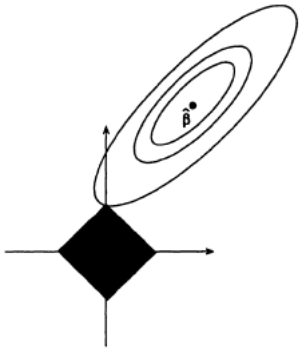
$$(\beta - \hat{\beta})X^T X(\beta - \hat{\beta}) \tag{2.70}$$

Η παραπάνω τετραγωνική εξίσωση αντιστοιχεί στις ελυπτικές καμπύλες οι οποίες κεντράρονται στις εκτιμήσες των ελαχίστων τετραγώνων του $\hat{\beta}$. Για κάθε μία από τις δύο μεθόδους, έχουμε να εξετάσουμε πως το penalty μπορεί να απεικονιστεί στο χώρο. Ένας τρόπος να κάνουμε κάτι τέτοιο είναι η απεικόνιση μιας p-norm στον L^p χώρο. Αποδεικνύεται ότι ότι η $l-1$ νόρμα ενός διανύσματος απεικονίζεται σαν μοναδιαίο ορθογώνιο παραλληλόγραμμο, ενώ η $l-2$ νόρμα σαν μοναδιαίος κύκλος. Η εικόνα (2.8) για την $l-1$ νόρμα φαίνεται παρακάτω:

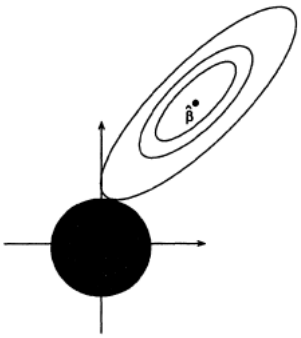


Σχήμα 2.8: Νόρμα $l-1$ στο L^p. Εικόνα: wikipedia.org

Όταν πλοτάρουμε το penalty της τετραγωνικής συνάτησης, παρατηρούμε (όπως φαίνεται και στα σχήματα 2.9 και 2.10) ότι στην περίπτωση του Lasso, το περίγραμμα της τετραγωνικής συνάρτησης τέμνει το ορθογώνιο της $l-1$ νόρμας. Αυτό μάλιστα συμβαίνει σε μία γωνία παράγοντας μηδενικό συντελεστή. Από την άλλη το Ridge το περίγραμμα τέμνει το μοναδιαίο κύκλο σε σημεία κοντά στο μηδέν αλλά όχι ακριβώς μηδέν. Αυτό το γεγονός περιγράφει διαισθητικά το λόγο για τον οποίο το Lasso δίνει αραιές λύσεις.



Σχήμα 2.9: Lasso Regression.



Σχήμα 2.10: Ridge Regression.

2.4.6 Εφαρμογή

Στην παρούσα εφαρμογή, έχουμε το παρακατω μοντέλο regression:

$$y = Ax$$

με το $x \in \mathbb{R}^n$ να είναι ένα διάνυσμα αγνώστων, το $y \in \mathbb{R}^m$ είναι το διάνυσμα των παρατηρήσεων και $A \in \mathbb{R}^{n \times n}$ ο πίνακας δεδομένων.
Το y είναι το target μας, δηλαδή οι δύο κατηγορίες στις οποίες μπορεί να ανήκουν τα samples, A είναι ο πίνακας ο οποίος περιέχει τα δεδομένα, δηλαδή samples $\times$ genes και x είναι τα γονίδια τα οποία ψάχνουμε.

2.5 Fuzzy Clustering

2.5.1 Clustering

Το Clustering είναι μια διαδικασία κατά την οποία διαχωρίζουμε ένα set παρατηρήσεων σε υποσύνολα (subsets ή clusters) έτσι ώστε οι παρατηρήσεις στο ίδιο cluster να είναι παρόμοιες υπό κάποια μετρική. Clustering στο X είναι εύρεση ενός ακεραίου c, $2 \leq c < n$, και το σπάσιμο του X σε c αμοιβαίως αποκλειόμενα υποσύνολα του X. Τα μέλη του κάθε cluster έχουν μεγαλύτερη σχέση μεταξύ τους, από ότι με τα μέλη των υπόλοιπων clusters.

Υπάρχουν διάφορες μορφές clustering: Δύο σημαντικές κατηγορίες είναι το Hard Clustering και το Fuzzy Clustering. Στην παρούσα εργασία χρησιμοποιούμε Fuzzy Clustering. Η διαφορά μεταξύ του hard και του fuzzy clustering είναι ότι στο hard clustering τα δεδομένα χωρίζονται σε διακριτά clusters όπου το κάθε στοιχείο δεδομένων ανήκει σε ακριβώς μία κατηγορία, σε αντίθεση με το fuzzy όπου τα δεδομένα μπορούν να ανήκουν σε περισσότερες από μία κατηγορίες.

2.5.2 Fuzzy Clustering

Έστω X οποιοδήποτε πεπερασμένο σύνολο. Το V_{cn} είναι ένα σύνολο πινάκων μεγέθους $c \times n$, με c ακέραιο μεταξύ $2 \leq c < n$. Ο fuzzy c-partition χώρος είναι ένα σύνολο:

$$M_{f_{cn}} = \{U \in V_{cn} \mid u_{ik} \in [0,1] \; \forall i,k; \; \sum_{i=1}^{c} u_{ik} = 1 \; \forall \; k; 0 < \sum_{k=1}^{n} u_{ik} < n \; \forall \; i\} \tag{2.71}$$

Η κάθε σειρά του πίνακα $U \in M_{f_{cn}}$ δείχνει το $i^{\text{στο}}$ βαθμό συμμετοχής u_i στο fuzzy c-partition U του X. Επειδή το άθροισμα της κάθε στήλης ισούται με 1, ο συνολικός βαθμός συμμετοχής του κάθε x_k στο X είναι 1. Λόγω του γεγονότος ότι το $0 \leq u_{ik} \leq 1 \forall i,k$ είναι πιθανό το x_k να παρουσιάσει ένα αυθαίρετο βαθμό συμμετοχής στα c υποσύνολα u_i του X. Ακόμη είναι πιθανό να υπάρχουν κάποιες στήλες στις οποίες να εκχωρήσουν το συνολικό ποσοστό συμμετοχής κάποιων x_k σε ένα μόνο u_i. Κάτι τέτοιο μας δείχνει ότι το M_{cn} (χώρος του hard clustering) είναι ένα πεπερασμένο υποσύνολο του $M_{f_{cn}}$

Έστω $R = [r_{ij}]_{n \times n}$ οποιαδήποτε σχέση η οποία ικανοποιεί τα ακόλουθα κριτήρια:

- $r_{ij} \geq 0, \forall i,j$
- $r_{ii} = 0$
- $r_{ij} = r_{ji}$

Καλούμε λοιπόν την R, σχέση ανομοιότητας (dissimilarity relation). Η σχέση R μπορεί να μην ικανοποιεί την τριγωνική ανισότητα, στην περίπτωση όμως που την ικανοποιεί τότε θεωρείται συνάρτηση απόστασης. Μια σχέση ανομοιότητας $R = [r_{ij}]_{n \times n}$ καλείται ευκλείδια, αν υπάρχει ένα σύνολο από διανύσματα $\{x_1, ..., x_n\} \subset R^{n-1}$ έτσι ώστε το r_{ij} να είναι η ευκλείδια απόσταση μεταξύ του x_i και του x_j. Σε αντίθετη περίπτωση η σχέση δεν είναι ευκλείδια.

Στην περίπτωσή μας θέλουμε να εξετάσουμε κατά πόσο υπάρχουν γονίδια τα οποία είναι μέχρι κάποιο βαθμό θετικά συσχετισμένα. Αν έχουμε λοιπόν m δείγματα και n γονίδια, τότε θα έχουμε n διανύσματα, κάθε ένα στο R^m, $\{x_1, ..., x_n\} \subset R^m$. Πρέπει να σημειωθεί ότι δεν μιλάμε για τις ευκλείδιες αποστάσεις μεταξύ διανυσμάτων, καθώς η απόσταση μεταξύ δύο υψηλά συσχετισμένων διανυσμάτων μπορεί να είναι γενικά πολύ μεγάλη. Ακόμη τόσο δύο θετικά συσχετισμένα, όσο και δύο αρνητικά συσχετισμένα γονίδια μπορεί να οδηγήσουν σε αντίστοιχες αποστάσεις, κάτι το οποίο δεν είναι επιθυμητό στην παρούσα εφαρμογή.

Για τους παραπάνω λόγους υπολογίζουμε τους Pearsons συντελστές συσχέτισης μεταξύ ζευγαριών διανυσμάτων, δηλαδή x_i και x_j. Ο υπολογισμός των Pearsons συντελστών μεταξύ δύο μεταβλητών X και Y με μέση τιμή μ_X και μ_U και τυπική απόκλιση σ_X και σ_U αντίστοιχα δίνεται από τον ακόλουθο τύπο:

$$\rho_{X,Y} = corr(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} \tag{2.72}$$

Ενώ στην περίπτωση όπου έχουμε n μετρήσεις του X και Y, ως x_i και y_i με $i = 1, 2, ..., n$ τότε θα έχουμε:

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{2.73}$$

Αυτή η σχέση R μπορεί να έχει τόσο θετικές, όσο και αρνητικές τιμές και έτσι δεν είναι σχέση αυτοσυσχέτισης. Για τη μετατροπή της σε σχέση αυτοσυσχέτισης κάνουμε τον ακόλουθο μετασχηματισμό: $R = 1 - R$, δηλαδή $r_{ij} = 1 - r_{ij}$. Αυτός ο μετασχηματισμός κάνει κάνει τα $r_{ii} = 0$ και $r_{ij} \geq 0$. Η συμμερτία της σχέσης R θα διατηρηθεί, ενώ τα αρνητικά συσχτισμένα διανύσματα θα εμφανίσουν μεγαλύτερες τιμές ανομοιότητας ενώ τα θετικά μικρότερες. Συνεπώς έχουμε μια έγκυρη σχέση ανομοιότητας, όχι όμως απαραίτητα ευκλείδεια. Θέλουμε να ομαδοποιήσουμε την παραπάνω την παραπάνω σχέση σε c clusters.

2.5.3 Ο αλγόριθμος NERFCM

Ο αλγόριθμος NERFCM [17] (Non-Euclidean Relational Fuzzy c-means) βρίσκει ένα fuzzy partition πίνακα $U = [u_{i,k}]_{c,n}$. Το $u_{i,k}$ δείχνει το βαθμό συμμετοχής του x_k στο $i^{\text{στο}}$ cluster, με $u_{i,k} \geq 0$ και $\sum_{i=1}^{i=c} u_{i,k} = 1$. Ο βαθμός συμμετοχής μας βοηθάει στο πόσο συμπαγές είναι κάθε cluster. Ο αλγόριθμος μετατρέπει μια μη ευκλείδια σχέση σε ευκλείδια χρησιμοποιώντας μια μετατροπή γνωστή ως $\beta - transformation$. Ο $\beta - transformation$ είναι η

$$R \Rightarrow R_\beta = R + \beta(M - I_n) \tag{2.74}$$

με $M_{i,j} = 1, \forall i, j$, ενώ I_n ο $n \times n$ μοναδιαίος πίνακας και β μια κατάλληλα επιλεγμένη μεταβλητή κλιμάκωσης.
Στη συνέχεια με τη χρήση ενός επαναληπτικού αλγορίθμου υπολογίζουμε τον πίνακα U.

2.6 Genotator

Το Genotator [18] είναι μία μετά-βάση δεδομένων που συγκεντρώνει και βαθμολογεί συσχετίσεις μεταξύ γονιδίων και ασθενειών από αξιόπιστες βάσεις δεδομένων που βρίσκονται στο διαδίκτυο. Το Genotator είναι λοιπόν ένα εργαλείο που ενσωματώνει αυτόματα δεδομένα από 11 εξωτερικά προσβάσιμες κλινικές πηγές και με τη χρήση μιας φόρμουλας χρησιμοποιεί αυτά τα δεδομένα ώστε να ταξινομήσει τα γονίδια ανάλογα με τη σχετικότητά τους για κάποια συγκεκριμένη ασθένεια.

Ο τρόπος με τον οποίο το genotator δίνει τιμές (GS -Genotator Scores-) στα γονίδια είναι ο ακόλουθος:

$$GS = GAD_Y - GAD_N + \phi(GPS) + \frac{1}{\gamma}(DB) + \frac{1}{\kappa}(REF) \tag{2.82}$$

όπου

Πίνακας 2.3: NERF c-means algorithm

Δεδομένου σχεσιακών δεδομένων D, ορίζουμε το c μεταξύ $2 \leq c \leq n$, το $m > 1$, αρχικοποιούμε το $\beta = 0$ και $U^{(0)} \in M_{f_{cn}}$.

