ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΜΕΤΑ–ΑΝΑΛΥΣΗ ΜΕ ΜΗΧΑΝΕΣ ΔΙΑΝΥΣΜΑΤΩΝ ΥΠΟΣΤΗΡΙΞΗΣ ΣΕ ΓΟΝΙΔΙΑΚΑ ΔΕΔΟΜΕΝΑ

ΜΙΧΑΛΟΠΟΥΛΟΥ ΕΛΕΥΘΕΡΙΑ

Επιβλέπων Καθηγητής:

κος Κουκουβίνος Χρήστος

Αθήνα, 2016

«Προς γαρ το τελευταίον εκβάν,

έκαστον των πριν υπαρξάντων κρίνεται»

Δημοσθένης ο Ολυνθιακός Α’

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΕΡΙΕΧΟΜΕΝΑ ΠΙΝΑΚΩΝ

ΠΕΡΙΕΧΟΜΕΝΑ ΣΧΗΜΑΤΩΝ

ΠΕΡΙΛΗΨΗ

Εξαιτίας της αυξανόμενης διαθεσιμότητας μελετών και δεδομένων, και ειδικά των συνόλων με microarray δεδομένα, των οποίων η ανάλυση έχει μετατραπεί τα τελευταία χρόνια σε μια περιοχή έντονης έρευνας (Vardhanabhuti et al., 2006), εμφανίζεται μεγάλη η ανάγκη για ολοκληρωμένες υπολογιστικές μεθόδους, οι οποίες θα αξιολογούν τα πολλαπλά και ανεξάρτητα αυτά σύνολα δεδομένων και θα διεξάγουν συμπεράσματα με βάση αυτά (Fishel et al., 2007). Η συμβολή της μετα–ανάλυσης και των μεθόδων μηχανικής μάθησης εμφανίζεται σε αυτό το ζήτημα αρκετά μεγάλη. Συγκεκριμένα, η τεχνική της μετα–ανάλυσης βοηθά στο να αντιμετωπιστούν θέματα, όπως το μικρό μέγεθος δείγματος και η ύπαρξη μεροληψίας, παράγοντας πιο έγκυρα και ενημερωτικά αποτελέσματα (Fishel et al., 2007). Επίσης, οι μέθοδοι μηχανικής μάθησης αποτελούν ισχυρό εργαλείο για την ανάλυση των προφίλ των γονιδιακών εκφράσεων, με σκοπό την πρόγνωση, τη διάγνωση και την αντιμετώπιση του καρκίνου (Fishel et al., 2007).

Έτσι, στην παρούσα διπλωματική εργασία, στο κεφάλαιο 1 αρχικά, κάνουμε μια εισαγωγή στην μετα–ανάλυση. Θα δούμε τι είναι η μετα–ανάλυση, ποιος την όρισε για πρώτη φορά και ποια είναι τα βασικά βήματα για την πραγματοποίηση της. Θα αναφερθούμε λεπτομερώς στα μοντέλα που χρησιμοποιούν κυρίως οι ερευνητές όταν διεξάγουν μια μετα–ανάλυση, καθώς και το πώς οι ίδιοι μετρούν και αντιμετωπίζουν την ετερογένεια (*heterogeneity*) μεταξύ των διαφόρων μελετών.

Στο κεφάλαιο 2, ορίζουμε το τι είναι ταξινόμηση και αναφέρουμε κάποιες βασικές μεθόδους της, όπως τα δέντρα απόφασης, τα τεχνητά νευρωνικά δίκτυα, την λογιστική παλινδρόμηση, τα Μπεϋζιανά μοντέλα δικτύου και τις μηχανές διανυσμάτων υποστήριξης. Θα δούμε συνοπτικά πως αυτές λειτουργούν και ποια είναι τα πλεονεκτήματα τους.

Το τρίτο κεφάλαιο είναι αφιερωμένο στις μηχανές διανυσμάτων υποστήριξης και στον αλγόριθμο SVM – RFE. Αρχικά, αναφερόμαστε αναλυτικά στο πως γίνεται η δυαδική ταξινόμηση και η παλινδρόμηση με χρήση των SVM. Στη συνέχεια, γίνεται μια εισαγωγή στην επιλογή χαρακτηριστικών (*feature selection*), καθώς ο αλγόριθμος SVM – RFE αποτελεί μέθοδο για feature selection. Έπειτα, περιγράφουμε την λειτουργία του αλγορίθμου και παρέχουμε μια εφαρμογή του σε δεδομένα που αποκομίσαμε από την UCI machine learning repository.

Στο κεφάλαιο 4 παραθέτουμε μεθόδους αξιολόγησης για την ταξινόμηση, αλλά και αναλύσεις, οι οποίες όταν εμπεριέχονται σε μια μετα–ανάλυση την

κάνουν πιο αξιόπιστη και βοηθούν τους ερευνητές να αξιολογήσουν το τελικό της συμπέρασμα.

Στο πέμπτο και τελευταίο κεφάλαιο εισάγεται μια νέα μέθοδος, η οποία προτάθηκε από τον Fishel et al. (2007) και αποτελεί μια predictor – based μετα–ανάλυση microarray δεδομένων, όπου ο predictor είναι ένας SVM ταξινομητής για τον διαχωρισμό των πνευμονικών ιστών σε καρκινικούς και φυσιολογικούς. Αρχικά, παραθέτονται κάποιες βασικές έννοιες σχετικά με το τι είναι τα DNA microarrays, οι γονιδιακές εκφράσεις κ.α., και στη συνέχεια παραθέτεται το πρόβλημα της χαμηλής overlap, το οποίο και στοχεύει να αντιμετωπίσει η μέθοδος που εισάγεται στη μελέτη του Fishel et al. (2007). Τέλος, περιγράφουμε αναλυτικά την μέθοδο και παρουσιάζουμε τα αποτελέσματα της.

ABSTRACT

Due to the increasing availability of studies and data sets, specific of microarray data sets whose analysis has become the last years an area of intense research (Vardhanabhuti et al., 2006), appears a growing need for integrative computational methods that evaluate those multiple and independent data sets and based on them conduct conclusions (Fishel et al., 2007). The contribution of meta–analysis and machine learning methods is large in this matter. Particularly the meta–analysis technique helps the investigators to deal with problems, such as the small sample size and the existence of biases, for the production of more reliable and informative results (Fishel et al., 2007). Also the machine learning methods have proven to be o powerful tool for the analysis of gene expression profiling in order to predict, detect and cure cancer.

So, in this thesis, we represent in chapter 1 an introduction to meta–analysis. We study what is a meta–analysis, who defined her and which are the basic steps for her construction. We refer in detail to the models, that the investigators mostly use when they conduct a meta–analysis and also how they compute and deal with heterogeneity between the studies.

In chapter 2 we give a definition of classification and we describe some classification methods, such as the Decision trees, the Artificial Neural Networks, the Logistic Regression, the Bayesian network models and the Support Vector Machines (SVM). We discuss in summary about their function and their advantages.

The third chapter is dedicated to Support Vector Machines (SVM) and SVM – RFE algorithm. At first, we refer analytically to how binary classification and regression are made using SVM. Then, we make an introduction to feature selection, due to the fact that the SVM – RFE is a method for feature selection. Afterwards, we describe the function of SVM – RFE and we present an application of this method using data from the UCI machine learning repository.

In chapter 4 we present evaluation methods for classification and specific analysis, those give reliability to meta–analysis and help the investigators evaluate her final conclusion.

In fifth and final chapter we introduce a new method suggested by Fishel et al. (2007), which consists a predictor – based meta–analysis of microarray data sets, where the predictor is a SVM classifier for the separation of tumor and normal lung tissues. At first we offer some basic definitions (what is a DNA microarray, a gene expression etc) and then we describe the problem of low overlap, that Fishel's et al. (2007) method is trying to solve. At the end we describe the new method and her results.

ΕΥΧΑΡΙΣΤΙΕΣ

Σε αυτό το σημείο θα ήθελα να ευχαριστήσω πρωτίστως τον καθηγητή μου, κύριο Χρήστο Κουκουβίνο, για την ευκαιρία που μου έδωσε να συνεργαστούμε και να μελετήσω ενδιαφέροντα αντικείμενα, μέχρι στιγμής άγνωστα για εμένα. Μέσα από την προσφορά και καθοδήγηση του έμαθα, ωρίμασα, αλλά και καλλιέργησα μέσα μου τι είναι τελικά αυτό που από εδώ και πέρα θέλω να κάνω, να ασχοληθώ με την Βιοστατιστική και την έρευνα. Τον ευχαριστώ πολύ λοιπόν για τον χρόνο που μου αφιέρωσε και για την προσοχή του καθ' όλη την διάρκεια εκπόνησης της διπλωματικής μου εργασίας.

Επίσης, ένα μεγάλο ευχαριστώ για την πολύτιμη βοήθεια της, οφείλω να πω στην υποψήφια διδάκτορα Κρυσταλλένια Δρόσου. Χωρίς τις συμβουλές της δεν θα τα είχα καταφέρει, γι' αυτό την ευχαριστώ από καρδιάς που ήταν δίπλα μου όλο αυτό το διάστημα και απαντούσε σε όποια απορία και αν είχα.

Φυσικά, δεν θα μπορούσα να μην ευχαριστήσω τους γονείς μου και την αδερφή μου, χωρίς την αγάπη και στήριξη των οποίων δεν θα είχα πραγματοποιήσει κανένα στόχο μου. Εξαιτίας τους κατάφερα να φέρω εις πέρας αυτή την εργασία, αλλά και να πάρω την απόφαση να συνεχίσω τις σπουδές μου στο αντικείμενο που με ενδιαφέρει. Όλα τα χρόνια των σπουδών μου η υπομονή τους και η εμπιστοσύνη που μου έδειξαν μου έδινε δύναμη να συνεχίσω και να τα καταφέρω, όχι μόνο για εμένα, αλλά και για να κάνω εκείνους χαρούμενους και περήφανους.

Τέλος, θα ήθελα να ευχαριστήσω και τους φίλους μου, που με το χιούμορ τους με κάνουν πάντα να χαμογελώ και να ξεχνάω ότι με προβληματίζει και με αγχώνει. Τους ευχαριστώ που είναι πάντα δίπλα μου στις καλές αλλά και στις κακές στιγμές.

Μιχαλοπούλου Ελευθερία

Αθήνα, 2016

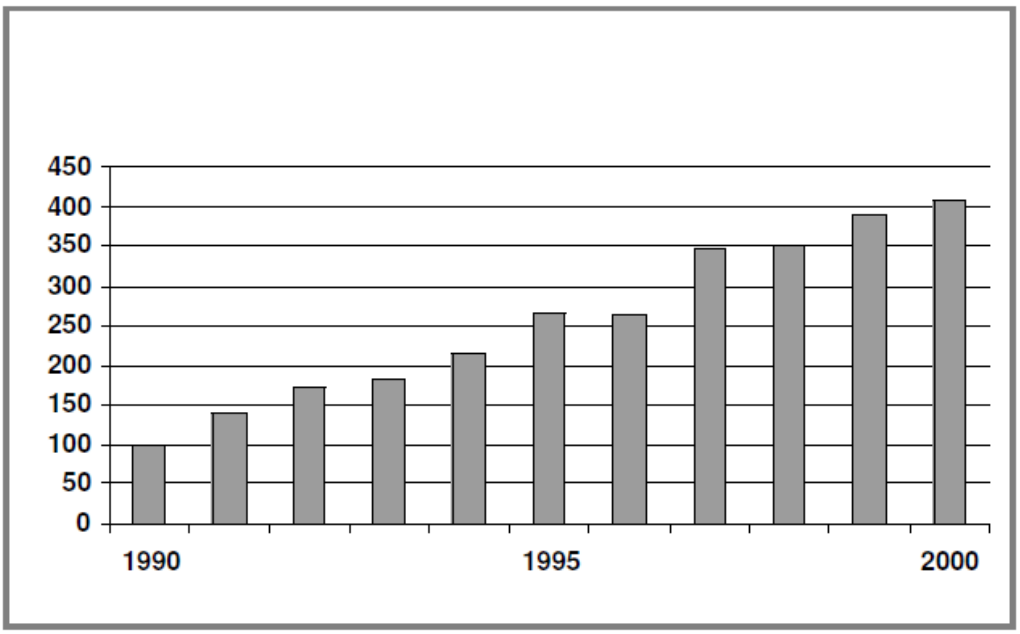
ΚΕΦΑΛΑΙΟ 1 : ΜΕΤΑ-ΑΝΑΛΥΣΗ

1.1 Εισαγωγή στη Μετα–Ανάλυση

Εξαιτίας της έκρηξης πληροφοριών σχετικών με διάφορες επιστημονικές υποθέσεις κατά τη διάρκεια των τελευταίων 70 ετών, η ανάγκη για έγκυρες και έγκαιρες αποφάσεις, τόσο σε θέματα δημόσιας υγείας όσο και στην καθημερινή κλινική πρακτική, καθιστά απαραίτητη την σύνθεση των αποτελεσμάτων μελετών που διεξάγονται καθημερινά ανά τον κόσμο (Γαλάνης, 2009). Υπάρχουν φορές όμως που το πλήθος των ερευνών που έχουν γίνει σε ένα πεδίο είναι αρκετά μεγάλο και συχνά τα αποτελέσματα που προκύπτουν είναι όχι μόνο διαφορετικά μεταξύ τους αλλά και αντικρουόμενα, δυσχεραίνοντας έτσι τη διατύπωση ενός τελικού συμπεράσματος. Επίσης, σύμφωνα με την Normand (1998), η κάθε έρευνα ξεχωριστά δεν έχει την δύναμη από μόνη της να ανιχνεύσει ένα αποτέλεσμα, ή ακόμη και αν το ανιχνεύσει, να είναι σίγουρη για αυτό και να μπορέσει να το εδραιώσει.