1. Υπολογισμός των c-means διανυσμάτων $v_i = v_i^r$ χρησιμοποιώντας $U = U^{(r)}$ και τις εξισώσεις για $1 \leq i \leq c$:

$$v_i = \frac{U_{i1}^m, U_{i2}^m, ..., U_{in}^m}{U_{i1}^m + U_{i2}^m + ... + U_{in}^m} \tag{2.75}$$

2. Υπολογίζουμε το:

$$d_{ik} = (D_\beta v_i)_k - \frac{v_i^T D_\beta v_i}{2}, \text{για } 1 \leq c \text{ και } 1 \leq k \leq n \tag{2.76}$$

Αν $d_{ik} < 0$ για κάθε i και k τότε:

$$\text{Υπολογίζουμε το } \Delta\beta = \max \frac{-2d_{ik}}{||v_i - e_k||^2} \tag{2.77}$$

$$\text{Ανανεώνουμε το } d_{ik} \leftarrow d_{ik} + \frac{\Delta\beta}{2} ||v_i - e_k||^2 \text{ για } 1 \leq c \text{ και } 1 \leq k \leq n \tag{2.78}$$

$$\text{Ανανεώνουμε και το } \beta = \beta + \Delta\beta \tag{2.79}$$

3. Ανανεώνουμε το $U^{(r)}$ σε $U = U^{(r+1)} \in M_{f_{cn}}$ για κάθε $k = 1, ..., n$:
Αν $d_{ik} > 0$ για όλα τα i τότε:

$$U_{ik} = \frac{1}{(\frac{d_{ik}}{d_{1k}} + \frac{d_{ik}}{d_{2k}} + ... + \frac{d_{ik}}{d_{ck}})^{\frac{1}{m-1}}} \tag{2.80}$$

$$\text{εναλλακτικά } U_{ik} = 0 \text{ αν } d_{ik} > 0,\ U_{ik} \in [0,1] \text{και } (U_{1k} + ... + U_{ck}) = 1 \tag{2.81}$$

4. Ελέγχουμε για σύγκλιση με τη χρήση οποιασδήποτε νόρμας μας βολεύει:
Αν $\left|\left|U^{(r+1)} - U^{(r)}\right|\right| \leq \epsilon$ σταματάμε
εναλλακτικά $r = r + 1$ και επιστρέφουμε στο βήμα 1.

- GAD_Y είναι ο συνολικός αριθμός των ¨Ναι¨ συσχετίσεων μεταξύ γονιδίου και ασθενειας στην Genetic Association Database.
- GAD_N είναι ο συνολικός αριθμός των ¨Οχι¨ συσχετίσεων μεταξύ γονιδίου και ασθενειας στην Genetic Association Database.
- GPS είναι το Gene Prospector's Score για τη σχέση γονιδίου - ασθένειας.
- DB είναι ο συνολικός αριθμός των βάσεων δεδομένων (εκ των 11), όπου το γονίδιο εμφανίζεται.
- REF είναι ο συνολικός αριθμός αναφορών στο συγκεκριμένο γονίδιο.
- ϕ, γ και κ είναι σταθερές που χρησιμοποιούνται για τα GPS, DB και REF αντίστοιχα.

Έτσι ανάλογα με το GS γίνεται και ταξινόμηση των γονιδίων με βάση.

Κεφάλαιο 3

Αποτελέσματα

Με τη χρήση του dataset για τον καρκίνο του μαστού [1] και την εφαρμογή των αλγορίθμων που περιγράψαμε παραπάνω καταλήξαμε σε τέσσερις γονιδιακές υπογραφές. Ο αριθμός των γονιδίων της κάθε υπογραφής διαφέρει, όμως τα ποσοστά επιτυχίας των υπογραφών είναι κοντινά. Όταν αναφερόμαστε σε ποσοστό επιτυχίας εννοούμε το πόσο καλά έγινε η ταξινόμηση (classification) με τη χρήση ενός SVM. Στα αρχικά 24188 γονίδια του data set εκπαιδεύουμε το SVM με τα δείγματα του train set και ελέγχουμε το ποσοστό επιτυχίας με το test set. Το ποσοστό επιτυχίας το οποίο επιτυγχάνεται κυμαίνεται μεταξύ 78-79%. Το συγκεκριμένο ποσοστό επιτυχίας είναι αρκετά σημαντικό καθώς με αυτό τον τρόπο μπορούμε να έχουμε ένα κριτήριο σύγκρισης για τις υπογραφές στις οποίες καταλήγουμε.

3.1 Πρώτη γονιδιακή υπογραφή

Η πρώτη γονιδιακή υπογραφή στην οποία καταλήξαμε προκύπτει από την εφαρμογή του αλγορίθμου RFE-LNW. Έχουμε αναφερθεί στην αλγοριθμική παρουσίαση στο 2.2. Όπως φαίνεται στο βήμα 6 του αλγορίθμου κάνουμε μια εκτίμηση της ακρίβειας της ταξινόμησης με τη χρήση ενός γραμμικού SVM. Επομένως μπορούμε να καταλήξουμε σε μια γραφική παράσταση η οποία να μας δείχνει το ποσοστό επιτυχίας του SVM καθώς τα γονίδια ελαττώνονται. Η συγκεκριμένη γραφική παράσταση φαίνεται στο σχήμα 3.1. Πιο συγκεκριμένα το ποσοστό επιτυχίας για τα τελευταία χίλια γονίδια φαίνεται στο σχήμα 3.2.

Μελετώντας προσεκτικά τις γραφικές παραστάσεις 3.2 και 3.1, παρατηρούμε ότι περίπου στα 200 γονίδια και πιο συγκεκριμένα στα 190 εμφανίζεται το υψηλότερο ποσοστό επιτυχίας, περίπου 82-83%, υψηλότερο δηλαδή από αυτό το οποίο ξεκινήσαμε αρχικά. Τα 190 λοιπόν γονίδια στα οποία καταλήξαμε εμφανίζεται το μεγαλύτερο ποσοστό επιτυχίας και αυτά είναι αυτά που κρατάμε ως γονιδιακή υπογραφή από την εφαρμογή του αλγορίθμου RFE-LNW.

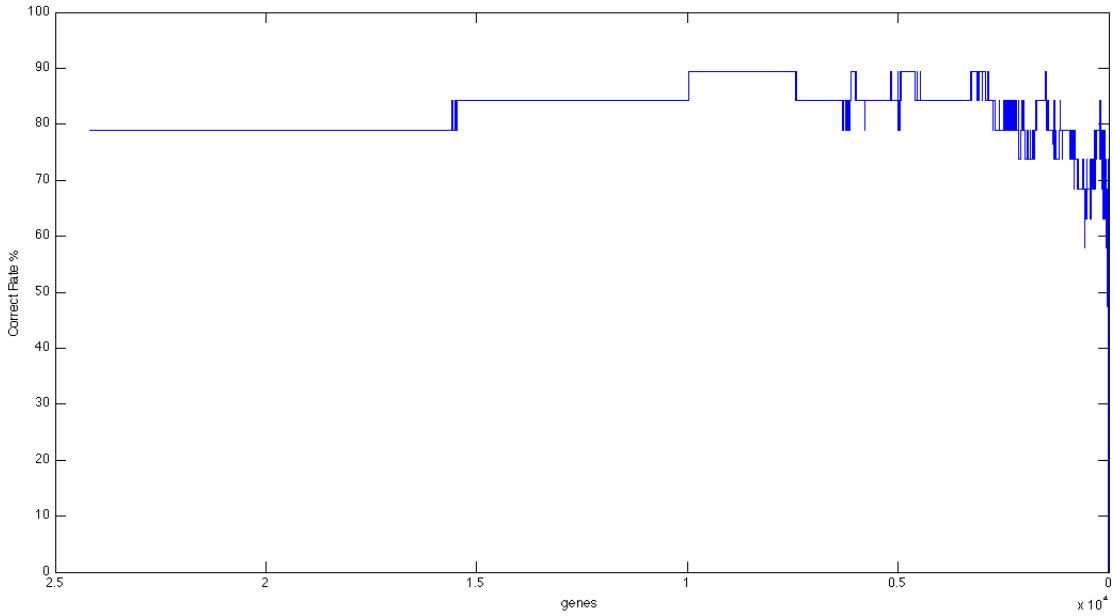


Figure 3.1: Πσοστό επιτυχίας καθώς ελαττώνουμε γονίδια με RFE-LNW.

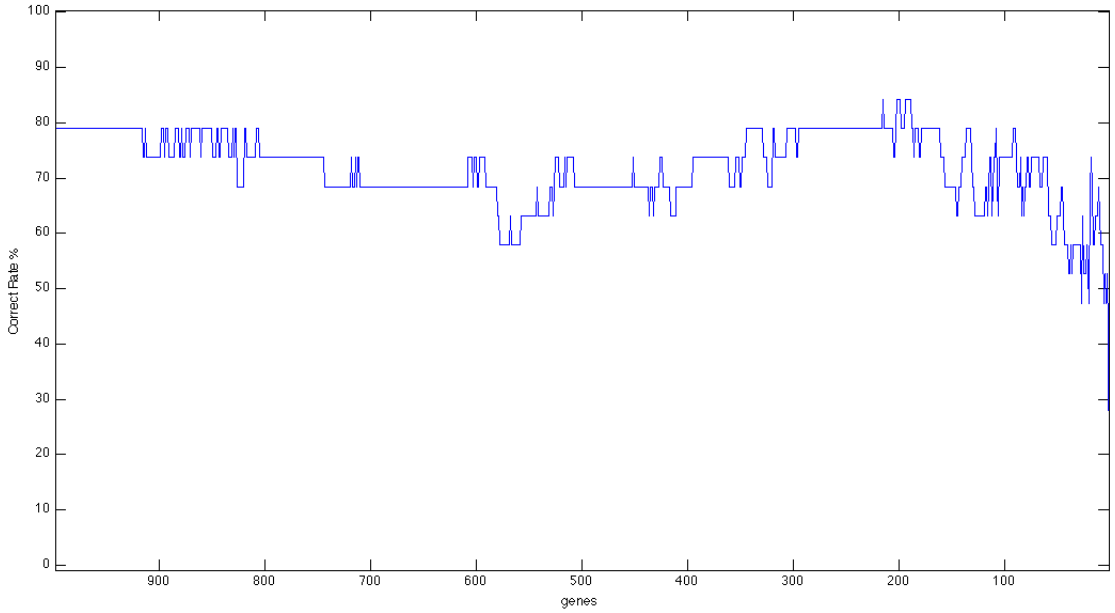


Figure 3.2: Πσοστό επιτυχίας για τα τελευταία 1000 γονίδια με τη μέθοδο RFE-LNW.

Παρακάτω φαίνονται τα 190 γονίδια από την εφαρμογή του αλγορίθμου RFE-LNW πάνω στο dataset.

GeneSeq	GeneID	Gene Symbol
462	NM_003079	SMARCE1
838	Contig13548	0
1163	NM_003150	STAT3
1202	NM_002434	MPG
1242	NM_001715	BLK
1724	Contig49185	SMAC
1899	Contig16772	0
1927	NM_001808	CELL
2078	NM_001842	CNTFR
2580	NM_012075	CGTHBA
2608	Contig61325	PRO0149
2641	AB007915	KIAA0446
2648	Contig44286	KIAA0780
2663	NM_003344	UBE2H
2727	NM_002629	PGAM1
2773	NM_003366	UQCRC2
2804	NM_002642	PIGC
2896	Contig37327	HLA-F
3153	Contig31164	0
3251	Contig42933	0
3598	NM_002720	PPP4C
3600	NM_003450	ZNF174
3661	NM_003468	FZD5
3745	Contig26988	0
3772	Contig19089	0
3805	Contig53488	0
3995	AF000416	EXTL2
4092	NM_012225	NUBP2
4346	NM_002808	PSMD2
4456	NM_002833	PTPN9
4710	AF043324	NMT1
4873	AL133447	0
4952	NM_005007	NFKBIL1
4979	NM_020321	ACCN3
4991	NM_005016	PCBP2
4996	NM_021055	TSC2
5092	NM_012326	MAPRE3
5156	NM_004314	ART1
5377	Contig54742	0
5415	Contig63079	CD74
5928	Contig54345	0
5932	Contig5681_	0
6367	AL117590	0
6404	NM_005176	ATP5G2
6473	Contig15355	0
6683	Contig29529	0
6786	Contig27722	0
6788	AF131773	GTR2
6895	AL117606	0
7238	Contig43621	0
7369	AL109698	0
7422	Contig15795	0
7453	X07618	CYP2D7AP
7615	AF131828	0
7729	Contig1699_	FLJ12438
7777	Contig26816	0
7874	NM_013323	SNX11
7951	NM_006049	SNAPC5
8011	Contig55302	LOC57020
8763	NM_006110	CD2BP2
8910	NM_013438	UBQLN1
9005	NM_014185	HSPC165
9380	Contig1403_	0
9467	NM_014230	SRP68
9476	NM_014233	UBTF
9639	Contig51498	0
9655	NM_004804	CIAO1
9704	NM_014293	NPTXR
9721	NM_014299	BRD4
9761	NM_006289	TLN1
9859	Contig20184	0
9874	NM_005594	NACA
10077	AB029012	KIAA1089
10377	NM_007066	PKIG
10389	NM_014359	OPTC
10458	NM_006354	TADA3L
11010	Contig7976_	0
11065	AL110147	DKFZP586C13
11100	NM_007108	TCEB2
11145	AB002324	KIAA0326
11342	NM_005705	PHEMX
11354	NM_007167	ZNF258
11441	NM_014473	HSA9761
11481	NM_006462	XAP4
11680	NM_005799	INADL
11731	Contig25546	0
11893	Contig33442	0
11984	AB011157	KIAA0585
12175	NM_006523	XPNPEPL
12259	NM_006544	SEC10L1
12405	D86971	KIAA0217

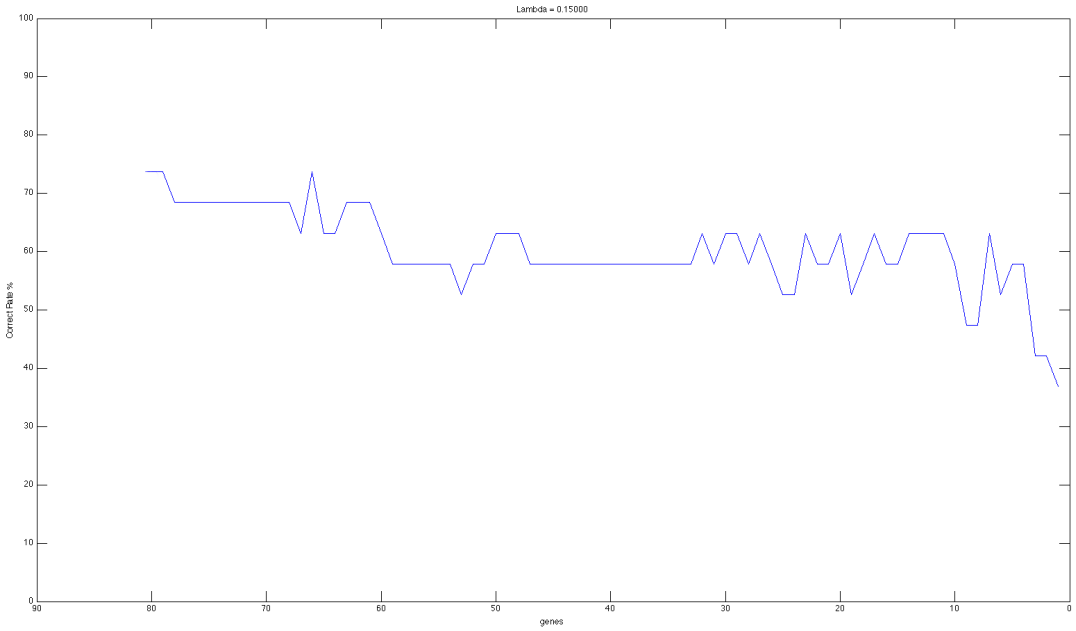
12461	Contig31153	0
12628	Contig34367	0
13170	NM_013941	OR10C1
13189	NM_014675	KIAA0445
13485	AK001092	0
13508	Contig54606	0
13530	U50383	RAI15
13690	NM_016119	LOC51131
13810	NM_016163	KIAA0917
13818	NM_016166	DDXBP1
13835	NM_014714	KIAA0590
13838	NM_016173	HEMK
13917	Contig65439	FLJ21939
14030	NM_014772	KIAA0427
14059	NM_006761	YWHAE
14076	NM_006768	BRAP
14196	Contig13551	0
14324	AL162078	DKFZp761H229
14331	Contig36715	0
14374	AL137323	0
14473	Contig50194	EST00098
14786	NM_006838	MNPEP
14796	Contig34634	GCN1L1
15155	Contig56569	0
15226	Contig319	0
15244	NM_016302	LOC51185
15320	NM_016326	LOC51192
15394	AB028950	TLN1
15495	NM_014928	KIAA1046
15557	Contig39195	0
15601	NM_015681	B9
15795	Contig45024	0
15966	Contig25945	0
16195	Contig21936	0
16214	NM_016444	ZNF226
16260	Contig40988	DJ102H19.4
16340	Contig33224	0
16398	Contig25057	0
16827	AL137690	DKFZp434O032
17187	AK000822	DKFZP564M182
17319	Contig41536	FLJ21988
17336	AL137717	0
17493	NM_018061	FLJ10330
17622	NM_016641	MIR16
17661	NM_015929	LOC51601
17745	Contig57662	0
17762	NM_015978	LOC51086
17799	NM_015996	LOC51092
17806	Contig52940	0
17945	Contig713_R	SF3B1
17963	Contig40670	0
18143	Contig31656	0
18331	NM_017456	PSCD1
18405	Contig48402	0
18539	AF161451	HSPC333
18700	AF007128	0
19947	NM_017685	FLJ20139
20206	Contig41392	0
20257	NM_019109	HMT-1
20434	NM_017705	FLJ20190
20442	NM_017708	FLJ20200
20595	NM_018463	MDS028
20693	AL049985	0
20705	NM_018486	HDAC8
20870	AF195883	GBL
20891	BE739817_R	IFNAR1
21045	AL050071	DKFZP566B0
21107	Contig45343	0
21199	NM_001130	AES
21591	Contig45212	0
21630	Contig43183	0
21963	AB014543	KIAA0643
22107	NM_000523	HOXD13
22213	NM_000540	RYR1
22547	Contig28393	0
22565	AL050201	DKFZP586E1
22614	Contig26685	0
22635	Contig35802	0
22680	Contig20391	0
22691	NM_002019	FLT1
22753	Contig37236	0
22809	Contig16250	0
22836	NM_002066	GML
23084	Contig52414	MIG-6
23112	Contig53736	0
23124	Contig36735	0
23249	Contig57957	MAP3K13
23286	Contig28952	0
23322	Contig32002	0
23386	Contig52993	0
23415	Contig44105	FLJ13164
23631	NM_019597	HNRPH2

23676	NM_000765	CYP3A7	
23689	Contig50059	FLJ21108	
23876	Contig56716		0
23889	Contig22253	FLJ21062	
23941	Contig49785	CG1I	
23951	AF009267		0
24138	Contig42746		0

3.2 Δεύτερη γονιδιακή υπογραφή

Η δεύτερη γονιδιακή υπογραφή είναι το αποτέλεσμα της εφαρμογής του αλγορίθμου LASSO πάνω στο dataset. Για την επίλυση του ℓ_1 - regularized least squares προβλήματος χρησιμοποιούμε ένα απλό Matlab solver [19]. Όπως έχουμε αναφέρει και πιο πάνω ψάχνουμε τα γονίδια x του μοντέλου (2.71), όπου A είναι ο πίνακας ο οποίος περιέχει τα δεδομένα, δηλαδή samples $\times$ genes και y τα targets του εκάστωτε sample.

Τρέχουμε τον αλγόριθμο αρκετές φορές πάνω στο train set και με τη χρήση SVM ελέγχουμε το classification accuracy στο test set. Ο λόγος για τον οποίο εκτελούμε πολλές φορές τον αλγόριθμο είναι ότι θέλουμε να βρούμε ένα λ το οποίο να μας δίνει μικρό μη μηδενικό αριθμό γονιδίων για τον οποίο το ποσοστό επιτυχίας του SVM να είναι ικανοποιητικό. Καλήγουμε λοιπόν ότι για $\lambda = 0.15$ επιτυγχάνεται ένα καλό ποσοστό επιτυχίας, λίγο κάτω από 74%, για 82 γονίδια. Ο αριθμός των γονιδίων είναι κατά πολύ μικρότερος από αυτόν που προέκυψε από την πρώτη γονιδιακή υπογραφή, βέβαια και το ποσοστό επιτυχίας είναι μικρότερο. Αν συγκρίνουμε όμως το ποσοστό επιτυχίας της δευτερης υπογραφής με τα αρχικά γονίδια παρατηρούμε ότι η διαφορά δεν είναι ιδιαίτερα μεγάλη. Ακολουθεί γραφική παράσταση 3.3 η οποία μας δείχνει το ποσοστό επιτυχίας για τα γονίδια στα οποία καταλήξαμε με την εφαρμογή του Lasso για $\lambda = 0.15$.



Σχήμα 3.3: Ποσοστό επιτυχίας για τα γονίδια που προέκυψαν με τη μέθοδο LASSO.

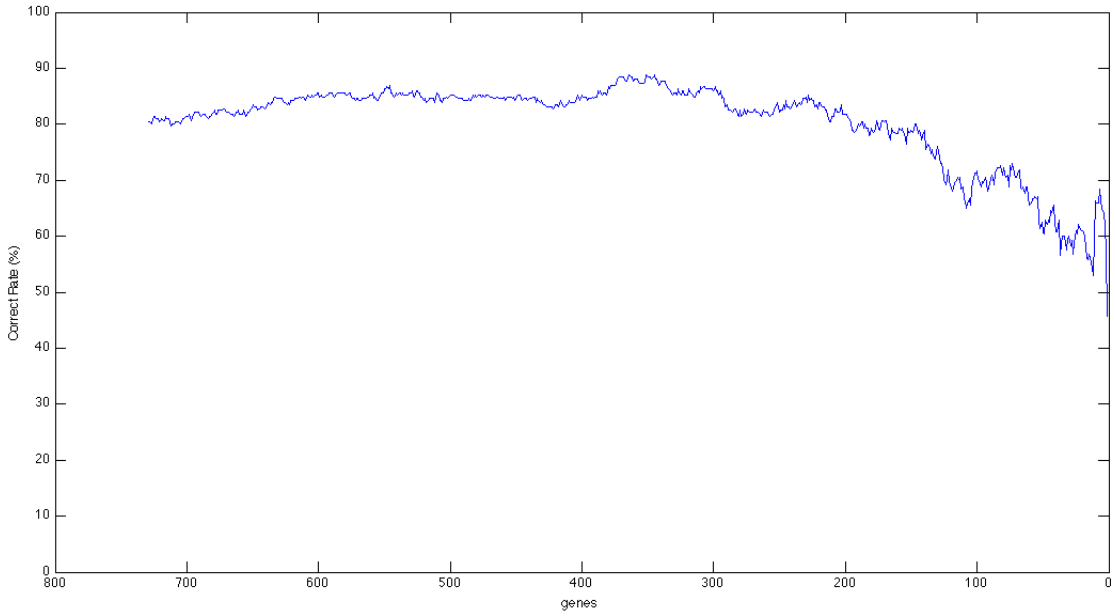
Παρακάτω φαίνονται τα 82 γονίδια από την εφαρμογή του αλγορίθμου LASSO πάνω στο dataset.

GeneSeq	GeneID	Gene Symbol
1037	Z24725	MIG2
1242	NM_001715	BLK
1552	Contig35752	0
1747	Contig43133	0
2391	Contig38834	FLJ14084
3041	NM_001981	EPS15
3153	Contig31164	0
3224	NM_020120	UGCGL1
4108	Contig31122	0
5113	AF000560	0
5293	NM_003607	PK428
5314	AB032976	KIAA1150
5377	Contig54742	0
5527	Contig9446_	0
5628	NM_003676	DEGS
6243	AL117568	0
6471	Contig22645	0
6862	Contig30410	0
6895	AL117606	0
7336	AF174605	FBXO25
7777	Contig26816	0
8257	NM_020680	NTKL
8525	Contig25706	0
8910	NM_013438	UBQLN1
9100	NM_004733	ACATN
9281	Contig39469	0
9380	Contig1403_	0
10075	AA555029_R	0
10078	AB029013	WHSC1
10657	Contig17103	0
10776	Contig49632	FGF11
10786	NM_004975	KCNB1
10983	AF038185	0
11010	Contig7976_	0
11769	Contig51888	0
11964	AI127067_RC	0
12175	NM_006523	XPNPEPL
12222	NM_014554	SENP1
12259	NM_006544	SEC10L1
12278	NM_007279	U2AF65
12405	D86971	KIAA0217
13609	AF039196	HR
13857	NM_006702	NTE
13874	NM_014726	KIAA0775
14226	AF070543	0
14294	Contig56137	0
14544	Contig27001	0
15795	Contig45024	0
16102	Contig15992	0
16264	Contig4825_	PP1201
16726	Contig56007	DDXBP1
16794	Contig48355	0
16963	Contig48097	0
17298	AK000895	FLJ10033
17359	NM_018011	FLJ10154
17963	Contig40670	0
18164	NM_017415	KLHL3
18736	AF007151	MMS19L
18955	NM_000142	FGFR3
19123	Contig45685	0
19394	AL122040	0
20026	Contig32810	0
20253	Contig5822_	0
20794	Contig26537	0
20891	BE739817_R	IFNAR1
20982	Contig34792	0
21025	Contig48006	0
21501	NM_001198	PRDM1
21714	AF072928	MTMR6
22090	AB014578	KIAA0678
22448	NM_017990	FLJ10079
22565	AL050201	DKFZP586E19
22643	Contig33967	0
22691	NM_002019	FLT1
23047	Contig53151	0
23084	Contig52414	MIG-6
23322	Contig32002	0
23570	Contig44143	FLJ20080
23903	Contig9736_	0
23917	AL157489	0
23951	AF009267	0
24019	NM_002225	IVD

3.3 Τρίτη γονιδιακή υπογραφή

Η τρίτη γονιδιακή υπογραφή είναι αποτέλεσμα του συνδυασμού του αλγορίθμου RFE-LNW με ένα Multilayer Perceptron νευρωνικό δίκτυο με gate opening. Επειδή τα αρχικά μας γονίδια ήταν πάρα πολλά και αν τα χρησιμοποιούσαμε ως έχουν, η εκπαίδευση του νευρωνικού μας δικτύου θα καθυστερούσε πολύ, κάναμε μία ελάττωση στα γονίδια τα οποία χρησιμοποιήσαμε ως είσοδο στο νευρωνικό δίκτυο. Ο περιορισμός αυτός έγινε με τη χρήση του αλγορίθμου RFE-LNW. Η μόνη διαφορά στον τρόπο εφαρμογής του RFE-LNW συγκριτικά με τη χρήση του RFE-LNW για την πρώτη υπογραφή είναι ότι δε χρησιμοποιήσαμε τόσο το train όσο και το test set αλλά μόνο το train. Ο λόγος είναι ότι το test set το χρησιμοποιήσαμε στην συνέχεια για να ελέγξουμε τα γονίδια τα οποία μας έβγαλε το νευρωνικό δίκτυο. Έτσι χρησιμοποιήσαμε μόνο το train set με τη χρήση 10-fold cross validation, στο οποίο έχουμε αναφερθεί σε παραπάνω κεφάλαιο, για τον έλεγχο του πόσο καλά έτρεξε ο αλγόριθμός μας. Με αυτό τον τρόπο κρατήσαμε 1500 γονίδια. Θεωρούμε ότι ο αριθμός των 1500 γονιδίων είναι ικανοποιητικός καθώς δεν είναι μεγάλος συγκριτικά με τον αρχικό αριθμό των γονιδίων ώστε να καθυστερήσει την εκπαίδευση του νευρωνικού στη συνέχεα, αλλά ούτε και μικρός ώστε να χαθούν γονίδια τα οποία θεωρούνται σημαντικά.