Ιστορική αναδρομή

Ο πρώτος που έθιξε την αναγκαιότητα συνδυασμού των πληροφοριών ήταν ο Άγγλος στατιστικολόγος και βιολόγος Ronald Fisher την δεκαετία του 1920 (Olkin, 1995). Όμως, όπως αναφέρεται και στο σύγγραμμα του Gioacchino (2005), πριν από αυτόν, συγκεκριμένα το 1904, ο Άγγλος μαθηματικός Karl Pearson πραγματοποίησε την πρώτη μετα–ανάλυση συνδυάζοντας τα δεδομένα πέντε διαφορετικών μελετών, με σκοπό να εξετάσει την συσχέτιση μεταξύ του εμβολιασμού για τον εντερικό πυρετό και της θνητότητας εξαιτίας αυτού του είδους πυρετού. Επίσημα όμως ο όρος «μετα–ανάλυση» επινοήθηκε το 1976 από τον Αμερικανό στατιστικολόγο Gene Glass, ο οποίος και την όρισε ως μια συγκεκριμένη τεχνική, διαχωρίζοντας την από την πρωτογενή ανάλυση των ερευνητικών δεδομένων.



Σχήμα 1: Ο αριθμός των παραγόμενων άρθρων ομαδοποιημένα ανά χρονολογία, από την στιγμή που ο όρος μετα–ανάλυση παρατάθηκε στην περίληψη (*abstract*) μιας δημοσίευσης, όπως παραθέτεται στο σύγγραμμα του Gioacchino (2005).

Ορισμός

Σύμφωνα με τον Gene Glass (1976):

"Η μετα-ανάλυση είναι η ανάλυση των αναλύσεων."

Γενικά, η μετα–ανάλυση αποτελεί ένα χρήσιμο εργαλείο αποσαφήνισης και σύνθεσης της υπάρχουσας γνώσης (Τσιάρα, 2011). Είναι μια αρχικά ερευνητική και στη συνέχεια μαθηματική διαδικασία, που συνδυάζει στατιστικά τα αποτελέσματα ανεξάρτητων και διαφορετικών μεταξύ τους μελετών, οι οποίες επιλέχθηκαν έπειτα από συστηματική ανασκόπηση και αφορούν ένα συγκεκριμένο επιστημονικό πεδίο (Γαλάνης, 2009). Ρόλος της μετα–ανάλυσης είναι η συγκέντρωση των αποτελεσμάτων αυτών σε ένα ενιαίο αποτέλεσμα, καθώς και η επαλήθευση και ερμηνεία τους.

Όπως περιγράφεται και στο άρθρο του Olkin (1995):

"Η μετα-ανάλυση δεν είναι μια αυτόματη διαδικασία. Αυτός που την εκτελεί χρειάζεται να καταναλώσει αρκετό χρόνο, αλλά και να κατέχει τις απαραίτητες γνώσεις για το αντικείμενο του θέματος και για μη συνήθεις

στατιστικές διαδικασίες. Είναι εύκολο να κάνεις μια μετα-ανάλυση. Είναι όμως δύσκολο να κάνεις μια καλή μετα-ανάλυση."

Πλεονεκτήματα

Σύμφωνα με τους Gioacchino (2005) και Olkin (1995), μέσω της σύνθεσης πληροφοριών:

- ✓ Επιτυγχάνεται η αύξηση της στατιστικής ισχύος μιας σύγκρισης.
- ✓ Συνδυάζονται μελέτες που είναι αντικρουόμενες.
- ✓ Εξετάζονται εκ νέου ερωτήματα, τα οποία δεν μπορούν να απαντηθούν μέσω μιας μοναδικής μελέτης, είτε γιατί είναι πολύ μικρή (δηλαδή έχει μικρό μέγεθος δείγματος), είτε γιατί είναι πολύ περιορισμένη (πχ. αναφέρεται μόνο σε άτομα κάτω από 18 ετών), ώστε να επιτρέψει την γενίκευση του αποτελέσματος της και σε άλλους πληθυσμούς (πχ. σε άτομα άνω των 18).
- ✓ Γίνεται κριτική αξιολόγηση των διαφόρων μελετών. Μέσω της μετα–ανάλυσης μπορούν να επιβεβαιωθούν ή να αναιρεθούν τα αποτελέσματα μιας μελέτης, της οποίας ο συγγραφέας μπορεί να μην είναι πρόθυμος να προσφέρει τα αυθεντικά δεδομένα του για επανα - αναλύσεις.
- ✓ Προσδιορίζονται περιοχές που χρήζουν επιπρόσθετης έρευνας δίνοντας νέες κατευθύνσεις.
- ✓ Εξετάζονται χαρακτηριστικά των μελετών, εξαιτίας των οποίων προκύπτουν υποσύνολα του πληθυσμού, όπου μπορεί μια θεραπεία να είναι πιο αποτελεσματική απ' ότι σε ολόκληρο τον πληθυσμό.
- ✓ Αναλύεται αν τελικά και πως οι προϋπάρχουσες μελέτες τροποποίησαν τις μέχρι τώρα γνώσεις μας πάνω σε ένα θέμα.
- ✓ Παρέχονται πιο αυθεντικές αποδείξεις. Τα αποτελέσματα που προκύπτουν είναι λιγότερο μεροληπτικά, σε σχέση με αυτά άλλων ανασκοπήσεων, καθώς η μετα-ανάλυση δεν προδικάζει και δεν αποκλείει από πριν μελέτες λόγω του ιδιαίτερου σχεδιασμού τους.

1.2 Βήματα για την πραγματοποίηση μιας μετα–ανάλυσης

Βασιζόμενοι στις διατυπώσεις των Gioacchino (2005) και Τσιάρα (2011) ορίζουμε τα παρακάτω βήματα:

- **1^{ο} Βήμα**: Προτού ξεκινήσει την οποιαδήποτε ερευνητική ή μαθηματική διαδικασία ο μετα–αναλυτής πρέπει να προσδιορίσει με σαφήνεια το αντικείμενο της έρευνας - τους στόχους της μετα–ανάλυσης. Θεμελιώνει αρχικές υποθέσεις

που βασίζονται σε ευρήματα προηγούμενων μελετών και στην ουσία ξεκινά με μια κριτική θεώρηση όλων των ερευνών που έχει στη διάθεση του. Ο μετα–αναλυτής υποδεικνύει έτσι τα θέματα που προέκυψαν και χρήζουν κατά τη γνώμη του περεταίρω διευκρίνισης και αποφασίζει με τι από αυτά θα ασχοληθεί.

- **2^{ο} Βήμα**: Αφού ο μετα–αναλυτής έχει αποφασίσει πλέον ποιο θέμα θα ερευνήσει, σειρά έχει η σύνταξη του ερευνητικού πρωτοκόλλου, δηλαδή:

i. Ο προσδιορισμός των πηγών πληροφορίας, στις οποίες θα διεξαχθεί η βιβλιογραφική έρευνα. Για παράδειγμα: μηχανές αναζήτησης στο διαδίκτυο (πχ. Medline, Embase, Εθνικό κέντρο τεκμηρίωσης, Index Medicus), γραπτές εργασίες, παραπομπές μελετών που έχουν δημοσιευτεί, καθώς και μελέτες ακαδημαϊκών, κρατικών ή ιδιωτικών οργανισμών που για διάφορους λόγους δεν έχουν δημοσιευτεί. Ο ερευνητής οφείλει να στηριχτεί σε όλες τις πηγές που προαναφέραμε, προκειμένου να αποφύγει την ύπαρξη σφάλματος δημοσίευσης (*publication bias*) στη μελέτη του (βλέπε 6^{ο} βήμα → ii) (Γαλάνης, 2009).

ii. Η εύρεση των κατάλληλων λέξεων κλειδιών, οι οποίες θα συγκεκριμενοποιήσουν περισσότερο το θέμα της έρευνας και θα την κάνουν πιο αποτελεσματική.

iii. Η επιλογή των κριτηρίων ένταξης ή αποκλεισμού μιας μελέτης στην μετα–ανάλυση (Normand, 1998). Η διαδικασία αυτή αποτελεί ένα από τα δυσκολότερα κομμάτια μιας μετα–ανάλυσης, καθώς οι κανόνες που χρησιμοποιούνται για την απόρριψη ή την αποδοχή μιας μελέτης δεν είναι ξεκάθαροι, αλλά αποτελούν υποκειμενική απόφαση των ερευνητών, με την προϋπόθεση βέβαια ότι η απόφαση αυτή είναι ορθολογική και εξυπηρετεί τις επιδιώξεις της επιστήμης και όχι προσωπικά ή οικονομικά συμφέροντα (Γαλάνης, 2009).

iv. Ο προσχεδιασμός των μεθόδων και των διαδικασιών της μετα–ανάλυσης. Ουσιαστικά ο μετα–αναλυτής κάνει μια πρώτη σκέψη για το πώς περίπου θα διαχειριστεί τα δεδομένα του και ποιες αναλύσεις θα πραγματοποιήσει. Αναλυτικά σε αυτές θα αναφερθούμε στο 5^{ο} και στο 6^{ο} βήμα.

- **3^{ο} Βήμα**: Έχοντας ως βάση το ερευνητικό πρωτόκολλο που σχεδίασε στο προηγούμενο βήμα, ο μετα–αναλυτής σε αυτό το βήμα διεξάγει την βιβλιογραφική έρευνα, λαμβάνοντας υπ' όψιν όλα τα δεδομένα και όχι μόνο αυτά στα οποία έχει εύκολη πρόσβαση ή τα θεωρεί ενδιαφέροντα. Η βιβλιογραφική αναζήτηση απαιτείται να είναι συστηματική και αναλυτική. Στόχος του μετα–αναλυτή είναι να αποφύγει τις μεροληψίες και τα σφάλματα, τα οποία μπορεί να επηρεάσουν το τελικό αποτέλεσμα, αλλά και να πείσει τον

αναγνώστη ότι με την βιβλιογραφική έρευνα που διεξήγαγε κατάφερε να συγκεντρώσει ένα τέτοιο αμερόληπτο δείγμα.

Στη συνέχεια ο μετα–αναλυτής επιλέγει ποιές μελέτες θα συμπεριλάβει στην ανάλυσή του. Στο σημείο αυτό ορισμένοι προτείνουν να εκτιμάται η ποιότητα των μελετών που επιλέχθηκαν, με σκοπό να προσδίδεται μεγαλύτερη βαρύτητα κατά τον υπολογισμό του συγκεντρωτικού αποτελέσματος στις μελέτες που θεωρούνται ποιοτικότερες (Γαλάνης, 2009). Πολλοί ερευνητές, όπως οι Cook και Cambel σύμφωνα με την Normand (1998), διατύπωσαν τα δικά τους πλαίσια για την απονομή του κατάλληλου επονομαζόμενου "*quality score*" σε μια μελέτη, αλλά γενικότερα δεν υπάρχει μέχρι στιγμής κάποιο από κοινού αποδεκτό κριτήριο για την απονομή αυτή. Για το λόγο αυτό είναι προτιμότερο να αποφεύγεται η εκτίμηση της ποιότητας των μελετών μέσω των quality scores, διότι μπορεί να οδηγήσει σε μεροληπτικά αποτελέσματα και μειωμένη εγκυρότητα.

Όλες οι παραπάνω διαδικασίες (βιβλιογραφική έρευνα, επιλογή και εκτίμηση της ποιότητας των μελετών) πρέπει να παρουσιάζονται αναλυτικά από τους συγγραφείς μιας μετα–ανάλυσης και να παρέχεται επαρκής αιτιολόγηση για το γιατί και με ποιόν τρόπο πραγματοποιήθηκε κάθε στάδιο. Οι πληροφορίες αυτές είναι πολύ βασικές, γι' αυτό και η καταγραφή τους προσδίδει στην μετα–ανάλυση μεγαλύτερη αξιοπιστία. Επιπλέον, σύμφωνα με την Normand (1998), είναι πολύ βασικό η βιβλιογραφική έρευνα και η επιλογή των μελετών να πραγματοποιούνται τουλάχιστον από δυο διαφορετικούς ερευνητές, των οποίων στη συνέχεια τα ευρήματα θα συνδυάζονται, προκειμένου να μειωθεί η πιθανότητα ύπαρξης μεροληψίας στην ανάλυση.

Μέθοδος "capture – recapture"

Πριν κλείσουμε με αυτό το βήμα είναι σημαντικό να αναφερθούμε στην μέθοδο "capture – recapture". Η "capture – recapture" είναι μια μέθοδος δειγματοληψίας που χρησιμοποιείται κυρίως από τους βιολόγους για την μέτρηση του αριθμού των μελών ενός ζωολογικού είδους σε μια γεωγραφική περιοχή. Για παράδειγμα, έστω ότι επιθυμούμε να εκτιμήσουμε τον αριθμό των καρχαριών στη Μεσόγειο θάλασσα. Για τον σκοπό αυτό πιάνουμε όσους καρχαρίες μπορούμε και τους «μαρκάρουμε» (στάδιο capture). Αυτοί αποτελούν το αρχικό δείγμα μας. Στη συνέχεια, απελευθερώνουμε τους «μαρκαρισμένους» καρχαρίες και μετά από το πέρας ενός χρονικού περιθωρίου, που είναι απαραίτητο για τον διασκορπισμό τους στη θάλασσα, συλλέγουμε ένα δεύτερο δείγμα καρχαριών (στάδιο recapture). Τέλος, μετράμε πόσοι μαρκαρισμένοι καρχαρίες βρέθηκαν ανάμεσα στους καρχαρίες του δεύτερου δείγματος.