Στη συνέχεια αυτά τα 1500 γονίδια τα βάζουμε σαν είσοδο σε ένα multilayer perceptron νευρωνικό δίκτυο με gate opening όπως αυτό έχει περιγραφεί σε πιο πάνω ενότητα. Το νευρωνικό δίκτυο έχει 1500 κόμβους στο επίπεδο εισόδου, 150 κόμβους στο κρυφό επίπεδο και 1 έξοδο. Στο συγκεκριμένο στάδιο θέλήσαμε να βγάλουμε μια γονιδιακή υπογραφή ανάλογου μεγέθους με τις δύο προηγούμενες. Δεν έγινε προσπάθεια βελτιστοποίησης του νευρωνικού δικτύου καθώς το θέμα δε είναι να βρούμε ένα βέλτιστο δίκτυο, αλλά ένα σύνολο από χρήσιμα γονίδια. Και η χρησιμότητα των γονιδίων δεν εξαρτάται τόσο από το μέγεθος του δικτύου οπότε η επιλογή των κόμβων του κρυφού επιπέδου δεν είναι τόσο σημαντική, παρόλα αυτά έγιναν αρκετές δοκιμές ώστε να καταλήξουμε σε 150 κόμβους. Λόγω της τυχαιότητας των νευρωνικών δικτύων, η εκπαίδευση του νευρωνικού δικτύου είναι σημαντικό να γίνει αρκετές φορές καθώς διαφορετικά γονίδα μπορεί να επιλεγούν κάθε φορά. Εκπαιδεύουμε λοιπόν το νευρωνικό δίκτυο με το train set ώστε σε κάθε εκτέλεση να βρούμε για ποιά γονίδια παρουσιάζετα το μεγαλύτερο gate opening. Η εκπαίδευση του νευρωνικού δικτύου ολοκληρώνεται είτε όταν το σφάλμα ταξινόμησης φτάσει το μηδεν, είτε όταν ξεπεράσουμε ένα συγκεκριμένο αριθμό επαναλήψεων, στην προκειμένη περίπτωση 5000. Στη συνέχεια εκπαιδεύουμε ένα SVM classifier με τα γονίδια του train set και ελέγχουμε το ποσοστό επιτυχίας με το test set. Η γραφική παράσταση 3.4 που ακολουθεί μας δείχνει κατά μέσο όρο το ποσοστό επιτυχίας του SVM classification. Καθώς προχωράμε από αριστερά προς τα δεξιά το gate opening τω γονιδίων αυξάνεται.



Σχήμα 3.4: Ποσοστό επιτυχίας για τα γονίδια που προέκυψαν με τη μέθοδο RFE-LNW και FSMLP.

Παρατηρούμε λοιπόν ότι στα 200 περίπου γονίδια παρουσιάζεται ένα καλό ποσοστό επιτυχίας, μεταξύ 80-85%. Από τα 200 αυτά γονίδια κοιτάμε ποια εμφανίζονται κατά μέσο όρο πάνω από 90% στις διαφορές εκτελέσεις του αλγορίθμου και καταλήγουμε σε μι γονιδιακή υπογραφή με 152 γονίδια. Τα γονίδια της τρίτης γονιδιακής υπογραφής ακολουθούν παρακάτω:

GeneSeq	GeneID	Gene Symbol
21	AF155648	LOC55818
39	Contig55949	0
367	Contig54839	0
466	Contig42593	0
987	Contig46316	0
1077	Contig35025	0
1283	NM_001724	BPGM
1356	NM_019859	HTR7
1426	Contig50669	KIF9
1545	Contig51291	0
1989	NM_001822	CHN1
2005	NM_001827	CKS2
2007	NM_001828	CLC
2062	Contig26610	0
2223	NM_001894	CSNK1E
2528	NM_003315	DNAJC7
2560	AF009674	AXIN1
2569	NM_004052	BNIP3
2663	NM_003344	UBE2H
3089	U23028	EIF2B5
3198	Contig62964	0
3224	NM_020120	UGCGL1
3236	NM_020125	SBBI42
3352	NM_020156	C1GALT1
3531	Contig2384_	0
3701	NM_003479	PTP4A2
3725	NM_002754	MAPK13
3734	NM_003488	AKAP1
4219	NM_012256	ZNF212
4274	Contig47004	0
4294	NM_003525	H2BFK
4574	Contig31449	0
4676	Contig49541	FLJ21901
4848	NM_021019	MYL6
4899	AL117418	DKFZp564G2263
5113	AF000560	0
5326	NM_003611	OFD1
5406	NM_003627	POV1
5440	Contig47042	FLJ20069
5547	NM_004380	CREBBP
6067	Contig33343	0
6303	NM_021198	NLI-IF
6384	AL117595	GIT1
6386	NM_004443	EPHB3
6668	Contig2262_	0
6815	Contig31839	0
7081	NM_004504	HRB
7358	NM_004562	PARK2
7500	Contig54029	0
7618	Contig33394	0
8060	Contig21852	0
8151	NM_005354	JUND
8365	NM_005393	PLXNB3
9077	NM_004729	ALTE
9240	Contig39858	0
9346	AI401061_RC	0
9550	NM_005507	CFL1
9610	NM_006251	PRKAA1
9760	NM_006288	THY1
9789	NM_005566	LDHA
10101	Contig54274	0
10183	Contig37858	0
10209	NM_014315	LCP
10362	Contig36879	0
10649	NM_005667	ZFP103
11042	AL110133	0
11498	NM_005737	ARL7
11530	NM_005741	ZNF263
11904	AB011115	KIAA0543
11905	AB011116	KIAA0544
12063	NM_014515	CNOT2
12222	NM_014554	SENP1
12465	NM_005861	STUB1
12522	Contig57805	TUSP
12577	NM_005899	M17S2
12680	U82987	BBC3
13019	NM_015368	PANX1
13760	NM_015417	DKFZP434I11
13801	NM_014700	KIAA0665
13925	NM_014739	KIAA0164
13994	Contig32036	0
14043	NM_014779	KIAA0669
14382	Contig56030	0
14558	NM_016243	LOC51706
14671	NM_006804	MLN64
14681	NM_016285	LOC51717
14822	NM_014863	BRAG
14850	NM_006854	KDELR2
15040	AK000543	DKFZP586M1
15041	U50536	0
15080	Contig19803	0

15327	Contig51740	DOCK1	
15545	NM_016398	CNOT2	
15835	Contig569_R		0
15981	AB037721	KIAA1300	
15996	AL049340		0
16161	Contig31647		0
16191	Contig20372		0
16264	Contig4825_	PP1201	
16329	Contig39877		0
16789	AB037853	KIAA1432	
16958	Contig33680		0
17169	Contig53806	FKBP5	
17179	AA470152_R		0
17336	AL137717		0
17346	Contig382_R	NLI-IF	
17436	Contig10688		0
17452	NM_018045	FLJ10276	
17674	NM_015932	HSPC014	
17799	NM_015996	LOC51092	
18152	Contig26659		0
18167	NM_018146	FLJ10581	
18358	Contig1804_		0
18752	Contig37188		0
18943	NM_018257	FLJ10883	
18953	Contig41198		0
19388	Contig1254_		0
19473	NM_019028	FLJ10852	
19565	NM_019057	FLJ10404	
19667	Contig52463		0
19906	NM_016946	JAM1	
19917	Contig45901		0
20386	NM_001038	SCNN1A	
20524	NM_001065	TNFRSF1A	
20655	NM_018478	HSMNP1	
20866	AF073519	SERF1A	
21109	NM_001101	ACTB	
21309	NM_017810	FLJ20417	
21548	NM_017863	FLJ20527	
21673	Contig749_R		0
21702	Contig58288		0
21908	AB014520	KIAA0620	
21923	Contig56457	TMEFF1	
21927	AB014531	KIAA0631	
22203	AB014595	CUL4B	
22213	NM_000540	RYR1	
22258	Contig39950		0
22279	NM_000559	HBG1	
22326	NM_018682	HDCMC04P	
22598	Contig16607		0
22704	AL121733	hypotheticalp	
22891	NM_002081	GPC1	
22927	Contig19865		0
23084	Contig52414	MIG-6	
23184	Contig42437	FLJ23316	
23317	AL050388	SOD2	
23688	X04526	GNB1	
23788	Contig41805		0
23810	Contig26760		0
24011	NM_002220	ITPKA	
24315	Contig49045		0
24352	Contig34477		0

3.4 Τέταρτη γονιδιακή υπογραφή

Σκοπός μας για την τέταρτη και τελευταία γονιδιακή υπογραφή ήταν να επιλέξουμε ένα πολύ μικρό αριθμό γονιδίων ο οποίος όμως θα επιτύγχανε καλό classification accuracy. Η διαδικασία την οποία ακολουθήσαμε (βλέπε σχήμα 3.6) περιλαμβάνει αρχικά feature selection με RFE-LNW και στη συνέχεια τη χρήση δύο (FSMLP) νευρωνικών δικτύων με gate opening, διαφορετικού μεγέθους το ένα από το άλλο, σε συνδυασμό με Fuzzy Clusterring (NERF c-means) ώστε να καταλήξουμε στο επιθυμητό αποτέλεσμα, δηλαδή τον μικρό αυτό αριθμό των γονιδίων. Θα μπορούσαμε να πούμε ότι η τεχνική την οποία ακολουθήσαμε είναι συνέχεια της τρίτης γονιδιακής υπογραφής.

Πιο συγκεκριμένα όπως και στην τρίτη υπογραφή εφαρμόζουμε τη μέθοδο RFE-LNW και κρατάμε 1500 γονίδια και στη συνέχεια εκπαιδεύουμε ένα FSMLP. Η διαφορά σε σχέση με την προηγούμενη υπογραφή είναι ότι δε θα κρατήσουμε 200 γονίδια αλλά μόλις 25. Ο λόγος για τον οποίο επιλέξαμε 25 γονίδια είναι ότι αυτά είχαν αρκετά μεγαλύτερο gate opening συγκριτικά με τα υπόλοιπα, κάτι το οποίο μας δείχνει ότι σε αυτά τα γονίδια υπάρχει περισσότερη πληροφορία συγκριτικά με τα υπόλοιπα.

Στη συνέχεια χρησιμοποιούμε αυτά τα 25 γονίδια ώστε να εκπαιδεύσουμε ένα νέο FSMLP νευρωνικό δίκτυο. Το τελευταίο έχει 25 κόμβους στο επίπεδο εισόδου, 5 κόμβους στο κρυφό επίπεδο και 1 έξοδο. Η διαδικασία εκπαίδευσης είναι παρόμοια με αυτή του προηγούμενου νευρωνικού δικτύου, δηλαδή η εκπαίδευση ολοκληρώνεται είτε όταν το σφάλμα ταξινόμησης φτάσει το μηδεν, είτε όταν ξεπεράσουμε ένα συγκεκριμένο αριθμό επαναλήψεων, στην προκειμένη περίπτωση 5000. Λόγω του μικρότερου μεγέθους του νευρωνικού δικτύου η ταξινόμηση είναι σαφώς ταχύτερη συγκριτικά με το προηγούμενο νευρωνικό δίκτυο. Από αυτή τη διαδικασία κρατάμε τα 10 γονίδια τα οποία παρουσιάζουν το μεγαλύτερο gate opening.

Στα 25 γονίδια τα οποία προκύπτουν από το πρώτο στάδιο του FSMLP, εφαρμόζουμε τον non-Euclidean relational fuzzy c-means αλγόριθμο ο οποίος βρίσκεται στο 2.3. Σε σχέση με το hard clustering ο αλγόριθμος NERFCM δείχνει το βαθμό συμμετοχής των εκάστωτε δεδομένων σε κάθε cluster. Διαφορετικές εκτελέσεις του NERFCM μπορεί να οδηγήσουν σε διαφορετικά partiotions τα οποία χρησιμοποιούντα για τον έλεγχο της συνοχής του cluster. Είναι σημαντικό να αναφέρουμε ότι εφαρμόζουμε το NERFCM στις correlation coefficient τιμές μεταξύ ζευγαριών διανυσμάτων γονιδίων. Ύστερα από αρκετές δοκιμές καταλήξαμε στη χρήση $c = 6$ clusters.