Όσον αφορά τώρα την χρήση της “capture – recapture” στη μετα–ανάλυση, σύμφωνα με τον Gioacchino (2005), η “capture – recapture” είναι μια τεχνική για την αξιολόγηση της πληρότητας της έρευνας. Συγκεκριμένα, μέσω του αποτελέσματος αυτής μπορεί να εκτιμηθεί ο αριθμός των μελετών που δεν βρέθηκαν, νούμερο το οποίο αποτελεί πολύ σημαντική πληροφορία για την ποιοτική αξιολόγηση της μετα–ανάλυσης. Ας υποθέσουμε λοιπόν ότι η έρευνα που διεξάχθηκε από τους μετα–αναλυτές στο Medline μια συγκεκριμένη χρονική περίοδο απέφερε M μελέτες σχετικές με το αντικείμενο ενδιαφέροντος (στάδιο capture), ενώ η έρευνα που έγινε στο Index Medicus και στις αναφορές των συγκεντρωθέντων μελετών απέφερε n papers (στάδιο recapture). Από τις συνολικά $M + n$ μελέτες θα συμπεριληφθούν εν τέλει στην μετα–ανάλυση, έστω, m μελέτες. Λαμβάνοντας υπ’ όψιν την ανεξαρτησία μεταξύ των ερευνητικών πηγών, μπορούμε τώρα να εκτιμήσουμε την αναμενόμενη τιμή του συνολικού αριθμού των μελετών (αυτές που βρέθηκαν + αυτές που δεν βρέθηκαν), N, από την σχέση που ακολουθεί.

$$N = M\,\frac{n}{m}$$

με διακύμανση

$$\mathrm{Var}(N) = \frac{M \cdot n(M-m)(n-m)}{m^3}$$

Η παραπάνω εκτιμήτρια αποτελεί αποτέλεσμα της μεθόδου της μέγιστης πιθανοφάνειας (*maximum likelihood method*). Καθώς όμως αυτή παραμορφώνεται στην περίπτωση που τα M, n, m είναι μικρά, συνίσταται η χρήση της ακόλουθης εκτιμήτριας, που προκύπτει από την μέθοδο του Chapman:

$$N = \frac{(M+1)(n+1)}{m+1} - 1$$

με διακύμανση

$$\mathrm{Var}(N) = \frac{(M+1)(n+1)(M-m)(n-m)}{2(m+1)(m+2)}$$

Υποχρεωτική απαίτηση για την χρήση της μεθόδου “capture – recapture” είναι η ανεξαρτησία μεταξύ των ερευνητικών πηγών (Gioacchino, 2005).

- **4^{ο} Βήμα**: Σε αυτό το βήμα ο μετα–αναλυτής, έχοντας επιλέξει τις μελέτες που θα συμπεριλάβει στην ανάλυση του, προχωρά στην εξαγωγή των δεδομένων από αυτές και στην καταγραφή τους. Συγκεκριμένα:

i. Καταγράφει στοιχεία που αφορούν τον τύπο της μελέτης (*study design*), για παράδειγμα τυχαιοποιημένες δοκιμές (*randomized trials*), μη πειραματικές μελέτες (*non experimental studies*), επισκόπηση (*survey*) κ.ο.κ., το έτος δημοσίευσης και τους συγγραφείς της, καθώς και τον αριθμό συμμετεχόντων σε αυτή.

ii. Προσδιορίζει το πρωτεύον αποτέλεσμα, το οποίο υπάρχει σε όλες τι μελέτες και αφορά πχ. την θνητότητα εξαιτίας μιας ασθένειας ή την αποτελεσματικότητα ενός φαρμάκου.

iii. Αναζητά δευτερεύοντα αποτελέσματα, που αφορούν υποσύνολα του πληθυσμού κάθε μελέτης, τα οποία όμως δεν είναι απαραίτητο να περιέχονται σε όλες τις μελέτες. Αυτού του είδους τα δεδομένα θα βοηθήσουν στη συνέχεια της ανάλυσης να απαντηθούν ερωτήματα σχετικά με τις πιθανές αιτίες ύπαρξης ετερογένειας μεταξύ των μελετών .

iv. Δίνει "λειτουργικούς" ορισμούς στα αποτελέσματα των μελετών**.** Ορίζει, δηλαδή, ποιος θα είναι ο πληθυσμός και ποιες οι μεταβλητές της μετα–ανάλυσης (Normand, 1998).

- **5^{ο} Βήμα**: Σειρά για τον μετα–αναλυτή έχει η στατιστική ανάλυση των δεδομένων που συνέλεξε και ο υπολογισμός του ενιαίου - συνολικού αποτελέσματος. Για να γίνει αυτό, ο μετα–αναλυτής:

i. Επιλέγει ποιό μέτρο αποτελέσματος (*effect size*) θα χρησιμοποιήσει ανάλογα με το αν τα δεδομένα που θα επεξεργαστεί είναι συνεχή (*continuous*), δίτιμα (*dichotomous*) ή άλλου τύπου. Το effect size είναι ένα μέτρο έντασης της σχέσης μεταξύ δύο μεταβλητών. Πιθανά μέτρα αποτελέσματος σύμφωνα με τον Kim (2011) είναι:

- Για συνεχή δεδομένα, όταν δηλαδή οι μελέτες υπολογίζουν μέσες τιμές και αποκλίσεις:

1. **Μέση διαφορά** (*mean difference*)
2. **Τυποποιημένη μέση διαφορά** (*standardized mean difference*): υπολογίζεται από την σχέση:

$$\delta = \frac{\mu_e - \mu_c}{\sigma}$$

όπου $\boldsymbol{\mu_e}$ η μέση τιμή της πειραματικής ομάδας (ασθενείς που πάσχουν από μια συγκεκριμένη νόσο), $\boldsymbol{\mu_c}$ η μέση τιμή της ομάδας αναφοράς – ελέγχου (υγιείς ασθενείς) και **σ** η διασπορά.

Η τυποποιημένη μέση τιμή δ εκτιμάται είτε μέσω του στατιστικού g του Cohen, είτε μέσω του στατιστικού d του Hedges (Hedges & Olkin, 1985). Συγκεκριμένα:

$$g = \frac{\overline{Y}_e - \overline{Y}_c}{\sqrt{(s_e{}^2 + s_c{}^2)/2}}$$

όπου $\boldsymbol{\overline{Y}_e}$ και $\boldsymbol{s_e{}^2}$ η μέση τιμή και η διακύμανση αντίστοιχα της πειραματικής ομάδας και $\boldsymbol{\overline{Y}_c}$ και $\boldsymbol{s_c{}^2}$ η μέση τιμή και η διακύμανση αντίστοιχα της ομάδας αναφοράς.

$$d = \frac{\overline{Y}_e - \overline{Y}_c}{s} \times J(N-2)$$

με

$$s = \sqrt{\frac{(n_e-1)\, s_e{}^2 + (n_c-1)\, s_c{}^2}{n_e+n_c-2}}$$

όπου $\boldsymbol{\overline{Y}_e}$ και $\boldsymbol{\overline{Y}_c}$ η μέση τιμή της πειραματικής ομάδας και της ομάδας ελέγχου αντίστοιχα, **s** η συνοπτική τυπική απόκλιση και $\boldsymbol{J(N-2)}$ σταθερά που δίνεται για $m \geq 2$ από τον τύπο:

$$J(m) = 1 - \frac{3}{4m-1}$$

Μια θετική τιμή του εκτιμητή g ή d δηλώνει ότι η πειραματική ομάδα υπερτερεί έναντι της ομάδας ελέγχου, ενώ μια αρνητική τιμή δείχνει το αντίστροφο.

- Για δίτιμα δεδομένα, όταν δηλαδή στις μελέτες καταγράφεται η πιθανότητα εμφάνισης κάποιου γεγονότος σε μια ομάδα:

1. **Περιττοί Λόγοι ή Λόγος Συμπληρωματικών Πιθανοτήτων** (*odds ratio* → OR): εκφράζει κατά πόσο μια διαδικασία (πχ. η εφαρμογή μιας θεραπείας) αυξάνει την πιθανότητα εμφάνισης ενός γεγονότος, μιας έκβασης (πχ. του θανάτου) (Gioacchino, 2005). Ως odd ορίζεται ο λόγος της πιθανότητας να συμβεί ένα γεγονός ως προς την πιθανότητα να μην συμβεί, δηλαδή: $\mathrm{p}/(1-\mathrm{p})$, όπου p η πιθανότητα επιτυχίας. Κατ' επέκταση ως odds ratio (OR) ορίζεται ο λόγος των odds για ένα γεγονός που συμβαίνει σε μια ομάδα προς τα odds αυτού του γεγονότος σε μια άλλη ομάδα, δηλαδή:

$$\mathrm{OR} = \frac{\mathrm{p}/(1-\mathrm{p})}{\mathrm{q}/(1-\mathrm{q})}$$

όπου **p** η πιθανότητα επιτυχίας στην πρώτη ομάδα και **q** η πιθανότητα επιτυχίας στην δεύτερη. Οι ομάδες αυτές μπορεί να είναι άνδρες και γυναίκες, μια πειραματική και μια ομάδα αναφοράς ή οποιαδήποτε άλλη διχοτομική ταξινόμηση. Στην περίπτωση που τα δεδομένα είναι της μορφής που φαίνεται στον πίνακα 1, τότε το odds ratio προκύπτει, όμοια με την παραπάνω σχέση, από τον τύπο:

$$\mathrm{OR} = \frac{\mathrm{a}/\mathrm{b}}{\mathrm{c}/\mathrm{d}}$$

Η διακύμανση της αντίστοιχης μελέτης τότε θα είναι:

$$\mathrm{s}^2 = \frac{\mathrm{a}+\mathrm{b}+\mathrm{c}+\mathrm{d}}{\mathrm{b}\cdot\mathrm{c}}$$

Ένα OR ίσο με τη μονάδα δηλώνει ότι μεταξύ των δύο ομάδων δεν υπάρχει στατιστικά σημαντική διαφορά, καθώς η εμφάνιση του γεγονότος είναι εξίσου πιθανή και στις δύο ομάδες. Όταν προκύψει $\mathrm{OR} > 1$ συμπεραίνεται ότι στην πρώτη ομάδα υπάρχει μεγαλύτερη πιθανότητα να εμφανιστεί το γεγονός απ' ότι στη δεύτερη. Αντίθετα, όταν $\mathrm{OR} < 1$ η δεύτερη ομάδα είναι πιο πιθανό να εμφανίσει το γεγονός από ότι η πρώτη. Προφανώς, οι τιμές που παίρνει το odds ratio είναι μη αρνητικές.

Πίνακας 1: Γενική μορφή πίνακα δεδομένων για τις μελέτες μιας μετα-ανάλυσης.

	Εμφάνιση γεγονότος	Απουσία γεγονότος	Σύνολο
Πειραματική ομάδα (experimental group)	a	b	a + b
Ομάδα αναφοράς – ελέγχου (control group)	c	d	c + d
Σύνολο	a + c	b + d	

2. **Σχετικός κίνδυνος** (*risk ratio* ή *relative risk* → RR): χρησιμοποιείται όπως και το OR για τη σύγκριση των πιθανοτήτων εμφάνισης ενός γεγονότος μεταξύ δύο ομάδων. Βασιζόμενοι στο ότι τα δεδομένα της μελέτης είναι της μορφής του Πίνακα 1 το RR υπολογίζεται από τη σχέση:

$$\mathrm{RR} = \frac{\mathrm{a}/(\mathrm{a}+\mathrm{c})}{\mathrm{b}/(\mathrm{b}+\mathrm{d})}$$

Ένα RR ίσο με τη μονάδα δηλώνει ότι μεταξύ των δύο ομάδων δεν υπάρχει στατιστικά σημαντική διαφορά, καθώς η εμφάνιση του γεγονότος είναι εξίσου πιθανή και στις δύο ομάδες. Όταν προκύπτει $\mathrm{RR} > 1$ συμπεραίνεται ότι στην πειραματική ομάδα υπάρχει μεγαλύτερη πιθανότητα να εμφανιστεί το γεγονός απ' ότι στην ομάδα αναφοράς. Αντίθετα, όταν $\mathrm{RR} < 1$, η ομάδα αναφοράς είναι πιο πιθανό να εμφανίσει το γεγονός από ότι η πειραματική ομάδα. Προφανώς, οι τιμές που παίρνει το risk ratio είναι μη αρνητικές. Στην περίπτωση που οι εκβάσεις που μελετάμε είναι αρκετά σπάνιες τείνει να ισχύει: $\mathrm{RR} \approx \mathrm{OR}$.

3. **Διαφορά κινδύνου** (*risk difference* → RD): και αυτό το effect size υπολογίζεται σε δεδομένα που εμφανίζονται σε 2 x 2 πίνακες, όπως ο Πίνακας 1. Δίνεται από τη σχέση

$$\mathrm{RD} = \frac{\mathrm{a}}{\mathrm{a}+\mathrm{b}} - \frac{\mathrm{c}}{\mathrm{c}+\mathrm{d}}$$

Μια αρνητική τιμή του RD υποδηλώνει ότι η πιθανότητα εμφάνισης του γεγονότος είναι μεγαλύτερη στην ομάδα αναφοράς παρά στην ομάδα ελέγχου. Μια θετική τιμή αντίθετα υποδηλώνει το αντίστροφο. Αν το risk

difference πάρει την τιμή 0 συμπεραίνουμε ότι δεν υπάρχει στατιστικά σημαντική διαφορά μεταξύ των δύο ομάδων και η εμφάνιση του γεγονότος είναι ισοπίθανη σε κάθε μία από αυτές.

- Για άλλου είδους δεδομένα:

1. **Λόγος επικινδυνότητας** (*hazard ratio*): χρησιμοποιείται όταν η μελέτη αφορά δεδομένα επιβίωσης (*survival data*).

2. **Συντελεστής συσχέτισης r του Pearson** (*Pearson's r Correlation coefficient*): χρησιμοποιείται όταν οι μελέτες μετρούν συσχετίσεις μεταξύ ποσοτικών μεταβλητών. Ισχύει ότι $-1 \leq r \leq 1$, όπου το -1 δείχνει αρνητική γραμμική σχέση, το 1 τέλεια θετική γραμμική και το 0 δείχνει τη μη ύπαρξη γραμμικής σχέσης μεταξύ των δυο μεταβλητών. Ο Cohen (1988) υποστήριξε ότι συσχετίσεις της τάξης του 0,1 είναι μικρές, της τάξης του 0,3 μεσαίες, ενώ συσχετίσεις της τάξης του 0,5 και πάνω θεωρούνται μεγάλες. Ο συντελεστής συσχέτισης δύο μεταβλητών x και y υπολογίζεται από τη σχέση:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

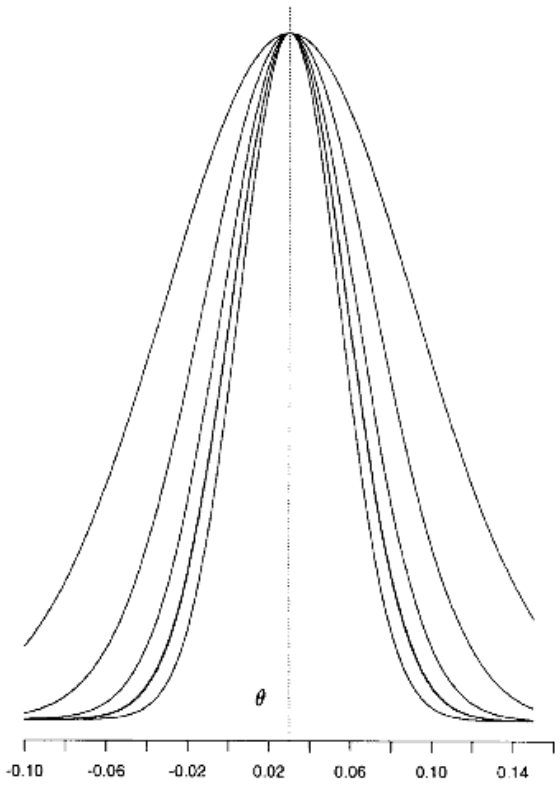
ii. Προσαρμόζει ένα μοντέλο στα μέτρα αποτελέσματος των μελετών που υπολόγισε. Σε μια μετα-ανάλυση κατά κύριο λόγο χρησιμοποιούνται τα μοντέλα των σταθερών και των τυχαίων επιδράσεων (*fixed effects model* και *random effects model* αντίστοιχα), ανάλογα με την ύπαρξη ή μη στατιστικής ετερογένειας μεταξύ των μελετών. Σε αυτή την επιλογή θα αναφερθούμε αναλυτικά στο 6^{ο} βήμα (6i: [ΕΠΙΛΟΓΗ ΜΟΝΤΕΛΟΥ]). Και στα δύο αυτά μοντέλα, εκτός από τα εκτιμώμενα μέτρα αποτελέσματος, υπολογίζεται για την κάθε μελέτη ξεχωριστά μια τιμή w_i, όπου $i = 1, 2, \ldots, n$ η μελέτη, η οποία καλείται βάρος και εκφράζει την βαρύτητα της i - οστής μελέτης στη μετα - ανάλυση, δηλαδή πόσο αυτή συμβάλει στον υπολογισμό του εκτιμώμενου συνολικού αποτελέσματος (*estimated total effect*). Συνήθως οι μελέτες με το μεγαλύτερο μέγεθος δείγματος έχουν και την μεγαλύτερη βαρύτητα (Kim, 2011). Πιο αναλυτικά:

- **Μοντέλο σταθερών επιδράσεων (fixed effects model):** θεωρεί ότι υπάρχει ένα μοναδικό πραγματικό αποτέλεσμα για όλες τις μελέτες και είναι σταθερό

(*fixed*). Δηλαδή ότι τα εκτιμώμενα μέτρα αποτελέσματος (*estimated effect sizes*) που προκύπτουν από την κάθε μελέτη προέρχονται από την κανονική κατανομή και έχουν την ίδια μέση τιμή, όπως φαίνεται και στο Σχήμα 2 (Normand,1998). Συγκεκριμένα ισχύει:

$$\widehat{y_i} \sim N(\Theta, s_i^2)$$

όπου $\widehat{\boldsymbol{y_i}}$ το εκτιμώμενο μέτρο αποτελέσματος της i μελέτης, **Θ** η μέση τιμή για οποιοδήποτε $\widehat{y_i}$ και $\mathbf{s_i^2}$ η διασπορά του $\widehat{y_i}$, με $i = 1, 2, 3, \ldots, n$, όπου **n** ο συνολικός αριθμός των μελετών που περιέχονται στη μετα-ανάλυση.

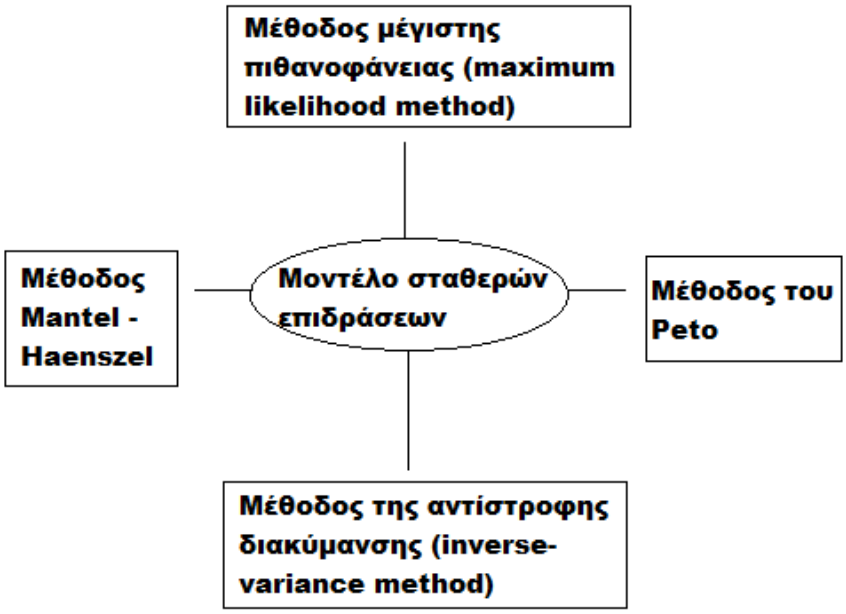


Σχήμα 2: Η κατανομή 5 υποθετικών στατιστικών μελετών υπό τις παραδοχές του fixed effects model, όπως απεικονίζεται στο άρθρο της Normand (1998). Η κάθετη διακεκομμένη γραμμή αναπαριστά την μέση τιμή Θ για κάθε μελέτη.

Η υπόθεση αυτή, σύμφωνα με τον Gioacchino (2005), ισοδυναμεί με την υπόθεση ύπαρξης στατιστικής ομοιογένειας μεταξύ των μελετών, την οποία θα περιγράψουμε αναλυτικά στο 6^{ο} βήμα (6i). Επομένως, όταν ο μετα-αναλυτής χρησιμοποιεί το μοντέλο των σταθερών επιδράσεων ενδιαφέρεται μόνο για την

“ενδο-ετερογένεια” (*within study variation*), από την οποία χαρακτηρίζεται η κάθε μελέτη ξεχωριστά.

Το fixed effects model διαθέτει τέσσερις μεθόδους για την εκτίμηση των μέτρων αποτελέσματος. Αυτές είναι:



Σχήμα 3: Μέθοδοι εκτίμησης των μέτρων αποτελέσματος όταν χρησιμοποιείται το fixed effects model.

Η μέθοδος που χρησιμοποιείται συχνότερα από τους αναλυτές είναι η Mantel – Haenszel, διότι θεωρείται πολύ ακριβής. Όσον αφορά τη μέθοδο του Peto, με βάση την μελέτη του Γαλάνη (2009), χρησιμοποιείται κυρίως στις περιπτώσεις όπου:

- ✓ ως effect size έχει επιλεχθεί το odds ratio.
- ✓ η μετα-ανάλυση αφορά κλινικές δοκιμές.
- ✓ ο αριθμός των περιπτώσεων έκβασης είναι σχετικά μεγάλος.
- ✓ ο αριθμός των συμμετεχόντων στην ομάδα αναφοράς και στην πειραματική ομάδα είναι ίσος.

Για το μοντέλο των σταθερών επιδράσεων χρησιμοποιώντας ενδεικτικά το odds ratio (OR) ως μέτρο αποτελέσματος και την μέθοδο Mantel – Haenszel προκύπτουν οι εξής εκτιμήτριες, όπως καταγράφονται στο σύγγραμμα του Gioacchino (2005):

Το **εκτιμώμενο μέτρο αποτελέσματος** (εδώ το OR) για την κάθε i μελέτη από τις n υπολογίζεται μέσω της σχέσης:

$$\hat{y}_i = \frac{a_i / b_i}{c_i / d_i}$$

με $i = 1, 2, \dots, n$, όπου τα $\mathbf{a_i}$, $\mathbf{b_i}$, $\mathbf{c_i}$ και $\mathbf{d_i}$ προκύπτουν από τον Πίνακα 1 αντίστοιχα για την i μελέτη.

Η **διασπορά του εκτιμώμενου μέτρου αποτελέσματος** (εδώ του OR) για κάθε i μελέτη από τις n δίνεται από τη σχέση:

$$s_i^2 = \frac{a_i + b_i + c_i + d_i}{b_i \cdot c_i}$$

με $i = 1, 2, \dots, n$, όπου τα $\mathbf{a_i}$, $\mathbf{b_i}$, $\mathbf{c_i}$ και $\mathbf{d_i}$ προκύπτουν από τον Πίνακα 1 αντίστοιχα για την i μελέτη.

Το **βάρος** υπολογίζεται από τον γενικό τύπο:

$$w_i = \frac{1}{s_i^2}$$

Ο **εκτιμητής του συνολικού αποτελέσματος** υπολογίζεται από τη σχέση:

$$\hat{Y} = \frac{\sum_{i=1}^{n} w_i \hat{y}_i}{\sum_{i=1}^{n} w_i}$$

Η **διασπορά του εκτιμητή του συνολικού αποτελέσματος** προκύπτει από τον τύπο:

$$S^2 = \frac{1}{\sum_{i=1}^{n} w_i}$$

Το **διάστημα εμπιστοσύνης του εκτιμώμενου συνολικού αποτελέσματος** ορίζεται ως:

$$\hat{\Upsilon} \pm 1{,}96 \frac{1}{\sqrt{\sum_{i=1}^{n} w_i}}$$

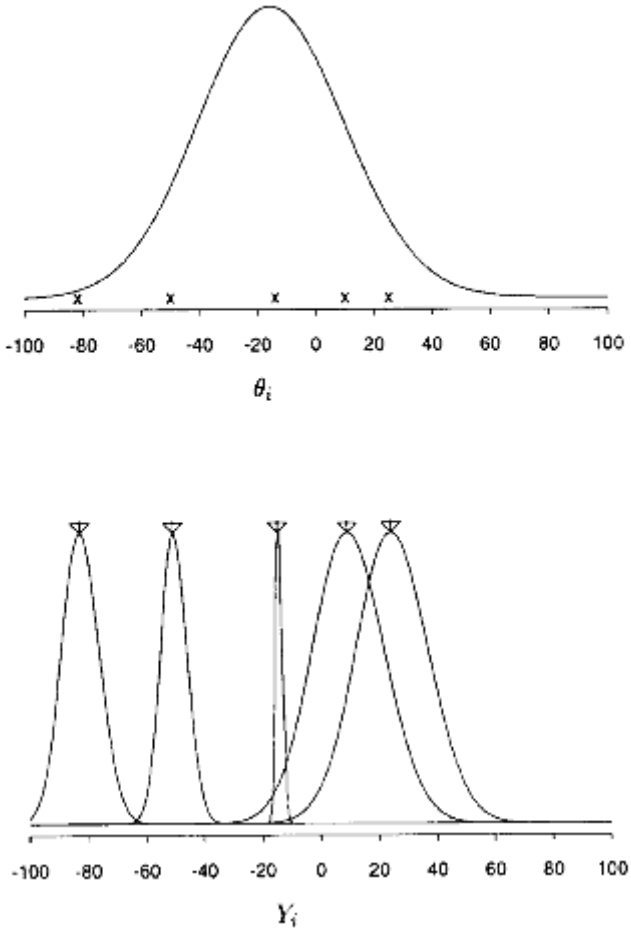
- **Μοντέλο τυχαίων επιδράσεων (random effects model):** θεωρεί ότι τα εκτιμώμενα μέτρα αποτελέσματος της κάθε μελέτης είναι τυχαίες μεταβλητές με τη δική τους μέση τιμή η κάθε μια. Η διακύμανση αυτού του μοντέλου είναι μεγαλύτερη και τα διαστήματα εμπιστοσύνης πιο ευρέα (Gioacchino, 2005). Συγκεκριμένα, σύμφωνα με την Normand (1998), ισχύει:

$$\hat{y_i} \sim N(\hat{\theta_i}, s_i^2)$$

με

$$\hat{\theta_i} \sim N(\Theta, \tau^2)$$

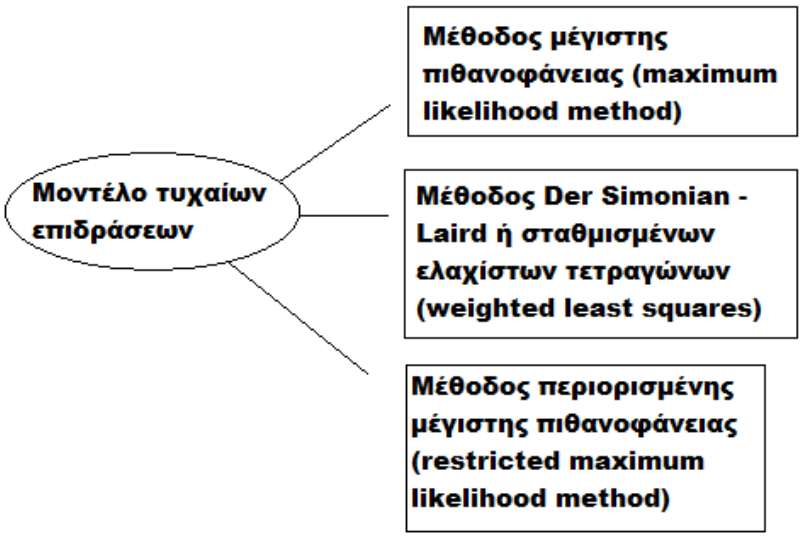
όπου $\hat{\boldsymbol{y_i}}$ το εκτιμώμενο μέτρο αποτελέσματος της i – μελέτης, $\hat{\boldsymbol{\theta_i}}$ η μέση τιμή του $\hat{y_i}$, $\mathbf{s_i^2}$ η διασπορά του $\hat{y_i}$, **Θ** η μέση τιμή των $\hat{\theta_i}$ και $\boldsymbol{\tau^2}$ η διασπορά του $\hat{\theta_i}$ (*between study variation*) με $i = 1, 2, 3, \dots, n$, όπου **n** ο συνολικός αριθμός των μελετών που περιέχονται στη μετα–ανάλυση.