Συνδυάζοντας τώρα τα δεδομένα που μας έδωσε το δεύτερο στάδιο του FSMLP (δηλαδή τα 10 γονίδια) με τα αποτελέσματα από την ανάλυση σε clusters καταλήγουμε σε ένα μικρό αριθμό γονιδίων, συγκεκριμένα 6. Η λογική που ακολουθούμε είναι η ακόλουθη: Αν σε κάποιο cluster δεν ανήκει κάποιο από τα δέκα γονίδια με το μεγαλύτερο gate opening, τότε επιλέγουμε το γονίδιο με το μεγαλύτερο gate opening. Στην περίπτωση που περισσότερα από ένα μέλη κάποιου cluster ανήκουν στα top 10 γονίδια και παράλληλα τα γονίδια έχουν πολύ ψηλές τιμές τότε τα κρατάμε όλα. Για να βρούμε τώρα το ποσοστό επιτυχίας εκπαιδεύουμε ένα γραμμικό SVM με το train set και ελέγχουμε το ποσοστό ταξινόμησης με το test set. Το ποσοστό επιτυχίας το οποίο επιτυγχάνεται με τη χρήση

των 6 γονιδίων στα οποία καταλήξαμε είναι πάνω από 89%. Ποσοστό ιδιαίτερα ψηλό, ειδικότερα αν αναλογιστούμε τον μικρό αριθμό γονιδίων (μόλις 6) της υπογραφής. Τα γονίδια στα οποία καταλήξαμε φαίνονται στο σχήμα 3.5.

GeneSeq	GeneID	Gene Symbol
1356	NM_019859	HTR7
1989	NM_001822	CHN1
2005	NM_001827	CKS2
2569	NM_004052	BNIP3
3734	NM_003488	AKAP1
15040	AK000543	DKFZP586M1523

Σχήμα 3.5: Η τέταρτη γονιδιακή υπογραφή.

3.5 Σύγκριση αποτελεσμάτων

3.5.1 Στατιστικά αποτελέσματα

Μελετώντας τα αποτελέσματα στα οποία καταλήξαμε φτάνουμε στο συμπέρασμα ότι η κάθε μέθοδος οδηγεί σε διαφορετικά γονίδια με διαφορετικά ποσοστά επιτυχίας. Συνοψίζοντας λοιπόν το μεγαλύτερο ποσοστό επιτυχίας επιτυγχάνεται στην τέταρτη υπογραφή παρότι τα γονίδια για τα οποία επιτυγχάνεται κάτι τέτοιο είναι μόλις 6. Ο πίνακας 3.5.1 συνοψίζει τα ποσοστά επιτυχίας της κάθε υπογραφής.

Πίνακας 3.1: Συγκριτικός πίνακας γονιδιακών υπογραφών

Μέθοδος	**Αριθμός Γονιδίων**	**Ποσοστό επιτυχίας**
RFE-LNW	190	82-83%
LASSO	82	74%
RFE-LNW, FSMLP	152	80-85%
RFE-LNW, FSMLP, NERFCM	6	89%

Κάτι άλλο που παρουσιάζει ενδιαφέρον, είναι ποιά από τα γονίδια εμφανίζονται σε παραπάνω από μία υπογραφές. Δηλαδή ποιές υπογραφές έχουν κοινά γονίδια. Τα κοινά γονίδια της πρώτης μεθόδου (RFE-LNW) με τη δεύτερη μέθοδο (LASSO) είναι 19 και φαίνονται στο σχήμα 3.7. Τα κοινά γονίδια μεταξύ της πρώτης (RFE-LNW) και της τρίτης (RFE-LNW, FSMLP) μεθόδου είναι 5 και φαίνονται στο σχήμα 3.8. Κοινά γονίδια ανάμεσα στην πρώτη (RFE-LNW) και στην τέταρτη (RFE-LNW, FSMLP, NERFCM) μέθοδο δεν υπάρχουν. Τα κοινά γονίδια της δεύτερης (LASSO) και της τρίτης (RFE-LNW, FSMLP) μεθόδου είναι 5 και φαίνονται στο σχήμα 3.9. Κοινά γονίδια ανάμεσα στη δεύτερη (LASSO) και στην τέταρτη (RFE-LNW, FSMLP, NERFCM) μέθοδο δεν υπάρχουν. Τέλος όλα τα γονίδια που προκύπτουν από την τέταρτη (RFE-LNW, FSMLP, NERFCM) μέθοδο εμφανίζονται και στην τρίτη (RFE-LNW, FSMLP) μέθοδο. Αυτά τα γονίδια φαίνονται στο σχήμα 3.5.

Ένα γονίδιο εμφανίζεται στην πρώτη, στη δεύτερη αλλά και στην τρίτη γονιδιακή υπογραφή. Είναι το γονίδιο με Gene Sequence: 23084, Gene ID: Contig52414_RC και Gene Symbol: MIG-6. Ένας λόγος για τον οποίο περισσότερα κοινά γονίδια εμφανίζονται μεταξύ των τριών πρώτων υπογραφών και όχι με την τέταρτη υπογραφή είναι ότι στις τρεις πρώτες υπογραφές ο αριθμός των γονιδίων είναι αρκετά μεγάλος οπότε είναι πολύ πιθανόν γονίδια να συμπίπτουν. Αντίθετα η τέταρτη υπογραφή έχει κοινά γονίδια μόνο με την τρίτη καθώς και οι δύο ξεκινούν από μία κοινή βάση δηλαδή από 1500 ίδια γονίδια.

3.5.2 Βιολογικά αποτελέσματα

Το genotator (που περιγράψαμε σε πιο πάνω ενότητα) είναι ένα πολύ χρήσιμο εργαλείο για να εξετάσουμε κατά πόσο τα δεδομένα μας σχετίζονται με κάποια ασθένεια. Συγκρίναμε λοιπόν τα αποτελέσματα των γονιδιακών μας υπογραφών με τη λίστα που μας δίνει το genotator για τα γονίδια που σχετίζονται με καρκίνο του μαστού.

- Για την πρώτη υπογραφή (RFE-LNW) τα γονίδια τα οποία είναι γνωστά είναι 125 από τα 190 γονίδια της υπογραφής. Από αυτά, 29 γονίδια σχετίζονται με τον καρκίνο του μαστού. Ποσοστό 23.2% εκ των γνωστών γονιδίων δηλαδή.

- Για την δεύτερη υπογραφή (LASSO) τα γονίδια τα οποία είναι γνωστά είναι 44 από τα 82 γονίδια της υπογραφής. Από αυτά 12, γονίδια σχετίζονται με τον καρκίνο του μαστού. Ποσοστό 27.27% εκ των γνωστών γονιδίων δηλαδή.

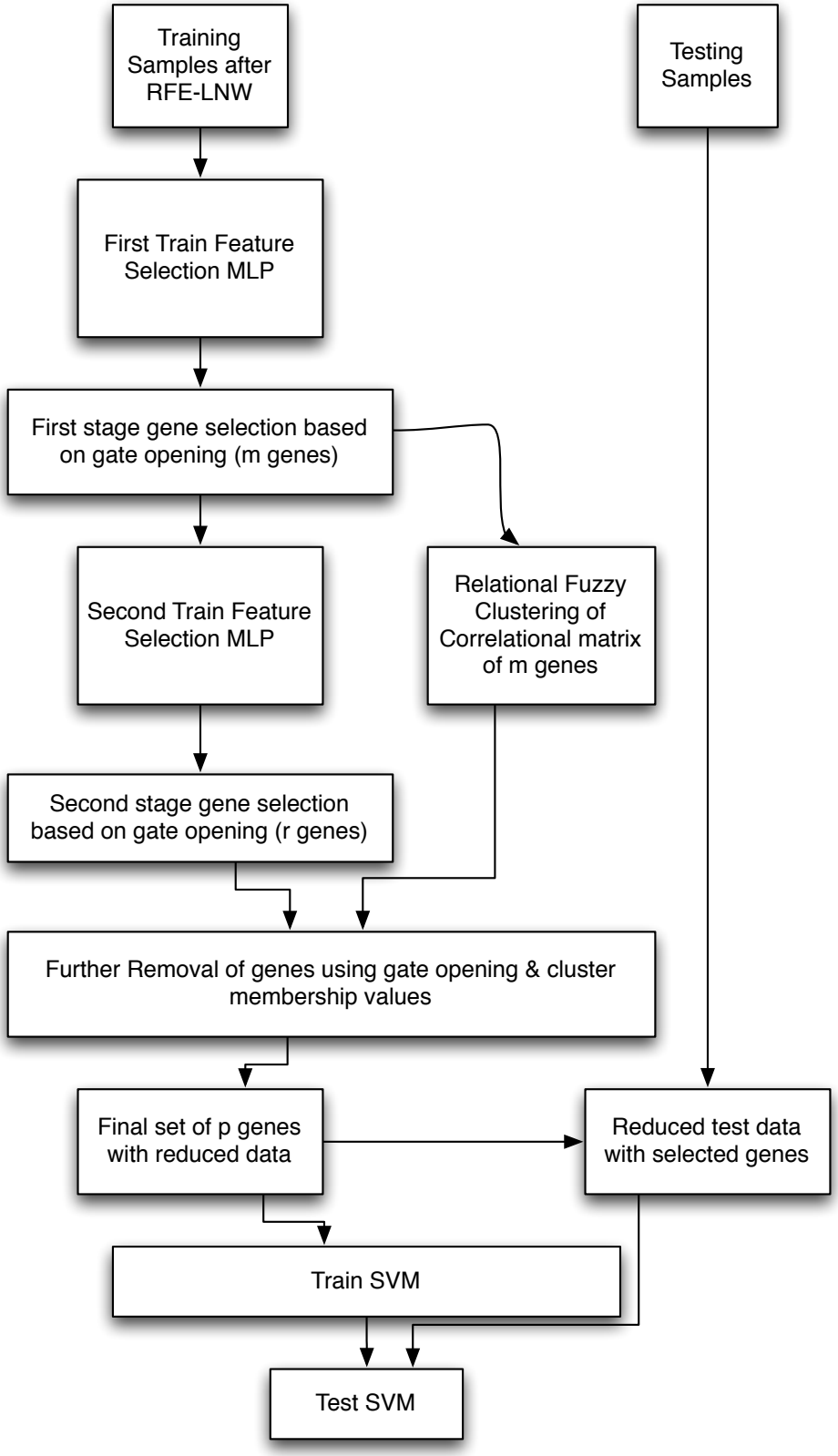
- Για την τρίτη υπογραφή (RFE-LNW, FSMLP) τα γονίδια τα οποία είναι γνωστά είναι 102 από τα 152 γονίδια της υπογραφής. Από αυτά, 39 γονίδια σχετίζονται με τον καρκίνο του μαστού. Ποσοστό 38.24% εκ των γνωστών γονιδίων δηλαδή.

- Για την τέταρτη υπογραφή (RFE-LNW, FSMLP, NERFCM) τα γονίδια τα οποία είναι γνωστά είναι 6 από τα 6 γονίδια της υπογραφής. Από αυτά, 2 γονίδια σχετίζονται με τον καρκίνο του μαστού. Ποσοστό 33.33% εκ των γνωστών γονιδίων δηλαδή.

Ο πίνακας 3.5.2 συνοψίζει τα παραπάνω αποτελέσματα.

Πίνακας 3.2: Γονίδια στο Genotator .

Μέθοδος	**Αριθμός Γονιδίων**	**Γνωστά Γονίδια**	**Σχετιζόμενα με BC**	**Ποσοστό**
1η	190	125	29	23.2%
2η	82	44	12	27.27%
3η	152	102	39	38.24%
4η	6	6	2	33.33%



Σχήμα 3.6: Η μεθοδολογία που ακολουθήσαμε για να καταλήξουμε στην τέταρτη γονιδιακή υπογραφή.

GeneSeq	GeneID	Gene Symbol
1242	NM_001715	BLK
3153	Contig31164_RC	0
5377	Contig54742_RC	0
6895	AL117606	0
7777	Contig26816_RC	0
8910	NM_013438	UBQLN1
9380	Contig1403_RC	0
11010	Contig7976_RC	0
12175	NM_006523	XPNPEPL
12259	NM_006544	SEC10L1
12405	D86971	KIAA0217
15795	Contig45024_RC	0
17963	Contig40670_RC	0
20891	BE739817_RC	IFNAR1
22565	AL050201	DKFZP586E1923
22691	NM_002019	FLT1
23084	Contig52414_RC	MIG-6
23322	Contig32002_RC	0
23951	AF009267	0

Σχήμα 3.7: Κοινά γονίδια πρώτης και δεύτερης μεθόδου.

GeneSeq	GeneID	Gene Symbol
2663	NM_003344	UBE2H
17336	AL137717	0
17799	NM_015996	LOC51092
22213	NM_000540	RYR1
23084	Contig52414_RC	MIG-6

Σχήμα 3.8: Κοινά γονίδια πρώτης και τρίτης μεθόδου.

GeneSeq	GeneID	Gene Symbol
3224	NM_020120	UGCGL1
5113	AF000560	0
12222	NM_014554	SENP1
16264	Contig4825_RC	PP1201
23084	Contig52414_RC	MIG-6

Σχήμα 3.9: Κοινά γονίδια δεύτερης και τρίτης μεθόδου.

Κεφάλαιο 4

Συμπεράσματα και μελλοντική εργασία

4.1 Συμπεράσματα

Στην συγκεκριμένη εργασία μελετήσαμε διάφορες μοντέρνες τεχνικές για feature selection. Η κάθε τεχνική συνδύαζε ένα αριθμό αλγορίθμων και κατέληξαμε σε διαφορετικά αποτελέσματα ανάλογα με την εκάστωτε μεθοδολογία.