Σχήμα 4: Η κατανομή 5 υποθετικών στατιστικών μελετών υπό τις παραδοχές του random effects model, όπως απεικονίζεται στο άρθρο της Normand (1998). Κάθε $\widehat{\boldsymbol{\theta}_i}$, στο σχήμα $\boldsymbol{\theta_i}$, αντλείται από ένα υπερπληθυσμό με μέση τιμή Θ και διακύμανση $\boldsymbol{\tau^2}$ (πάνω γράφημα). Τα $\widehat{\boldsymbol{y_i}}$ αναπαριστούνται στο κάτω διάγραμμα με $\mathbf{Y_i}$. Αυτά παράγονται από μια κατανομή με μέση τιμή $\widehat{\boldsymbol{\theta_i}}$ ($\boldsymbol{\theta_i}$) (υποδηλώνεται με x στο πάνω γράφημα) και διακύμανση $\mathbf{s_i}^2$.

Όταν, λοιπόν, ο μετα–αναλυτής χρησιμοποιεί το μοντέλο των τυχαίων επιδράσεων δεν ενδιαφέρεται μόνο για την "ενδο–ετερογένεια" (*within study variation*), όπως στο μοντέλο των σταθερών επιδράσεων, αλλά και για την ετερογένεια μεταξύ των μελετών (*between study variation ή heterogeneity*).

Το random effects model διαθέτει τρεις μεθόδους για την εκτίμηση των μέτρων αποτελέσματος. Αυτές είναι:



Σχήμα 5: Μέθοδοι εκτίμησης των μέτρων αποτελέσματος όταν χρησιμοποιείται το random effects model.

Η μέθοδος που χρησιμοποιείται κατά κύριο λόγο από τους αναλυτές είναι η Der Simonian Laird, η οποία είναι μη επαναληπτική.

Για το μοντέλο των τυχαίων επιδράσεων χρησιμοποιώντας ενδεικτικά την μέθοδο Der Simonian Laird και ως effect size το odds ratio οι εκτιμήτριες των αποτελεσμάτων προκύπτουν ως εξής:

Το **εκτιμώμενο μέτρο αποτελέσματος** (εδώ το OR) για την κάθε i μελέτη από τις n υπολογίζεται μέσω της σχέσης:

$$\hat{y}_i = \frac{a_i / b_i}{c_i / d_i}$$

με $i = 1, 2, \dots, n$, όπου τα $\mathbf{a_i}$, $\mathbf{b_i}$, $\mathbf{c_i}$ και $\mathbf{d_i}$ προκύπτουν από τον Πίνακα 1 αντίστοιχα για την i μελέτη.

Η **διασπορά του εκτιμώμενου μέτρου αποτελέσματος** (εδώ του OR) για κάθε i μελέτη από τις n δίνεται από τη σχέση:

$$s_i^2 = \frac{a_i + b_i + c_i + d_i}{b_i \cdot c_i}$$

με $i = 1, 2, \ldots, n$, όπου τα $\mathbf{a_i}$, $\mathbf{b_i}$, $\mathbf{c_i}$ και $\mathbf{d_i}$ προκύπτουν από τον Πίνακα 1 αντίστοιχα για την i μελέτη.

Ο **εκτιμητής της between study variation** τ^2 προκύπτει μέσω της σχέσης:

$$\tau^2 = \frac{[Q - (n-1)] \cdot \sum_{i=1}^{n} w_i}{(\sum_{i=1}^{n} w_i)^2 - \sum_{i=1}^{n} w_i^2}$$

όπου $\mathbf{w_i}$ το βάρος της i - οστής μελέτης για το μοντέλο των σταθερών επιδράσεων, $\mathbf{n}$ ο αριθμός των μελετών της μετα-ανάλυσης και **Q** η τιμή του στατιστικού του Cochrane (βλέπε βήμα 6i).

Το **βάρος** υπολογίζεται από τον γενικό τύπο:

$$w_i^* = \frac{1}{s_i^2 + \tau^2}$$

Ο **εκτιμητής του συνολικού αποτελέσματος** υπολογίζεται από τη σχέση

$$\hat{\Upsilon} = \frac{\sum_{i=1}^{n} w_i^* \hat{y}_i}{\sum_{i=1}^{n} w_i^*}$$

Η **διασπορά του εκτιμητή του συνολικού αποτελέσματος** προκύπτει από τον τύπο:

$$S^2 = \frac{1}{\sum_{i=1}^{n} w_i^*}$$

Το **διάστημα εμπιστοσύνης του εκτιμώμενου συνολικού αποτελέσματος** ορίζεται ως:

$$\hat{\Upsilon} \pm 1{,}96 \frac{1}{\sqrt{\sum_{i=1}^{n} w_i^*}}$$

- **6^{ο} Βήμα:** Ο μετα–αναλυτής οφείλει να διεξάγει κάποιες επιπλέον στατιστικές αναλύσεις, οι οποίες είναι απαραίτητες για τον σωστό και αξιόπιστο σχεδιασμό μιας μετα–ανάλυσης. Αυτές είναι:

i. **Διερεύνηση ύπαρξης στατιστικής ετερογένειας (*test for heterogeneity*):** βοηθά τον αναλυτή να καταλάβει αν οι μελέτες μετρούν μια κοινή παράμετρο (Normand, 1998) και να επιλέξει το καταλληλότερο στατιστικό μοντέλο (σταθερών ή τυχαίων επιδράσεων) για την περιγραφή των δεδομένων του. Η στατιστική ετερογένεια οφείλεται στην χρήση διαφορετικών μεθόδων από τους αναλυτές και στην απόκρυψη και μη καταγραφή χαρακτηριστικών των μελετών που επηρεάζουν το αποτέλεσμα (Kim, 2011). Δεν πρέπει σε καμία περίπτωση να συγχέεται με την κλινική ετερογένεια, που προκύπτει εξαιτίας των διαφορών μεταξύ των πληθυσμών των μελετών και της οποίας η ύπαρξη σε μια μελέτη είναι σχεδόν πάντα προφανής, λαμβάνοντας υπόψη την μοναδικότητα κάθε οργανισμού. Το παρακάτω σχεδιάγραμμα (Σχήμα 6) δείχνει πως πρέπει να αντιμετωπίζεται η ύπαρξη ετερογένειας.

Σχήμα 6: Τρόποι αντιμετώπισης ύπαρξης στατιστικής ετερογένειας.

Γενικά, όταν ανιχνεύεται στατιστική ετερογένεια δεν πρέπει ποτέ να αγνοείται από τους αναλυτές, όπως τονίζει στο βιβλίο του ο Gioacchino (2005), καθώς αυτό προσδίδει σε μια μελέτη αναξιοπιστία. Οι αιτίες της ετερογένειας θα πρέπει να ερευνηθούν και αν εντοπιστούν να ακολουθηθούν από τους μετα-αναλυτές οι κατάλληλες περαιτέρω αναλύσεις (βλέπε *subgroup analyses* και *sensitivity analysis*). Όσον αφορά την επιλογή μοντέλου βάσει της ύπαρξης ή μη ετερογένειας βλέπε παρακάτω “[ΕΠΙΛΟΓΗ ΜΟΝΤΕΛΟΥ]“.

Κατά κύριο λόγο το τεστ που χρησιμοποιείται για την ανίχνευση της στατιστικής ετερογένειας είναι το Q – test του Cochrane, το οποίο μετρά την απόκλιση του κάθε εκτιμώμενου μέτρου αποτελέσματος $\widehat{y_i}$ από τον εκτιμητή του συνολικού αποτελέσματος $\widehat{Y}$ (Kim, 2011). Η αρχική υπόθεση του Q – test είναι η H_0: $\theta_1 = \theta_2 = \theta_3 = ... = \theta_n = \Theta$, όπου n ο αριθμός των μελετών, έναντι της H_1 : τουλάχιστον ένα θ_i είναι διαφορετικό του Θ και των υπολοίπων, όπου $i = 1, 2, ..., n$. Δηλαδή, η H_0 υποθέτει ότι υπάρχει ομοιογένεια μεταξύ των μελετών, ενώ η H_1 υποθέτει ότι υπάρχει ετερογένεια (Gioacchino, 2005). Όταν ο μετα-αναλυτής χρησιμοποιεί το μοντέλο των τυχαίων επιδράσεων η H_0 ισοδυναμεί με την υπόθεση $\tau^2 = 0$ και αντίστοιχα η H_1 ισοδυναμεί με την υπόθεση $\tau^2 \neq 0$. Υπό την H_0, σύμφωνα με την Normand (1998), έχουμε:

$$Q = \sum_{i=1}^{n} w_i \left(\widehat{y_i} - \widehat{Y}\right)^2 \sim X^2_{n-1,\alpha}$$

όπου **n** ο αριθμός των μελετών της μετα–ανάλυσης, **α** το επίπεδο σημαντικότητας, $\widehat{y_i}$ το εκτιμώμενο μέτρο αποτελέσματος της i – μελέτης, w_i το βάρος της i – μελέτης και $\widehat{Y}$ το εκτιμώμενο συνολικό αποτέλεσμα.

Όταν για το στατιστικό Q ή για την p – τιμή του ελέγχου προκύπτει:

- $Q > X^2_{n-1,\alpha}$ ή $p - τιμή < \alpha$, τότε η H_0 απορρίπτεται και άρα συμπεραίνεται ότι υπάρχει στατιστική ετερογένεια μεταξύ των μελετών (Normand, 1998).

[ΕΠΙΛΟΓΗ ΜΟΝΤΕΛΟΥ] Στην περίπτωση αυτή συνιστάται από την πλειοψηφία των μελετητών η χρήση του μοντέλου των τυχαίων επιδράσεων, καθώς θεωρείται πιο συμβατικό και δίνει πιο ευρέα διαστήματα εμπιστοσύνης. Στη συνέχεια όμως ο μετα–αναλυτής οφείλει να βρει τα επιπλέον χαρακτηριστικά των μελετών που αποτελούν αιτίες της ετερογένειας, να χωρίσει τις μελέτες σε όσο το δυνατόν πιο ομοιογενή υποσύνολα με βάση τα χαρακτηριστικά που συνέλλεξε και τότε να χρησιμοποιήσει το fixed effects model για να αναλύσει ξεχωριστά το κάθε υποσύνολο (*subgroup analysis*) (Gioacchino, 2005).

Στο σύγγραμμα του Gioacchino (2005) τονίζεται βέβαια ότι κάποιοι ερευνητές υποστηρίζουν ότι πάντα σε μια μετα–ανάλυση θα πρέπει να χρησιμοποιείται το μοντέλο των τυχαίων επιδράσεων, καθώς:
→ Ακόμα και όταν δεν υπάρχει ετερογένεια, τα αποτελέσματα που δίνει είναι πολύ κοντά σε αυτά που προκύπτουν με τη χρήση του μοντέλου σταθερών επιδράσεων.

→ Σε δεδομένα που αφορούν ασθενείς υπάρχει ενδογενής κλινική ετερογένεια, εξαιτίας της μοναδικότητας που χαρακτηρίζει τον κάθε οργανισμό. Αυτού του είδους η ετερογένεια, η οποία δεν σχετίζεται με την ύπαρξη στατιστικής ετερογένειας όπως προαναφέραμε, περιγράφεται σωστότερα από το μοντέλο των τυχαίων επιδράσεων.

- $\mathbf{Q} < \mathbf{X^2_{n-1,\alpha}}$ ή $\mathbf{p} - \mathbf{τιμ}$ή $> \alpha$, τότε δεν έχουμε αρκετές ενδείξεις ώστε να απορρίψουμε την H_0, επομένως υποτίθεται ότι δεν υπάρχει στατιστική ετερογένεια μεταξύ των μελετών.

[ΕΠΙΛΟΓΗ ΜΟΝΤΕΛΟΥ] Στην περίπτωση αυτή συνίσταται η χρήση του μοντέλου των σταθερών επιδράσεων, καθώς είναι πιο συντηρητικό σε σχέση με το μοντέλο των τυχαίων επιδράσεων. Παρ' όλα αυτά θεωρείται από την πλειοψηφία των μελετητών ότι οποιοδήποτε μοντέλο και να επιλέξει ο μετα–αναλυτής, είτε σταθερών είτε τυχαίων επιδράσεων, προκύπτουν στατιστικά όμοια αποτελέσματα, από τη στιγμή που δεν υπάρχει στατιστική ετερογένεια μεταξύ των μελετών.

Γενικά, όπως αναφέρει στο άρθρο της η Normand (1998), υπάρχει ένα μεγάλο debate όσον αφορά την επιλογή του καταλληλότερου μοντέλου για την περιγραφή των δεδομένων. Γι' αυτό και ένας μετα–αναλυτής οφείλει να χρησιμοποιήσει και τα δύο μοντέλα και στη συνέχεια να συγκρίνει τα αποτελέσματα τους για να καταλήξει εν τέλει στο ποιο μοντέλο θα χρησιμοποιήσει.

Καθώς το Q – test δεν ποσοτικοποιεί την στατιστική ετερογένεια και έχει μικρή στατιστική ισχύ όταν στη μετα–ανάλυση περιέχονται λίγες μελέτες (Gioacchino, 2005), ο αναλυτής οφείλει όταν πραγματοποιεί μια μετα–ανάλυση να υπολογίζει και τον συντελεστή I^2. Ο I^2 περιγράφει το ποσοστό της μεταβλητότητας μιας εκτιμήτριας που οφείλεται στην ύπαρξη στατιστικής ετερογένειας. Ισχύει ότι $0\% \leq I^2 \leq 100\%$, με το I^2 να προκύπτει από τον τύπο:

$$I^2 = \frac{Q-df}{Q} \cdot 100\%$$

όπου $\mathbf{Q}$ η τιμή του στατιστικού του Cochrane και $\mathbf{df} = \mathbf{n} - \mathbf{1}$ οι βαθμοί ελευθερίας του (n: ο αριθμός των μελετών της μετα–ανάλυσης).

Οι Higgins et al. (2003) πρότειναν ότι όταν:

- ✓ $I^2 < 25\% \rightarrow$ υπάρχει χαμηλή στατιστική ετερογένεια
- ✓ $25\% < I^2 < 50\% \rightarrow$ υπάρχει μέτρια στατιστική ετερογένεια
- ✓ $I^2 > 75\% \rightarrow$ υπάρχει υψηλή στατιστική ετερογένεια

Παρ' όλα αυτά, σύμφωνα με τον Gioacchino (2005), οι περισσότεροι αναλυτές υποστηρίζουν ότι δεν είναι αξιόπιστο να βασιστεί κανείς στις τιμές του συντελεστή I^2, ώστε να αποφασίσει κατά πόσο χαμηλή, μέτρια ή υψηλή είναι η στατιστική ετερογένεια που εμφανίζεται στη μελέτη του.

ii. **Εκτίμηση σφάλματος δημοσίευσης** (*publication bias*): αν και υπάρχουν πολλά είδη μεροληψίας, τα οποία μπορούν να επηρεάσουν το αποτέλεσμα μιας μετα–ανάλυσης, πχ. η μεροληψία έρευνας (*search bias*), η μεροληψία αναφοράς (*reference bias*) κ.α., οι ερευνητές κατά κανόνα τις αγνοούν και ενδιαφέρονται κυρίως για το σφάλμα δημοσίευσης. Αναλυτικά για αυτό θα αναφερθούμε στο κεφάλαιο 4.

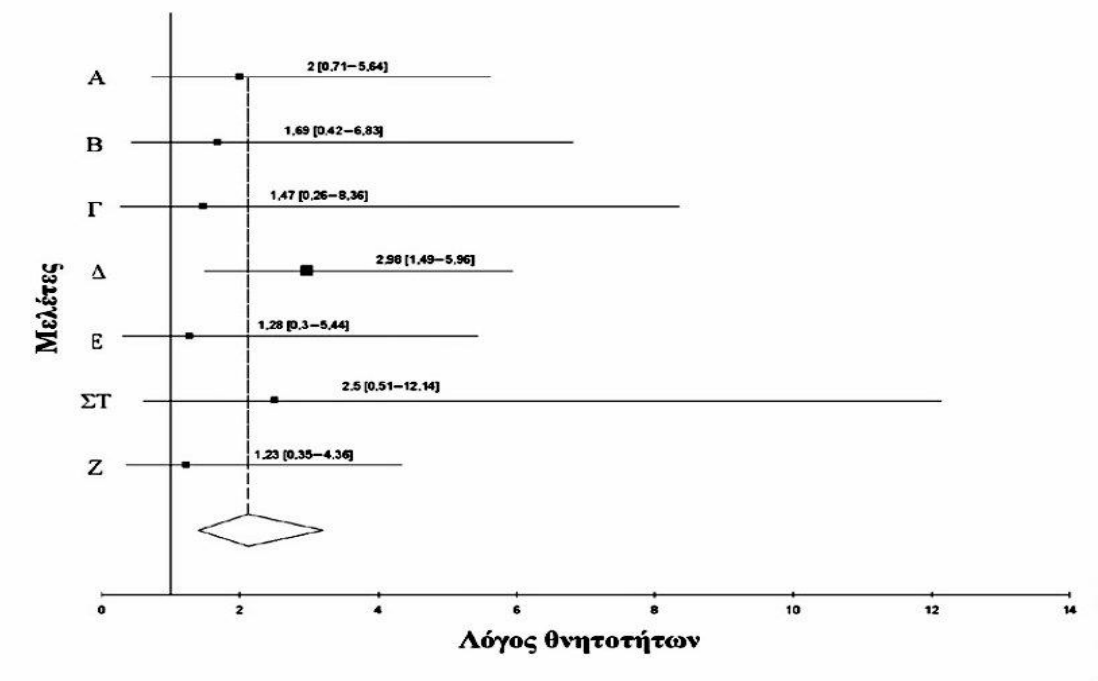
iii. **Ανάλυση υποσυνόλων** (*subgroup analysis*): η ανάλυση αυτή ρίχνει φως στις αιτίες ύπαρξης ετερογένειας (Whitehead & Whitehead, 1991). Αναλυτικά για την πραγματοποίηση αυτής θα αναφερθούμε στο κεφάλαιο 4.

iv. **Ανάλυση ευαισθησίας** (*sensitivity analysis*): αποτιμά την "ευαισθησία" των αποτελεσμάτων μιας μετα–ανάλυσης. Χρησιμεύει, δηλαδή, στην εκτίμηση της σταθερότητας των εκτιμώμενων μέτρων αποτελέσματος όταν προκύπτουν αλλαγές στα δεδομένα (Normand, 1998). Αναλυτικά για την πραγματοποίηση της συγκεκριμένης ανάλυσης θα αναφερθούμε στο κεφάλαιο 4.

- **7^{ο} Βήμα**: Αφού έχει τελειώσει με το στατιστικό κομμάτι, ο μετα–αναλυτής ερμηνεύει τα αποτελέσματα της μελέτης του. Στο σημείο αυτό δεν πρέπει να ξεχνάει ότι σκοπός του δεν είναι απλά ο υπολογισμός ενός περιληπτικού μεγέθους, αλλά η άντληση ουσιαστικών και σημαντικών αποδείξεων για κλινικά και όχι μόνο θέματα, για τα οποία μέχρι σήμερα δεν υπάρχει μια οριστική απάντηση. Γι' αυτό και οφείλει να είναι ιδιαίτερα αντικειμενικός και προσεκτικός κατά την ερμηνεία των αποτελεσμάτων (Gioacchino, 2005).

Κάθε μετα–ανάλυση συνοδεύεται από το λεγόμενο «διάγραμμα δάσος» (*forest plot*), το οποίο προσφέρει στον μετα–αναλυτή μια σαφή και άμεση εικόνα της μετα–ανάλυσης και τον βοηθά να ερμηνεύσει ευκολότερα τα αποτελέσματα της. Στο διάγραμμα αυτό απεικονίζονται τα αποτελέσματα τόσο των επιμέρους μελετών όσο και της μετα–ανάλυσης (Γαλάνης, 2009). Συγκεκριμένα, όπως

αναφέρει ο Γαλάνης (2009), στον κάθετο άξονα αντιστοιχούν οι επιμέρους μελέτες, ενώ στον οριζόντιο άξονα αντιστοιχεί το εκτιμώμενο μέτρο αποτελέσματος. Σε κάθε μελέτη αντιστοιχεί ένα τετράγωνο που αποτελεί το εκτιμώμενο μέτρο αποτελέσματος της αντίστοιχης μελέτης, καθώς και μια οριζόντια γραμμή διαμέσου του τετραγώνου, τα άκρα της οποίας αποτελούν το διάστημα εμπιστοσύνης (*confidence interval*) αυτού. Το μέγεθος των τετραγώνων εξαρτάται από το βάρος της κάθε μελέτης, με μεγαλύτερα τετράγωνα να αντιστοιχούν σε μελέτες με μεγαλύτερη βαρύτητα και μικρότερα τετράγωνα να αντιστοιχούν σε μελέτες με μικρότερη βαρύτητα. Η κάθετη γραμμή που φέρεται στην τιμή 1 ή στην τιμή 0, ανάλογα με το effect size που χρησιμοποιήθηκε (πχ. για το OR στο 1 και για την RD στο 0), αντιστοιχεί στην ύπαρξη μη στατιστικά σημαντικού αποτελέσματος ή ισοδύναμα στην μη ύπαρξη συσχέτισης μεταξύ προσδιοριστή και έκβασης. Έτσι, όταν η κάθετη αυτή γραμμή τέμνεται από την οριζόντια γραμμή κάποιας μελέτης, ή όμοια όταν το διάστημα εμπιστοσύνης περιέχει αντίστοιχα το 1 ή το 0, ο μετα–αναλυτής συμπεραίνει ότι η μελέτη αυτή δεν κατέληξε σε στατιστικά σημαντικό αποτέλεσμα, δηλαδή η διαφορά μεταξύ ομάδας αναφοράς και πειραματικής ομάδας δεν είναι στατιστικά σημαντική. Αυτό είναι πιθανό να οφείλεται στο μικρό μέγεθος ή στην χαμηλή στατιστική ισχύ που μπορεί να έχει μια μελέτη (Gioacchino, 2005). Όσον αφορά το εκτιμώμενο συνολικό αποτέλεσμα, αυτό απεικονίζεται στο forest plot με ένα " διαμάντι ", τα άκρα του οποίου αποτελούν το διάστημα εμπιστοσύνης του. Όλα αυτά τα κατανοεί κανείς καλύτερα παρατηρώντας το παρακάτω σχήμα (Σχήμα 7).



Σχήμα 7: «Διάγραμμα δάσος» μετα-ανάλυσης υποθετικών δεδομένων 7 κλινικών δοκιμών, όπως αυτό απεικονίζεται στο άρθρο του Γαλάνη (2009). Σύγκριση της αποτελεσματικότητας της στρεπτοκινάσης

(πειραματική ομάδα) έναντι του ενεργοποιητή του ιστικού προδρόμου της πλασμίνης (ομάδα αναφοράς) για την πρόληψη του θανάτου (έκβαση) έπειτα από οξύ έμφραγμα του μυοκαρδίου. Για την πραγματοποίηση της μετα-ανάλυσης επιλέχθηκε το μοντέλο των σταθερών επιδράσεων με χρήση της μεθόδου Mantel – Haenszel και του λόγου θνητοτήτων (OR) ως effect size.

Έτσι, στο Σχήμα 7 παρατηρούμε για παράδειγμα ότι για την μελέτη Α το εκτιμώμενο μέτρο αποτελέσματος είναι ίσο με 2 και έχει διάστημα εμπιστοσύνης [0.71, 5.64]. Εξαιτίας του ότι στο διάστημα εμπιστοσύνης περιέχεται η μονάδα συμπεραίνουμε ότι η μελέτη αυτή δεν καταλήγει σε στατιστικά σημαντικό αποτέλεσμα. Το τετράγωνο που απεικονίζει το μέτρο αποτελέσματος της μελέτης Δ παρατηρούμε ότι είναι το μεγαλύτερο σε σχέση με τα υπόλοιπα, άρα συμπεραίνουμε ότι η μελέτη Δ συμβάλλει περισσότερο από όλες τις μελέτες στον υπολογισμό του συνολικού αποτελέσματος. Παρά το γεγονός ότι 6 από τις 7 μελέτες κατέληξαν σε μη στατιστικά σημαντικά αποτελέσματα, το εκτιμώμενο συνολικό αποτέλεσμα φαίνεται ότι είναι στατιστικά σημαντικό, καθώς στο διάστημα εμπιστοσύνης του (αριστερό και δεξί άκρο διαμαντιού) δεν περιέχεται η μονάδα (το διαμάντι δεν τέμνετε από την κάθετη γραμμή που φέρεται στην τιμή 1). Οπότε ο μετα–αναλυτής οδηγείται στο συμπέρασμα ότι η θνητότητα σε αυτούς που ελάμβαναν στρεπτοκινάση ήταν περίπου δύο φορές μεγαλύτερη σε σχέση με εκείνους που ελάμβαναν ενεργοποιητή του ιστικού προδρόμου της πλασμίνης και ότι η σχέση αυτή είναι στατιστικά σημαντική (Γαλάνης, 2009).

Το παράδειγμα αυτό είναι χαρακτηριστικό της συνεισφοράς της μετα–ανάλυσης στην εξαγωγή ασφαλέστερων συμπερασμάτων και στον υπολογισμό ενός συγκεντρωτικού αποτελέσματος με μεγαλύτερη ακρίβεια και εγκυρότητα.

1.3 Κανόνες για μια πρώτη αξιολόγηση της μετα–ανάλυσης

Σε αυτή την παράγραφο θα τονίσουμε κάποιους βασικούς κανόνες, τους οποίους θα πρέπει να τηρούν οι συγγραφείς ώστε να συνθέσουν μια πιο αξιόπιστη μετα–ανάλυση. Συγκεκριμένα σύμφωνα με τον Gioacchino (2005) οι μετα–αναλυτές οφείλουν:

i. Να βασιστούν σε ένα συγκεκριμένο πρωτόκολλο για τον τρόπο εργασίας τους.
ii. Να περιγράψουν ξεκάθαρα την ερευνητική τους στρατηγική (πηγές, κριτήρια αποδοχής ή απόρριψης μιας μελέτης).
iii. Να αξιολογήσουν την ποιότητα των μελετών που εντάχθηκαν στην μετα–ανάλυση, καθώς και την ομοιογένεια στα κριτήρια απόρριψης και αποδοχής κάθε μελέτης σε κάθε paper.

iv. Να καταγράψουν επακριβώς το γιατί απέρριψαν τις μελέτες που απέρριψαν, αλλά και να αξιολόγησαν αναλυτικά αυτά τα dropouts.
v. Να παρουσιάσουν αναλυτικά τα original data που θα χρησιμοποιήσουν στην ανάλυση τους, καθώς η απουσία – απόκρυψη τους αποτελεί μεγάλο σφάλμα και υποδηλώνει την πιθανή πρόθεση των συγγραφέων να θέλουν να αποφύγουν την επανεξέταση των ευρημάτων τους από άλλους ερευνητές.
vi. Να παρουσιάσουν αναλυτικά και προσεκτικά τις ερωτήσεις οι οποίες απαντώνται μέσω της μετα–ανάλυσης τους.
vii. Να συμπεριλάβουν στην μελέτη τους τα χαρακτηριστικά των ασθενών που τυχόν καταγράφονται στα papers που εντάχθηκαν στην μετα–ανάλυση.
viii. Να χρησιμοποιήσουν γραφική αναπαράσταση των αποτελεσμάτων.
ix. Να ερευνήσουν το ποσό της ετερογένειας που προέκυψε από την ανάλυση, καθώς και τις αιτίες αυτού.
x. Να αναφέρουν πως υπολογίστηκε το συνολικό θεραπευτικό κέρδος, αν αυτό υπάρχει, καθώς και αν το αποτέλεσμα αυτό της μετα–ανάλυσης είναι σωστό και ολοκληρωμένο.
xi. Να λάβουν υπ’ όψιν τους την μεροληψία ή σφάλμα δημοσίευσης, αν αυτό υπάρχει.
xii. Να πραγματοποιήσουν ανάλυση ευαισθησίας.
xiii. Να ζητήσουν την γνώμη ενός έμπειρου στατιστικολόγου ώστε να επικυρώσουν τις μεθόδους και τα ευρήματα τους.