Κάτι το οποίο παρουσιάζει ιδιαίτερο ενδιαφέρον είναι η σχέση μεταξύ των στατιστικών και των βιολογικών αποτελεσμάτων. Δεν είναι απαραίτητο ότι όσο αυξάνεται το ποσοστό επιτυχίας του classification αυξάνεται και το ποσοστό των γονιδίων που σχετίζονται με καρκίνο του μαστού (βλέπε πίνακα 4.1). Ούτε για παρόμοια ποσοστά επιτυχίας στην ταξινόμηση (βλέπε πρώτη και τρίτη γονιδιακή υπογραφή) υπάρχει μεγάλη διαφορά μεταξύ των ποσοστών των γονιδίων που σχετίζονται με καρκίνο του μαστού. Κάτι τέτοιο μας δείχνει ότι δεν είναι απαραίτητο να υπάρχει άμεση συσχέτιση.

Πίνακας 4.1: Συγκριτικός πίνακας ποσοστού επιτυχίας ταξινόμησης και του ποσοστού των γνωστών γονιδίων σχετιζόμενο με καρκίνο του μαστού

Μέθοδος	**Ποσοστό ταξινόμησης**	**Ποσοστό σχετιζόμενο με BC**
RFE-LNW	82-83%	23.2%
LASSO	74%	27.27%
RFE-LNW, FSMLP	80-85%	38.24%
RFE-LNW, FSMLP, NERFCM	89%	33.33%

Η τέταρτη υπογραφή είναι ένα σύνολο από 6 μόλις γονίδια που παρουσιάζει ιδιαίτερα υψηλό ποσοστό επιτυχίας στο classification και τα γονίδιά της έχουν σχέση με τον καρκίνο του μαστού. Το μόνο πρόβλημα σε αυτή την περίπτωση είναι ότι δεν είναι εύκολο να εντοπιστούν κοινά γονίδια με άλλες υπογραφές, όπως έγινε στην περίπτωση των τριών άλλων υπογραφών όπου βρήκαμε ένα γονίδιο το οποίο ήταν κοινό και στις τρεις. Το συγκεκριμένο μάλιστα γονίδιο είναι ιδιαίτερα ενδιαφέρον καθώς σχετίζεται με τον καρκίνο του μαστού αλλά δεν έχει μελετηθεί τόσο όσο άλλα γνωστά γονίδια.

4.2 Μελλοντική εργασία

Στο κεφάλαιο 2 αναφερθήκαμε στο genotator, και πιο συγκεκριμένα στον αλγόριθμο με τον οποίο ταξινομεί τα γονίδια ανάλογα με τη σχετικότητά τους με την ασθένεια. Στη συγκεκριμένη εργασία δεν εξετάσαμε καθόλου τις συγκεκριμένες τιμές (GS - Genotator Scores -). Σε κάποια μελλοντική εργασία καλό θα ήταν να δοθεί βαρύτητα στις συγκεκριμένες τιμές. Παράδειγμα να εξεταστεί το αν μεγάλα genotator scores οδηγούν σε καλό classification των δεδομένων.

Επιπρόσθετα ενδιαφέρον θα ήταν να εξεταστεί το συγκεκριμένο dataset με κάποιο διαφορετικό κριτήριο από το SVM ώστε να γίνει σύγκριση με τα αποτελέσματα στα οποία καταλήξαμε. Ακόμη οι συγκεκριμένες τεχνικές μπορούν να εφαρμοστούν και σε άλλα dataset με διαφορετικούς τύπους καρκίνων.

Παράρτημα Α΄

Παράρτημα

Ακολουλουθούν τα αποτελέσματα της πρώτης, της δεύτερης και της τρίτης υπογραφής αντίστοιχα. Εξετάζονται ποια γονίδια βρίσκονται στο genotator, ποιά όχι και ποιά είναι τα άγνωστα γονίδια.

Genotator	Gene Symbol	Gene Name
YES	SMARCE1	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily e, member 1
YES	STAT3	signal transducer and activator of transcription 3 (acute-phase response factor)
YES	MPG	N-methylpurine-DNA glycosylase
YES	DIABLO	diablo homolog (Drosophila)
YES	CEL	carboxyl ester lipase (bile salt-stimulated lipase)
YES	KDM4C	lysine (K)-specific demethylase 4C
YES	UBE2H	ubiquitin-conjugating enzyme E2H (UBC8 homolog, yeast)
YES	PGAM1	phosphoglycerate mutase 1 (brain)
YES	PTPN9	protein tyrosine phosphatase, non-receptor type 9
YES	NMT1	N-myristoyltransferase 1
YES	TSC2	tuberous sclerosis 2
YES	CD74	CD74 molecule, major histocompatibility complex, class II invariant chain
YES	CYP2D6	cytochrome P450, family 2, subfamily D, polypeptide 6
YES	MIIP	migration and invasion inhibitory protein
YES	CD2BP2	CD2 (cytoplasmic tail) binding protein 2
YES	TADA3	transcriptional adaptor 3
YES	TCEB2	transcription elongation factor B (SIII), polypeptide 2 (18kDa, elongin B)
YES	TSPAN32	tetraspanin 32
YES	RBCK1	RanBP-type and C3HC4-type zinc finger containing 1
YES	PIAS1	protein inhibitor of activated STAT, 1
YES	CTIF	CBP80/20-dependent translation initiation factor
YES	YWHAE	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, epsilon polypeptide
YES	BRAP	BRCA1 associated protein
YES	HDAC8	histone deacetylase 8
YES	HOXD13	homeobox D13
YES	RYR1	ryanodine receptor 1 (skeletal)
YES	FLT1	fms-related tyrosine kinase 1 (vascular endothelial growth factor/vascular permeability factor receptor)
YES	ERRFI1	ERBB receptor feedback inhibitor 1
YES	CYP3A7	cytochrome P450, family 3, subfamily A, polypeptide 7

Genotator	Gene Symbol	Gene Name
NO	BLK	B lymphoid tyrosine kinase
NO	CNTFR	ciliary neurotrophic factor receptor
NO	NPRL3	nitrogen permease regulator-like 3 (S. cerevisiae)
NO	C16orf72	chromosome 16 open reading frame 72
NO	SLC25A44	solute carrier family 25, member 44
NO	UQCRC2	ubiquinol-cytochrome c reductase core protein II
NO	PIGC	phosphatidylinositol glycan anchor biosynthesis, class C
NO	HLA-F	major histocompatibility complex, class I, F
NO	PPP4C	protein phosphatase 4, catalytic subunit
NO	ZNF174	zinc finger protein 174
NO	FZD5	frizzled homolog 5 (Drosophila)
NO	EXTL2	exostoses (multiple)-like 2
NO	NUBP2	nucleotide binding protein 2 (MinD homolog, E. coli)
NO	PSMD2	proteasome (prosome, macropain) 26S subunit, non-ATPase, 2
NO	EDC3	enhancer of mRNA decapping 3 homolog (S. cerevisiae)
NO	NFKBIL1	nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor-like 1
NO	ACCN3	amiloride-sensitive cation channel 3
NO	PCBP2	poly(rC) binding protein 2
NO	MAPRE3	microtubule-associated protein, RP/EB family, member 3
NO	ART1	ADP-ribosyltransferase 1
NO	ATP5G2	ATP synthase, H+ transporting, mitochondrial Fo complex, subunit C2 (subunit 9)
NO	RRAGC	Ras-related GTP binding C
NO	C9orf25	chromosome 9 open reading frame 25
NO	SNX11	sorting nexin 11
NO	SNAPC5	small nuclear RNA activating complex, polypeptide 5, 19kDa
NO	C16orf62	chromosome 16 open reading frame 62
NO	UBQLN1	ubiquilin 1
NO	RANGRF	RAN guanine nucleotide release factor
NO	SRP68	signal recognition particle 68kDa
NO	UBTF	upstream binding transcription factor, RNA polymerase I

NO	CIAO1	cytosolic iron-sulfur protein assembly 1
NO	NPTXR	neuronal pentraxin receptor
NO	BRD4	bromodomain containing 4
NO	TLN1	talin 1
NO	NACA	nascent polypeptide-associated complex alpha subunit
NO	SMG5	Smg-5 homolog, nonsense mediated mRNA decay factor (C. elegans)
NO	PKIG	protein kinase (cAMP-dependent, catalytic) inhibitor gamma
NO	OPTC	opticin
NO	SPPL3	signal peptide peptidase 3
NO	ZNF629	zinc finger protein 629
NO	ZMYM6	zinc finger, MYM-type 6
NO	DIMT1L	DIM1 dimethyladenosine transferase 1-like (S. cerevisiae)
NO	INADL	InaD-like (Drosophila)
NO	JMJD6	jumonji domain containing 6
NO	XPNPEP1	X-prolyl aminopeptidase (aminopeptidase P) 1, soluble
NO	EXOC5	exocyst complex component 5
NO	LARP4B	La ribonucleoprotein domain family, member 4B
NO	OR10C1	olfactory receptor, family 10, subfamily C, member 1
NO	CROCC	ciliary rootlet coiled-coil, rootletin
NO	SMYD5	SMYD family member 5
NO	PHF11	PHD finger protein 11
NO	SCFD1	sec1 family domain containing 1
NO	IFT140	intraflagellar transport 140 homolog (Chlamydomonas)
NO	HEMK1	HemK methyltransferase family member 1
NO	AZI2	5-azacytidine induced 2
NO	KLC4	kinesin light chain 4
NO	C9orf16	chromosome 9 open reading frame 16
NO	METAP2	methionyl aminopeptidase 2
NO	GCN1L1	GCN1 general control of amino-acid synthesis 1-like 1 (yeast)
NO	CRBN	cereblon
NO	CKLF	chemokine-like factor
NO	TLN1	talin 1
NO	OTUD4	OTU domain containing 4
NO	B9D1	B9 protein domain 1
NO	ZNF226	zinc finger protein 226
NO	C6orf164	chromosome 6 open reading frame 164

NO	RSL1D1	ribosomal L1 domain containing 1
NO	NARFL	nuclear prelamin A recognition factor-like
NO	ANKRD36BP2	ankyrin repeat domain 36B pseudogene 2
NO	PRPF38B	PRP38 pre-mRNA processing factor 38 (yeast) domain containing B
NO	GDE1	glycerophosphodiester phosphodiesterase 1
NO	LIPT1	lipoyltransferase 1
NO	TNNI3K	TNNI3 interacting kinase
NO	SIDT2	SID1 transmembrane family, member 2
NO	SF3B1	splicing factor 3b, subunit 1, 155kDa
NO	CYTH1	cytohesin 1
NO	NSMCE1	non-SMC element 1 homolog (S. cerevisiae)
NO	FLJ20139	hypothetical protein FLJ20139; Gene ID: 54833, discontinued on 10-May-2005
NO	ALG1	asparagine-linked glycosylation 1, beta-1,4-mannosyltransferase homolog (S. cerevisiae)
NO	PAQR5	progestin and adipoQ receptor family member V
NO	FAM83E	family with sequence similarity 83, member E
NO	ITFG2	integrin alpha FG-GAP repeat containing 2
NO	RAB22A	RAB22A, member RAS oncogene family
NO	MLST8	MTOR associated protein, LST8 homolog (S. cerevisiae)
NO	IFNAR1	interferon (alpha, beta and omega) receptor 1
NO	PVRL3	poliovirus receptor-related 3
NO	AES	amino-terminal enhancer of split
NO	CLUAP1	clusterin associated protein 1
NO	C6orf162	chromosome 6 open reading frame 162
NO	GML	glycosylphosphatidylinositol anchored molecule like protein
NO	MAP3K13	mitogen-activated protein kinase kinase kinase 13
NO	OSBPL11	oxysterol binding protein-like 11
NO	HNRNPH2	heterogeneous nuclear ribonucleoprotein H2 (H')
NO	CTNNBL1	catenin, beta like 1
NO	C7orf63	chromosome 7 open reading frame 63
NO	ZNHIT1	zinc finger, HIT-type containing 1