Ενώ οι παραπάνω κανόνες συμβάλλουν στην πραγματοποίηση μιας αξιόπιστης μετα–ανάλυσης, δεν είναι όλοι κοινά αποδεκτοί από όλους τους ερευνητές. Αναλυτικά με την αξιολόγηση της μετα–ανάλυσης θα ασχοληθούμε στο κεφάλαιο 4.

ΚΕΦΑΛΑΙΟ 2 : ΤΑΞΙΝΟΜΗΣΗ

2.1 Εισαγωγή στην Ταξινόμηση

2.1.1 Τι είναι Ταξινόμηση

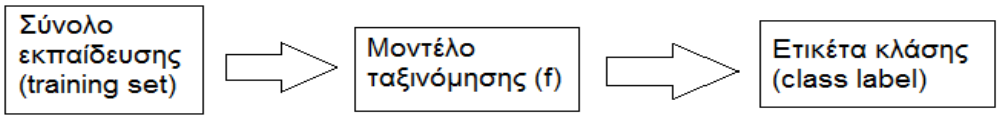
Η ταξινόμηση (*classification*) είναι μια από τις πιο συχνές ενέργειες του ανθρώπινου εγκεφάλου. Για παράδειγμα, όταν θέλουμε να αποφασίσουμε ποιό φαγητό θα διαλέξουμε σε ένα εστιατόριο ταξινομούμε στο μυαλό μας με βάση την όρεξη μας τις πιθανές επιλογές πιάτων σε ομάδες - κλάσεις (*classes*). Στη συνέχεια διακρίνουμε το καταλληλότερο πιάτο σκεπτόμενοι διάφορους παράγοντες, όπως τον χρόνο παρασκευής, το κόστος, τις θερμίδες κ.α. Η ταξινόμηση, δηλαδή, εξετάζει τα χαρακτηριστικά ενός νέου αντικειμένου (στο παράδειγμα μας τα χαρακτηριστικά των φαγητών) και με βάση αυτά το κατατάσσει σε ένα προκαθορισμένο σύνολο κλάσεων.

Επισήμως, στην στατιστική με τον όρο classification εννοούμε μια σειρά ενεργειών, κατά την οποία βρίσκονται οι κοινές ιδιότητες των αντικειμένων ενός συνόλου μιας βάσης δεδομένων και στη συνέχεια τα αντικείμενα αυτά ομαδοποιούνται σε διαφορετικές κλάσεις – τάξεις σύμφωνα με ένα μοντέλο ταξινόμησης. Πως προκύπτει όμως αυτό το μοντέλο ταξινόμησης; Χρησιμοποιώντας ένα δοθέν σώμα δεδομένων εφοδιασμένο με ετικέτες (*label*) κλάσης, το λεγόμενο σύνολο εκπαίδευσης (*training set*) (βλέπε Πίνακα 2), ο ταξινομητής (η επιλεγμένη από τον ερευνητή μέθοδος ταξινόμησης) «εκπαιδεύεται» αναλύοντας και αξιολογώντας τα χαρακτηριστικά του training set και έτσι αναπτύσσει για κάθε μια κλάση ένα μοντέλο, το οποίο εφαρμόζεται μελλοντικά για να κατατάξει όσο το δυνατόν σωστότερα νέα αντικείμενα της βάσης δεδομένων στις κλάσεις. Αυτός είναι ουσιαστικά και ο σκοπός της ταξινόμησης. Στη συνέχεια, για να μελετηθεί η ακρίβεια του προκύπτοντος μοντέλου χρησιμοποιείται ένα σύνολο ελέγχου (*test set*), δηλαδή ένα σώμα δεδομένων που δεν γνωρίζουμε την κλάση τους (βλέπε Πίνακα 3).

Σε μια πρόταση: **η ταξινόμηση είναι το γενικό πρόβλημα της ανάθεσης ενός αντικειμένου σε μία ή περισσότερες προκαθορισμένες κλάσεις** (Tan et al., 2006).

2.1.2 Μαθηματικός Ορισμός της Ταξινόμησης

Το πρόβλημα της ταξινόμησης συνίσταται στον προσδιορισμό της απεικόνισης $f : T \rightarrow C$. Με $T = \{t_1, t_2, \dots, t_n\}$ συμβολίζουμε το σύνολο εκπαίδευσης (*training set*), το οποίο προέρχεται από μια μεγαλύτερη βάση δεδομένων, έστω B. Κάθε t_i του συνόλου εκπαίδευσης καλείται στοιχείο ή εγγραφή ή παράδειγμα και αποτελείται από όλα τα πολλαπλά χαρακτηριστικά των εγγραφών της B καθώς και από μια ετικέτα (*label*) κλάσης, η οποία υποδεικνύει την κλάση που ανήκει το αντίστοιχο t_i. Με $C = \{c_1, c_2, \dots, c_n\}$ συμβολίζουμε το σύνολο των κλάσεων. Κάθε c_i ορίζει ένα σύνολο, στο οποίο περιέχονται οι εγγραφές – στοιχεία – παραδείγματα που ανήκουν στην κλάση αυτή. Η απεικόνιση f αποτελεί λοιπόν το μοντέλο ταξινόμησης και αντιστοιχεί κάθε t_i σε μια κλάση c_i, διαμερίζοντας έτσι το T σε κλάσεις ισοδυναμίας.



Σχήμα 8: Με χρήση ενός training set **T** καταλήγουμε μέσω ενός μοντέλου **f** σε ένα σύνολο **C** από ετικέτες κλάσης.

Πίνακας 2: Παράδειγμα συνόλου εκπαίδευσης (*training set*). Για την χορήγηση επιδόματος ενοικίου, στην περίπτωση όπου ο ενδιαφερόμενος μένει μόνος του και δεν έχει οικογένεια, ισχύουν οι παρακάτω προϋποθέσεις: α) ο ενδιαφερόμενος πρέπει να έχει ετήσιο εισόδημα το πολύ 2.400 €, β) η αξία της ακίνητης περιουσίας του να κυμαίνεται μεταξύ 90.000 € – 200.000 € και γ) οι τραπεζικές καταθέσεις του να είναι μέχρι και το διπλάσιο του ετήσιου εισοδήματος του.

	Χαρακτηριστικό 1	Χαρακτηριστικό 2	Χαρακτηριστικό 3	Κλάση
t_i	Ετήσιο εισόδημα (€)	Αξία ακίνητης περιουσίας (€)	Τραπεζικές καταθέσεις (€)	Δικαίωμα χορήγησης επιδόματος ενοικίου
t_1	1.500	95.000	800	Ναι
t_2	2.300	250.000	1.000	Όχι
t_3	4.000	100.000	5.000	Όχι
t_4	2.300	200.000	6.300	Όχι
t_5	2.000	185.000	1.800	Ναι
t_6	3.500	300.000	1.000	Όχι

Πίνακας 3: Παράδειγμα συνόλου ελέγχου (*test set*).

t_i	Χαρακτηριστικό 1	Χαρακτηριστικό 2	Χαρακτηριστικό 3	Κλάση
	Ετήσιο εισόδημα (€)	Αξία ακίνητης περιουσίας (€)	Τραπεζικές καταθέσεις (€)	Δικαίωμα χορήγησης επιδόματος ενοικίου
t_1	3.000	98.000	1.000	?
t_2	2.300	160.000	4.000	?
t_3	1.800	100.000	7.300	?
t_4	2.250	80.000	600	?
t_5	4.800	130.000	6.500	?
t_6	1.500	60.000	1.000	?

2.1.3 Εφαρμογές της Ταξινόμησης

Η ταξινόμηση αποτελεί αντικείμενο μελέτης της στατιστικής, της μηχανικής μάθησης (*machine learning*) και της εξόρυξης γνώσης (*data mining*) και βρίσκει εφαρμογή στην ιατρική, στην βιολογία, στο marketing, στα χρηματοοικονομικά, στην οπτική αναγνώριση χαρακτήρων (*optical character recognition -> OCR*) και αλλού. Πιο συγκεκριμένα στα συγγράμματα των Hand (1981) και Tan et al. (2006) αναφέρονται οι παρακάτω εφαρμογές της κατηγοριοποίησης:

i. Αναγνώριση της εθνικότητας του ιδιοκτήτη ενός αρχαίου κρανίου, μια από τις πρώτες προσπάθειες κατηγοριοποίησης.
ii. Εκτίμηση της απόδοσης καλλιεργειών και αναγνώριση πιθανών επιβλαβών ασθενειών από φωτογραφίες τραβηγμένες από μεγάλο υψόμετρο.
iii. Απόφαση του πιο κατάλληλου τύπου εγχείρησης σε γυναίκες που πάσχουν από καρκίνο του μαστού, για την πρόβλεψη ισχαιμικών ασθενειών, για την πρόβλεψη υποτροπιάζουσας φυματίωσης κ.α.
iv. Αναγνώριση της ομιλίας, κατά την οποία τα αντικείμενα που είναι προς ταξινόμηση είναι ηχητικές κυματομορφές.
v. Έγκαιρη αναγνώριση διαταραγμένων προσωπικοτήτων ή ψυχικών ασθενειών.
vi. Εντοπισμός spam emails με βάση πχ. την επικεφαλίδα τους ή το περιεχόμενο τους.
vii. Πρόβλεψη καρκινικών κυττάρων χαρακτηρίζοντας τα καλοήθη ή κακοήθη.
viii. Κατηγοριοποίηση συναλλαγών με πιστωτικές κάρτες ως νόμιμες ή προϊόν απάτης.
ix. Κατηγοριοποίηση δευτερευόντων δομών πρωτεΐνης ως alpha – helix, beta – sheet ή random coil.
x. Χαρακτηρισμός ειδήσεων ως οικονομικές, αθλητικές, πολιτιστικές, πρόβλεψης καιρού, κλπ.

2.2 Εισαγωγή στις Μεθόδους Ταξινόμησης

Οι πιο κοινές μέθοδοι ταξινόμησης είναι τα δέντρα απόφασης (*decision trees*), τα τεχνητά νευρωνικά δίκτυα (*Artificial Neural Networks*), η λογιστική παλινδρόμηση (*logistic regression*), τα Μπεϋζιανά μοντέλα δικτύου (*Bayesian network models*) και οι μηχανές διανυσμάτων υποστήριξης (*Support Vector Machines → SVM*).

2.2.1 Δέντρα απόφασης (decision trees)

Τα δέντρα απόφασης αποτελούν μία από τις πιο διαδεδομένες και ευρέως χρησιμοποιούμενες τεχνικές ταξινόμησης, καθώς είναι εύκολα στη χρήση και προσφέρουν σαφή και κατανοητά αποτελέσματα σε σύντομο χρονικό διάστημα. Οι εσωτερικοί κόμβοι ενός τέτοιου δέντρου αντιστοιχούν σε ένα από τα χαρακτηριστικά του προβλήματος και τα φύλλα αποτελούν τις κλάσεις, ενώ κάθε ακμή είναι μια πιθανή τιμή ενός χαρακτηριστικού.

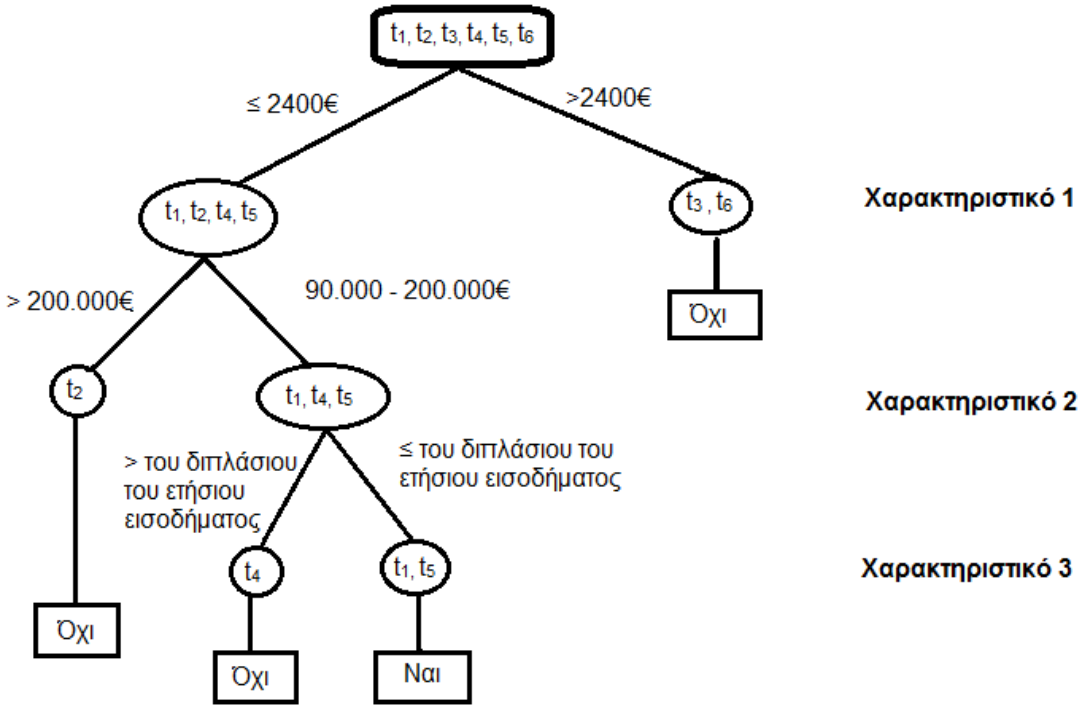
Κατασκευή

Λαμβάνοντας υπ' όψιν το σύγγραμμα του Tan et al. (2006), για την κατασκευή ενός δέντρου απόφασης ορίζουμε τα εξής βήματα:

1. Ξεκινάμε με έναν κόμβο που περιέχει όλες τις εγγραφές του συνόλου εκπαίδευσης.
2. Διασπάμε τον κόμβο με βάση μια συνθήκη διαχωρισμού, η οποία καθορίζεται από κάποιο κριτήριο, όπως το κέρδος πληροφορίας (*information gain*), τον δείκτη Gini (*Index Gini*) ή τον λόγο κέρδους (*gain ratio*). Δημιουργούμε δηλαδή υποσύνολα των εγγραφών, που το καθένα περιέχει λιγότερα στοιχεία από το αρχικό σύνολο.
3. Αναδρομικά σε κάθε κόμβο καλούμε το βήμα 2. Με αυτόν τον τρόπο πραγματοποιούμε την λεγόμενη ομοιογένεια υποσυνόλων, δηλαδή εξασφαλίζουμε ότι τα στοιχεία κάθε υποσυνόλου – οι εγγραφές κάθε κόμβου έχουν όσο το δυνατόν την ίδια τιμή στο χαρακτηριστικό, το οποίο αντιστοιχεί ο κόμβος.
4. Η διάσπαση ενός κόμβου σταματά όταν όλες οι εγγραφές του ανήκουν στην ίδια κλάση.