GeneID	Unknown Genes
Contig13548_RC	No annotation available; Reporter: Contig13548_RC (SPOT ID)
Contig16772_RC	No annotation available; Reporter: Contig16772_RC (SPOT ID)
Contig31164_RC	No annotation available; Reporter: Contig31164_RC (SPOT ID)
Contig42933_RC	No annotation available; Reporter: Contig42933_RC (SPOT ID)
Contig26988_RC	No annotation available; Reporter: Contig26988_RC (SPOT ID)
Contig19089_RC	The following term was not found in GEO Profiles: Contig19089_RC.
	No annotation available; Reporter: Contig53488 (SPOT ID)
Contig53488	TSA: Zea mays contig53488, mRNA sequence
Contig54742_RC	No annotation available; Reporter: Contig54742_RC (SPOT ID)
Contig54345_RC	No annotation available; Reporter: Contig54345_RC (SPOT ID)
Contig5681_RC	No annotation available; Reporter: Contig5681_RC (SPOT ID)
	LOC254413; hypothetical gene supported by AL117590/ Chromosome: 5/ID: 254413; LOC153372; hypothetical gene supported by AL117590/Chromosome: 5/Location: 5q14.1/ ID: 153372; LOC133815; hypothetical gene supported by AL117590/Chromosome: 5/Location: 5q14.1/ID: 133815; LOC90029; hypothetical gene supported by AL117590/Chromosome: 5/Location: 5q14.1/ID: 90029; LOC85808; Other Designations: hypothetical gene supported by AL117590/ XM_029084; Chromosome: 5/ID: 85808
AL117590	All records were discontinued
Contig15355_RC	The following term was not found in GEO Profiles: Contig15355_RC.
Contig29529_RC	No annotation available; Reporter: Contig29529_RC (SPOT ID)
Contig27722_RC	No annotation available; Reporter: Contig27722_RC (SPOT ID)
	LOC152441; hypothetical gene supported by AL117606/BC014110/Chromosome: 3/Location: 3q22.1/ID: 152441; LOC132242; hypothetical gene supported by AL117606/Chromosome: 3/Location: 3q22.1/ID: 132242; LOC90287; hypothetical gene supported by AL117606/Chromosome: 3/Location: 3p13-q13.33/ID: 90287; LOC87947; Other Designations: hypothetical gene supported by AL117606; XM_041630 Chromosome: 3/ID: 87947;
AL117606	All records were discontinued
Contig43621_RC	No annotation available; Reporter: Contig43621_RC (SPOT ID)
AL109698	The following term was not found in Gene: AL109698
Contig15795_RC	No annotation available; Reporter: Contig15795_RC (SPOT ID)
Contig26816_RC	No annotation available; Reporter: Contig26816_RC (SPOT ID)
Contig1403_RC	No annotation available; Reporter: Contig1403_RC (SPOT ID)
Contig51498_RC	No annotation available; Reporter: Contig51498_RC (SPOT ID)
Contig20184_RC	No annotation available; Reporter: Contig20184_RC (SPOT ID)

Contig7976_RC	No annotation available; Reporter: Contig7976_RC (SPOT ID)
Contig25546_RC	No annotation available; Reporter: Contig25546_RC (SPOT ID)
Contig33442_RC	No annotation available; Reporter: Contig33442_RC (SPOT ID)
Contig31153	No annotation available; Reporter: Contig31153 (SPOT ID)
Contig34367_RC	No annotation available; Reporter: Contig34367_RC (SPOT ID)
AK001092	The following term was not found in Gene: AK001092
Contig54606_RC	No annotation available; Reporter: Contig54606_RC (SPOT ID)
Contig13551_RC	No annotation available; Reporter: Contig13551_RC (SPOT ID)
Contig36715_RC	No annotation available; Reporter: Contig36715_RC (SPOT ID)
	LOC89439/Gene ID: 89439, discontinued on 10-May-2005
	This record was discontinued;
AL137323	LOC89439; hypothetical gene supported by AL137323; XM_049777
Contig56569_RC	No annotation available; Reporter: Contig56569_RC (SPOT ID)
Contig319	The following term was not found in Gene: Contig319
Contig39195_RC	No annotation available; Reporter: Contig39195_RC (SPOT ID)
Contig45024_RC	No annotation available; Reporter: Contig45024_RC (SPOT ID)
Contig25945_RC	No annotation available; Reporter: Contig25945_RC (SPOT ID)
Contig21936_RC	No annotation available; Reporter: Contig21936_RC (SPOT ID)
Contig33224_RC	No annotation available; Reporter: Contig33224_RC (SPOT ID)
Contig25057	No annotation available; Reporter: Contig25057 (SPOT ID)
	hypothetical gene supported by AL137690/
AL137690	Chromosome: 11/ID: 92352; This record was discontinued
Contig57662_RC	No annotation available; Reporter: Contig57662_RC (SPOT ID)
Contig52940_RC	No annotation available; Reporter: Contig52940_RC (SPOT ID)
Contig40670_RC	No annotation available; Reporter: Contig40670_RC (SPOT ID)
Contig31656_RC	No annotation available; Reporter: Contig31656_RC (SPOT ID)
Contig48402_RC	No annotation available; Reporter: Contig48402_RC (SPOT ID)
AF007128	The following term was not found in Gene: AF007128
Contig41392_RC	No annotation available; Reporter: Contig41392_RC (SPOT ID)
Contig45343_RC	No annotation available; Reporter: Contig45343_RC (SPOT ID)
Contig45212_RC	No annotation available; Reporter: Contig45212_RC (SPOT ID)
Contig43183	No annotation available; Reporter: Contig43183 (SPOT ID)
Contig28393_RC	No annotation available; Reporter: Contig28393_RC (SPOT ID)
Contig26685_RC	No annotation available; Reporter: Contig26685_RC (SPOT ID)
Contig35802_RC	No annotation available; Reporter: Contig35802_RC (SPOT ID)
Contig20391_RC	No annotation available; Reporter: Contig20391_RC (SPOT ID)
Contig37236_RC	No annotation available; Reporter: Contig37236_RC (SPOT ID)
Contig16250_RC	No annotation available; Reporter: Contig16250_RC (SPOT ID)
Contig53736_RC	No annotation available; Reporter: Contig53736_RC (SPOT ID)
Contig36735_RC	No annotation available; Reporter: Contig36735_RC (SPOT ID)
Contig28952_RC	No annotation available; Reporter: Contig28952_RC (SPOT ID)
Contig32002_RC	No annotation available; Reporter: Contig32002_RC (SPOT ID)
Contig52993_RC	No annotation available; Reporter: Contig52993_RC (SPOT ID)
Contig56716_RC	No annotation available; Reporter: Contig56716_RC (SPOT ID)
AF009267	The following term was not found in Gene: AF009267
Contig42746_RC	No annotation available; Reporter: Contig42746_RC (SPOT ID)

Genotator	Gene Symbol	Gene Name
YES	FERMT2	fermitin family member 2
YES	EPS15	epidermal growth factor receptor pathway substrate 15
YES	CDC42BPA	CDC42 binding protein kinase alpha (DMPK-like)
YES	UBE2K	ubiquitin-conjugating enzyme E2K (UBC1 homolog, yeast)
YES	SCYL1	SCY1-like 1 (S. cerevisiae)
YES	WHSC1	Wolf-Hirschhorn syndrome candidate 1
YES	FGF11	fibroblast growth factor 11
YES	PIAS1	protein inhibitor of activated STAT, 1
YES	FGFR3	fibroblast growth factor receptor 3
YES	PRDM1	PR domain containing 1, with ZNF domain
YES	FLT1	fms-related tyrosine kinase 1 (vascular endothelial growth factor/vascular permeability factor receptor)
YES	ERRFI1	ERBB receptor feedback inhibitor 1

Genotator	Gene Symbol	Gene Name
NO	BLK	B lymphoid tyrosine kinase
NO	TMEM35	transmembrane protein 35
NO	UGGT1	UDP-glucose glycoprotein glucosyltransferase 1
NO	ZNF787	zinc finger protein 787
NO	GATAD2B	GATA zinc finger domain containing 2B
NO	DEGS1	degenerative spermatocyte homolog 1, lipid desaturase (Drosophila)
NO	FBXO25	F-box protein 25
NO	UBQLN1	ubiquilin 1
NO	SLC33A1	solute carrier family 33 (acetyl-CoA transporter), member 1
NO	LOC286052	Annotation: LOC286052: Hypothetical protein LOC286052
NO	KCNB1	potassium voltage-gated channel, Shab-related subfamily, member 1
NO	XPNPEP1	X-prolyl aminopeptidase (aminopeptidase P) 1, soluble
NO	SENP1	SUMO1/sentrin specific peptidase 1
NO	EXOC5	exocyst complex component 5
NO	U2AF2	U2 small nuclear RNA auxiliary factor 2
NO	LARP4B	La ribonucleoprotein domain family, member 4B
NO	HR	hairless homolog (mouse)
NO	PNPLA6	patatin-like phospholipase domain containing 6
NO	TBKBP1	TBK1 binding protein 1
NO	TMBIM1	transmembrane BAX inhibitor motif containing 1
NO	C14orf118;	chromosome 14 open reading frame 118
NO	ARGLU1	arginine and glutamate rich 1
NO	KLHL3	kelch-like 3 (Drosophila)
NO	MMS19	MMS19 nucleotide excision repair homolog (S. cerevisiae)
NO	IFNAR1	interferon (alpha, beta and omega) receptor 1
NO	MTMR6	myotubularin related protein 6
NO	DNAJC13	DnaJ (Hsp40) homolog, subfamily C, member 13
NO	PDPR	pyruvate dehydrogenase phosphatase regulatory subunit

NO	C6orf162	chromosome 6 open reading frame 162
NO	AFTPH	aftiphilin
NO	C14orf48	chromosome 14 open reading frame 48 (RNAgene)
NO	IVD	isovaleryl-CoA dehydrogenase

GeneID	Unknown Genes
Contig35752	No annotation available; Reporter: **Contig35752** (SPOT ID)
Contig43133_RC	No annotation available; Reporter: **Contig43133_RC** (SPOT ID)
Contig31164_RC	No annotation available; Reporter: **Contig31164_RC** (SPOT ID);
Contig31122_RC	No annotation available; Reporter: Contig31122_RC (SPOT ID)
Contig54742_RC	No annotation available; Reporter: **Contig54742_RC** (SPOT ID)
Contig9446_RC	No annotation available; Reporter: **Contig9446_RC** (SPOT ID)
Contig22645_RC	No annotation available; Reporter: **Contig22645_RC** (SPOT ID)
Contig30410_RC	No annotation available; Reporter: **Contig30410_RC** (SPOT ID)
AL117606	LOC152441/hypothetical gene supported by **AL117606;** LOC132242/hypothetical gene supported by **AL117606**/Chromosome: 3; LOC90287/hypothetical gene supported by **AL117606**/Chromosome: 3; LOC87947/LOC87947/Other Designations: hypothetical gene supported by **AL117606**; XM_041630/Chromosome: 3/ID: 87947; All records were discontinued
Contig26816_RC	No annotation available; Reporter: **Contig26816_RC** (SPOT ID)
Contig25706_RC	No annotation available; Reporter: **Contig25706_RC** (SPOT ID)
Contig39469	No annotation available; Reporter: **Contig39469** (SPOT ID)
Contig1403_RC	No annotation available; Reporter: **Contig1403_RC** (SPOT ID)
Contig17103_RC	No annotation available; Reporter: **Contig17103_RC** (SPOT ID)
AF038185	Homo sapiens clone 23700 mRNA sequence; cDNA Sources: brain; eye; lung; uncharacterized tissue; kidney; pancreas; embryonic tissue; mixed; uterus
Contig7976_RC	No annotation available; Reporter: **Contig7976_RC** (SPOT ID);
Contig51888_RC	No annotation available; Reporter: **Contig51888_RC** (SPOT ID);
AI127067_RC	Transcribed locus; cDNA Sources: heart
AF070543	Homo sapiens Clone 24740 mRNA sequence; cDNA Sources: brain
Contig56137_RC	No annotation available;Reporter: **Contig56137_RC** (SPOT ID)
Contig27001_RC	No annotation available; Reporter: **Contig27001_RC** (SPOT ID)

Contig45024_RC	No annotation available; Reporter: **Contig45024_RC** (SPOT ID)
Contig15992_RC	No annotation available; Reporter: **Contig15992_RC** (SPOT ID)
Contig48355_RC	No annotation available; Reporter: **Contig48355_RC** (SPOT ID)
Contig48097_RC	No annotation available; Reporter: **Contig48097_RC** (SPOT ID)
Contig40670_RC	No annotation available; Reporter: **Contig40670_RC** (SPOT ID)
Contig45685_RC	No annotation available; Reporter: **Contig45685_RC** (SPOT ID)
AL122040	Homo sapiens mRNA; cDNA DKFZp434G1972 (from clone DKFZp434G1972)
Contig32810	No annotation available; Reporter: **Contig32810** (SPOT ID)
Contig5822_RC	No annotation available; Reporter: **Contig5822_RC** (SPOT ID)
Contig26537_RC	The following term was not found in GEO Profiles: **Contig26537_RC**
Contig34792_RC	No annotation available; Reporter: **Contig34792_RC** (SPOT ID)
Contig48006_RC	No annotation available; Reporter: **Contig48006_RC** (SPOT ID)
Contig33967_RC	No annotation available; Reporter: **Contig33967_RC** (SPOT ID)
Contig53151_RC	No annotation available; Reporter: **Contig53151_RC** (SPOT ID)
Contig32002_RC	No annotation available; Reporter: **Contig32002_RC** (SPOT ID)
Contig9736_RC	No annotation available; Reporter: **Contig9736_RC** (SPOT ID)
AF009267	Homo sapiens clone FBA1 Cri-du-chat region mRNA; uncharacterized tissue; pancreas; mammary gland; liver; mixed; brain; lymph node; prostate

Genotator	Gene Symbol	Gene Name
YES	CKS2	CDC28 protein kinase regulatory subunit 2
YES	CSNK1E	casein kinase 1, epsilon
YES	AXIN1	axin 1
YES	BNIP3	BCL2/adenovirus E1B 19kDa interacting protein 3
YES	UBE2H	ubiquitin-conjugating enzyme E2H (UBC8 homolog, yeast)
YES	EIF2B5	eukaryotic translation initiation factor 2B, subunit 5 epsilon, 82kDa
YES	PTP4A2	protein tyrosine phosphatase type IVA, member 2
YES	MAPK13	mitogen-activated protein kinase 13
YES	AKAP9	A kinase (PRKA) anchor protein (yotiao) 9
YES	CREBBP	CREB binding protein
YES	GIT1	G protein-coupled receptor kinase interacting ArfGAP 1
YES	EPHB3	EPH receptor B3
YES	AGFG1	ArfGAP with FG repeats 1
YES	PARK2	parkinson protein 2, E3 ubiquitin protein ligase (parkin)
YES	JUND	jun D proto-oncogene
YES	PLXNB3	plexin B3
YES	ZBED1	zinc finger, BED-type containing 1
YES	CFL1	cofilin 1 (non-muscle)
YES	PRKAA1	protein kinase, AMP-activated, alpha 1 catalytic subunit
YES	THY1	Thy-1 cell surface antigen
YES	LDHA	lactate dehydrogenase A
YES	STUB1	STIP1 homology and U-box containing protein 1, E3 ubiquitin protein ligase
YES	NBR1	neighbor of BRCA1 gene 1
YES	BBC3	BCL2 binding component 3
YES	RAB11FIP3	RAB11 family interacting protein 3 (class II)
YES	BCLAF1	BCL2-associated transcription factor 1
YES	STARD3	StAR-related lipid transfer (START) domain containing 3
YES	DOCK1	dedicator of cytokinesis 1
YES	FKBP5	FK506 binding protein 5
YES	F11R	F11 receptor
YES	TNFRSF1A	tumor necrosis factor receptor superfamily, member 1A

YES	ACTB	actin, beta
YES	TMEFF1	transmembrane protein with EGF-like and two follistatin-like domains 1
YES	RYR1	ryanodine receptor 1 (skeletal)
YES	GPC1	glypican 1
YES	ERRFI1	ERBB receptor feedback inhibitor 1
YES	CDC73	cell division cycle 73, Paf1/RNA polymerase II complex component, homolog (S. cerevisiae)
YES	SOD2	superoxide dismutase 2, mitochondrial
YES	GNB1	guanine nucleotide binding protein (G protein), beta polypeptide 1

Genotator	Gene Symbol	Gene Name
NO	KDM3A	lysine (K)-specific demethylase 3A
NO	BPGM	2,3-bisphosphoglycerate mutase
NO	HTR7	5-hydroxytryptamine (serotonin) receptor 7 (adenylate cyclase-coupled)
NO	KIF9	kinesin family member 9
NO	CHN1	chimerin (chimaerin) 1
NO	CLC	Charcot-Leyden crystal protein
NO	DNAJC7	DnaJ (Hsp40) homolog, subfamily C, member 7
NO	UGGT1	UDP-glucose glycoprotein glucosyltransferase 1
NO	SLAMF8	SLAM family member 8
NO	C1GALT1	core 1 synthase, glycoprotein-N-acetylgalactosamine 3-beta-galactosyltransferase, 1
NO	AKAP1	A kinase (PRKA) anchor protein 1
NO	ZNF212	zinc finger protein 212
NO	HIST1H2BI	histone cluster 1, H2bi
NO	FASTKD1	FAST kinase domains 1
NO	MYL6	myosin, light chain 6, alkali, smooth muscle and non-muscle
NO	ZNF787	zinc finger protein 787
NO	OFD1	oral-facial-digital syndrome 1
NO	SLC43A1	solute carrier family 43, member 1
NO	AHI1	Abelson helper integration site 1
NO	CTDSP1	CTD (carboxy-terminal domain, RNA polymerase II, polypeptide A) small phosphatase 1
NO	KLHDC2	kelch domain containing 2
NO	RNF103	ring finger protein 103
NO	ARL4C	ADP-ribosylation factor-like 4C
NO	ZNF263	zinc finger protein 263
NO	ZNF862	zinc finger protein 862
NO	MGRN1	mahogunin, ring finger 1
NO	CNOT2	CCR4-NOT transcription complex, subunit 2
NO	SENP1	SUMO1/sentrin specific peptidase 1
NO	TULP4	tubby like protein 4
NO	PANX1	pannexin 1
NO	SPEF1	sperm flagellar 1
NO	TSC22D2	TSC22 domain family, member 2
NO	CYB5R1	cytochrome b5 reductase 1
NO	KLF12	Kruppel-like factor 12
NO	CHST15	carbohydrate (N-acetylgalactosamine 4-sulfate 6-O) sulfotransferase 15

NO	KDELR2	KDEL (Lys-Asp-Glu-Leu) endoplasmic reticulum protein retention receptor 2
NO	C18orf10	chromosome 18 open reading frame 10
NO	CNOT2	CCR4-NOT transcription complex, subunit 2
NO	STARD9	StAR-related lipid transfer (START) domain containing 9
NO	YTHDC1	YTH domain containing 1
NO	TMBIM1	transmembrane BAX inhibitor motif containing 1
NO	KIAA1432	KIAA1432
NO	ANKRD36BP2	ankyrin repeat domain 36B pseudogene 2
NO	CTDSP1	CTD (carboxy-terminal domain, RNA polymerase II, polypeptide A) small phosphatase 1
NO	BSDC1	BSD domain containing 1
NO	POMP	proteasome maturation protein
NO	SIDT2	SID1 transmembrane family, member 2
NO	RNMTL1	RNA methyltransferase like 1
NO	PCMTD2	protein-L-isoaspartate (D-aspartate) O-methyltransferase domain containing 2
NO	ZDHHC13	zinc finger, DHHC-type containing 13
NO	FAM193B	family with sequence similarity 193, member B
NO	SCNN1A	sodium channel, nonvoltage-gated 1 alpha
NO	DBNDD2	dysbindin (dystrobrevin binding protein 1) domain containing 2
NO	SERF1A	small EDRK-rich factor 1A (telomeric)
NO	ZNF434	zinc finger protein 434
NO	CXorf48	chromosome X open reading frame 48
NO	PLXND1	plexin D1
NO	ACSBG1	acyl-CoA synthetase bubblegum family member 1
NO	CUL4B	cullin 4B
NO	HBG1	hemoglobin, gamma A
NO	MLL5	myeloid/lymphoid or mixed-lineage leukemia 5 (trithorax homolog, Drosophila)
NO	PHF13	PHD finger protein 13
NO	ITPKA	inositol 1,4,5-trisphosphate 3-kinase A

GeneID	Unknown Genes
Contig55949_RC	No annotation available; Reporter: **Contig55949_RC** (SPOT ID)
Contig54839_RC	No annotation available; Reporter: **Contig54839_RC** (SPOT ID)
Contig42593_RC	No annotation available; Reporter: **Contig42593_RC** (SPOT ID)
Contig46316_RC	No annotation available; Reporter: **Contig46316_RC** (SPOT ID)
Contig35025_RC	No annotation available; Reporter: **Contig35025_RC** (SPOT ID)
Contig51291_RC	No annotation available; Reporter: **Contig51291_RC** (SPOT ID)
Contig26610_RC	No annotation available; Reporter: **Contig26610_RC** (SPOT ID)
Contig62964_RC	No annotation available; Reporter: **Contig62964_RC** (SPOT ID)
Contig2384_RC	No annotation available; Reporter: **Contig2384_RC** (SPOT ID)
Contig47004_RC	No annotation available; Reporter: **Contig47004_RC** (SPOT ID)
Contig31449_RC	No annotation available; Reporter: **Contig31449_RC** (SPOT ID)
Contig33343_RC	No annotation available; Reporter: **Contig33343_RC** (SPOT ID)
Contig2262_RC	No annotation available; Reporter: **Contig2262_RC** (SPOT ID)
Contig31839_RC	The following term was not found in GEO profiles: **Contig31839_RC**.
Contig54029_RC	No annotation available; Reporter: **Contig54029_RC** (SPOT ID)
Contig33394_RC	No annotation available; Reporter: **Contig33394_RC** (SPOT ID)
Contig21852_RC	No annotation available; Reporter: **Contig21852_RC** (SPOT ID)
Contig39858_RC	No annotation available; Reporter: **Contig39858_RC** (SPOT ID)
AI401061_RC	Annotation: **Transcribed locus;** Reporter: **AI401061**
Contig54274_RC	No annotation available; Reporter: **Contig54274_RC** (SPOT ID)
Contig37858_RC	No annotation available; Reporter: **Contig37858_RC** (SPOT ID)
Contig36879_RC	No annotation available; Reporter: **Contig36879_RC** (SPOT ID)

AL110133	Annotation: MRNA; cDNA DKFZp564H072 (from clone DKFZp564H072) Reporter: **AL110133**
Contig32036	No annotation available; Reporter: **Contig32036** (SPOT ID)
Contig56030_RC	No annotation available; Reporter: **Contig56030_RC** (SPOT ID)
U50536	LOCUS: HSU50536; DEFINITION: Human BRCA2 region, mRNA sequence CG011; Annotation: BRCA2 region, mRNA sequence CG011; Reporter: **U50536**
Contig19803_RC	No annotation available; Reporter: **Contig19803_RC** (SPOT ID)
Contig569_RC	No annotation available; Reporter: **Contig569_RC** (SPOT ID)
Contig31647_RC	No annotation available; Reporter: **Contig31647_RC** (SPOT ID)
Contig20372_RC	The following term was not found in GEO profiles: **Contig20372_RC.**
Contig39877_RC	No annotation available; Reporter: **Contig39877_RC** (SPOT ID)
Contig33680_RC	No annotation available; Reporter: **Contig33680_RC** (SPOT ID)
AA470152_RC	The following term was not found in GEO profiles: **AA470152_RC.**
Contig10688_RC	No annotation available; Reporter: **Contig10688_RC** (SPOT ID)
Contig26659_RC	No annotation available; Reporter: **Contig26659_RC** (SPOT ID)
Contig1804_RC	No annotation available; Reporter: **Contig1804_RC** (SPOT ID)
Contig37188_RC	No annotation available; Reporter: **Contig37188_RC** (SPOT ID)
Contig41198_RC	No annotation available; Reporter: **Contig41198_RC** (SPOT ID)
Contig1254_RC	No annotation available; Reporter: **Contig1254_RC** (SPOT ID)
Contig52463_RC	No annotation available; Reporter: **Contig52463_RC** (SPOT ID)
Contig45901_RC	No annotation available; Reporter: **Contig45901_RC** (SPOT ID)
Contig749_RC	No annotation available; Reporter: **Contig749_RC** (SPOT ID)
Contig58288_RC	No annotation available; Reporter: **Contig58288_RC** (SPOT ID)

Contig39950_RC	No annotation available; Reporter: **Contig39950_RC** (SPOT ID)
Contig16607_RC	No annotation available; Reporter: **Contig16607_RC** (SPOT ID)
Contig19865_RC	No annotation available; Reporter: **Contig19865_RC** (SPOT ID)
Contig41805_RC	No annotation available; Reporter: **Contig41805_RC** (SPOT ID)
Contig26760	No annotation available; Reporter: **Contig26760** (SPOT ID)
Contig49045_RC	No annotation available; Reporter: **Contig49045_RC** (SPOT ID)
Contig34477_RC	No annotation available; Reporter: **Contig34477_RC** (SPOT ID)

Βιβλιογραφία

[1] Λ. Θ. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend, "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, pp. 530–536, 2002.

[2] Αλαχιώτης Σ. Ν., *Εισαγωγή στη σύγχρονη γενετική*. Φοιτητικά Συγγράμματα και Πανεπιστημιακές Σημειώσεις, 1989.

[3] Παναγιώτης Φ. Σαρρής, *Βιολογία και γενετική*. Αγρόπολις, 2004.

[4] Λουκάς Χ. Μαργαρίτης, *Κυτταρική Βιολογία*. Ιατρικές Εκδόσεις Λίτσας, 1996.

[5] ε. α. Clamp M, Fry B, "Distinguishing protein-coding and noncoding genes in the human genome." *Proc Natl Acad Sci USA*, vol. 104, pp. 19 428–33, 2007.

[6] B. B. e. a. Lander ES, Linton LM, "Initial sequencing and analysis of the human genome." *Nature*, vol. 409, pp. 860–921, 2001.

[7] Ron Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, pp. 273 – 324, 1997.

[8] S. Baker and B. Kramer, "Identifying genes that contribute most to good classification in microarrays," *BMC Bioinformatics*, vol. 7, p. 407, 2006.

[9] R. Doe. (2002) cancer incidence, mortality and prevalence worldwide. [Online]. Available: http://www-dep.iarc.fr

[10] J. Schneider. (1997) cross validation. [Online]. Available: http://www.cs.cmu.edu/ schneide/tut5/node42.html

[11] Κωνσταντίνος Διαμαντάρας, *Τεχνητά Νευρωνικά Δίκτυα*. Κλειδάριθμος, 2007.

[12] Ω. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Bulletin of Mathematical Biology*, vol. 52, pp. 99–115, 1990.

[13] C. K. Pal NR, "A connectionist system for feature selection." *Neural, Parallel and Scientific Computations*, vol. 5, pp. 359–382, 1997.

[14] Simon S. Haykin, *Neural networks and learning machines*. Prentice Hall, 2009.

[15] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, pp. 389–422, 2002.

[16] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Royal. Statist. Soc B.*, vol. 58, pp. 267–288, 1996.

[17] Richard J. Hathaway and J. C. Bezdek, "Nerf c-means: Non-euclidean relational fuzzy clustering," *Pattern Recognition*, vol. 27, pp. 429 – 437, 1994.

[18] M. T. J.-Y. J. V. A. F. T. F. D. Dennis P Wall, Rimma Pivovarov and P. J. Tonellato. (2010) genotator: A disease-agnostic tool for genetic annotation of disease. [Online]. Available: http://genotator.hms.harvard.edu/

[19] S.-J. K. Kwangmoo Koh and S. Boyd. (2008) simple matlab solver for l1-regularized least squares problems.