Λειτουργία

Για την ταξινόμηση ενός άγνωστου συνόλου εγγραφών ξεκινάμε από την ρίζα του δέντρου και κατευθυνόμαστε προς τους υπόλοιπους κόμβους. Στην πορεία αυτή εξετάζονται σε κάθε κόμβο οι τιμές των χαρακτηριστικών κάθε εγγραφής, μέχρι να καταλήξουμε σε ένα φύλλο, όπου και τελικά η εγγραφή ταξινομείται στην κλάση που αντιπροσωπεύει το φύλλο αυτό.



Σχήμα 9: Παράδειγμα δέντρου απόφασης. Χρησιμοποιώντας το training set του Πίνακα 2 κατασκευάσαμε ένα δέντρο απόφασης, με στόχο την εκπαίδευση αυτού για μετέπειτα ταξινομήσεις ατόμων, που ενδιαφέρονται για την χορήγηση επιδόματος ενοικίου.

Σοβαρό μειονέκτημα των δέντρων απόφασης ως τεχνική ταξινόμησης είναι ότι δεν μπορούν να χειριστούν περίπλοκες σχέσεις μεταξύ χαρακτηριστικών και επίσης στην περίπτωση ελλιπών δεδομένων παρουσιάζουν δυσλειτουργία.

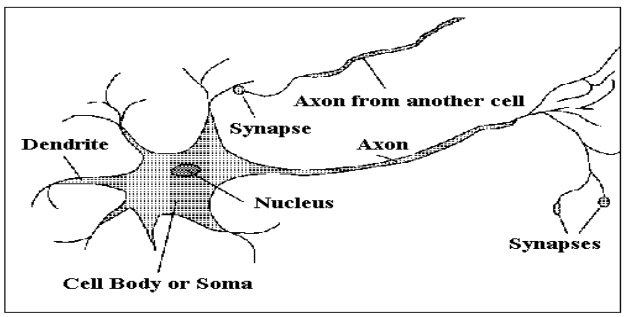
2.2.2 Τεχνητά Νευρωνικά Δίκτυα (Artificial Neural Networks)

Γενικά, νευρωνικό δίκτυο είναι ένα σύνολο κόμβων που συνεργάζονται για να επιτελέσουν κάποιο σκοπό. Τα τεχνητά νευρωνικά δίκτυα δεν είναι τίποτα άλλο παρά απομιμήσεις των βιολογικών νευρωνικών δικτύων και αναπτύχθηκαν στην προσπάθεια ανακάλυψης ενός νέου υπολογιστικού μοντέλου με δικτυακή δομή παρόμοια με αυτή του εγκεφάλου. Παράδειγμα τεχνητού νευρωνικού δικτύου είναι η εφαρμογή αναγνώρισης τραγουδιών Shazam.

Δομή

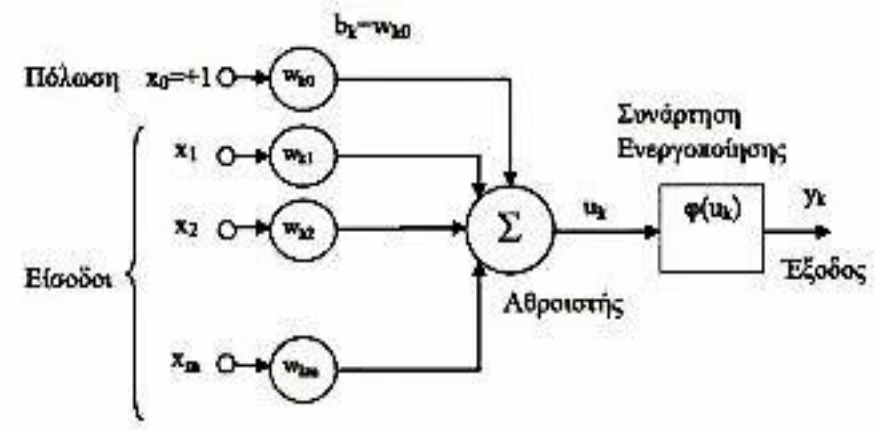
Τα τεχνητά νευρωνικά δίκτυα έχουν δομή παρόμοια, ή ίσως και απλούστερη, με αυτή ενός βιολογικού νευρωνικού δικτύου. Συγκεκριμένα, στον ανθρώπινο εγκέφαλο οι κόμβοι είναι τα νευρικά κύτταρα (νευρώνες). Οι νευρώνες αποτελούνται από 4 μέρη:

1. Το κυρίως σώμα (*body*): ο πυρήνας του νευρώνα. Σε αυτό επιτελούνται όλες οι διεργασίες.
2. Τον άξονα (*axon*): η πύλη εξόδου του νευρώνα. Μέσω αυτού κάθε νευρώνας στέλνει ηλεκτρικά σήματα σε όλους τους υπόλοιπους νευρώνες που είναι συνδεδεμένος.
3. Τους δενδρίτες (*dendrites*): οι πύλες εισόδου του νευρώνα. Μέσω αυτών ο νευρώνας λαμβάνει σήματα από γειτονικούς νευρώνες.
4. Τις συνάψεις (*synapses*): το «κενό» σημείο ένωσης του άξονα του νευρώνα με τους δενδρίτες άλλων νευρώνων. Μέσω αυτών μεταδίδεται το τελικό ποσοστό της ηλεκτρικής δραστηριότητας του άξονα, το λεγόμενο συναπτικό βάρος, στους δενδρίτες των γειτονικών νευρώνων. Οι συνάψεις χωρίζονται σε ανασταλτικές και ενισχυτικές ανάλογα με τον αν το φορτίο που εκλύεται από τη σύναψη καταστέλλει τον νευρώνα ή αντίστοιχα τον ερεθίζει να παράγει παλμούς με μεγαλύτερη συχνότητα.



Σχήμα 10: Δομή βιολογικού νευρωνικού δικτύου.

Όσον αφορά τώρα τον τεχνητό νευρώνα, αποτελείται από ένα σύνολο συνάψεων ή διασυνδέσεων, καθεμία από τις οποίες χαρακτηρίζεται από το δικό της βάρος (*weight*) w_{ki}, όπου $i = 1, 2, 3, \dots, m$ η σύναψη και k ο νευρώνας. Η τιμή βάρους μπορεί να είναι θετική ή αρνητική ανάλογα με το αν η λειτουργία της σύναψης είναι ενισχυτική ή ανασταλτική αντίστοιχα. Τα $x_1, x_2, \dots, x_m$ αποτελούν τα σήματα εισόδου που δέχεται το μοντέλο του τεχνητού νευρώνα.



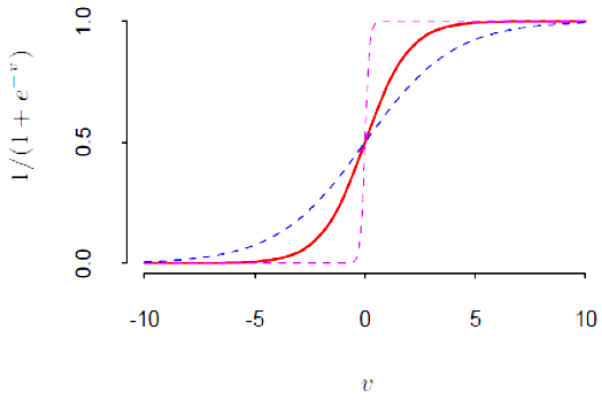
Σχήμα 11: Αναπαράσταση τεχνητού νευρώνα.

Το σώμα του τεχνητού νευρώνα αποτελείται από δύο μέρη:

1. Τον αθροιστή Σ, ο οποίος προσθέτει τα επηρεασμένα από τα βάρη σήματα εισόδου παράγοντας την ποσότητα $u_k = \sum_{i=0}^{m} w_{ki} x_i$.
2. Την συνάρτηση ενεργοποίησης (*activation function*) $\varphi(u_k)$, η οποία είναι ένας μη γραμμικός μετασχηματιστής, που δίνει το σήμα εξόδου y_k του νευρώνα ($\varphi(u_k) = y_k$). Συνήθως το εύρος τιμών του σήματος εξόδου είναι το διάστημα [0,1] ή το [-1,1] ανάλογα με την συνάρτηση ενεργοποίησης που χρησιμοποιείται. Διευκρινίζεται ότι η τιμή της εξόδου του νευρώνα είναι μοναδική.

Υπάρχουν πολλά είδη συναρτήσεων ενεργοποίησης (βηματική, γραμμική, μη γραμμική κ.α.), αλλά αυτή που χρησιμοποιείται συνήθως στην κατασκευή νευρωνικών δικτύων είναι η σιγμοειδής συνάρτηση (*sigmoid function*):

$$\sigma(v) = \frac{1}{1 + e^{-v}}$$



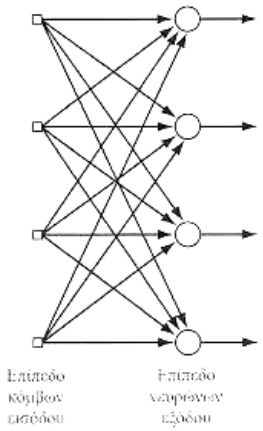
Σχήμα 12: Γράφημα σιγμοειδούς συνάρτησης (κόκκινη καμπύλη). Περιλαμβάνονται οι $\boldsymbol{\sigma}(\mathbf{sv})$ για s = ½ (μπλε διακεκομμένη καμπύλη) και s = 10 (μωβ διακεκομμένη καμπύλη) (Hastie et al., 2001).

Ένα τεχνητό νευρωνικό δίκτυο περιλαμβάνει επίσης μια εξωτερικά εφαρμοζόμενη πόλωση $b_{k,}$ η οποία ισούται με το βάρος w_{k0} της σταθερής εισόδου $x_0 = 1$. Ανάλογα με το αν είναι αρνητική ή θετική, η πόλωση συμβάλλει αντίστοιχα στην μείωση ή αύξηση της δικτυακής διέγερσης της συνάρτησης ενεργοποίησης.

Αρχιτεκτονική

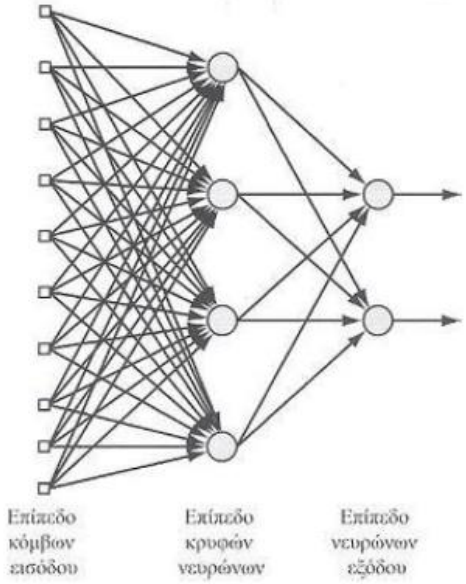
Οι νευρώνες ενός νευρωνικού δικτύου οργανώνονται με τη μορφή επιπέδων και ο τρόπος με τον οποίο αυτοί είναι δομημένοι εξαρτάται από το πώς έχει εκπαιδευτεί το μοντέλο. Υπάρχουν τρεις κατηγορίες αρχιτεκτονικών δικτύων:

1. Δίκτυα πρόσθιας τροφοδότησης ενός επιπέδου (*feedforward*)



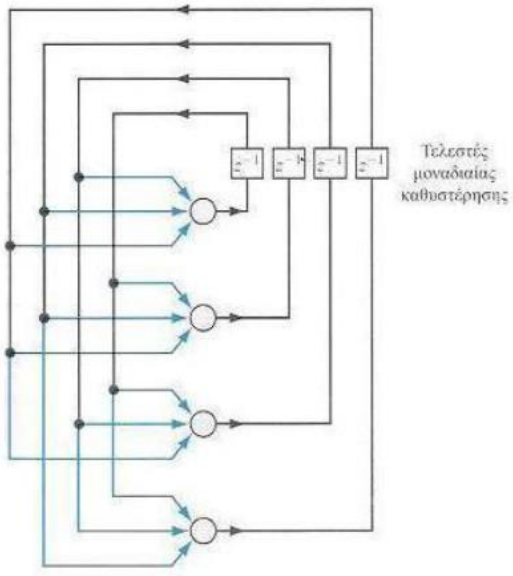
Σχήμα 13: Δίκτυα πρόσθιας τροφοδότησης ενός επιπέδου (*feedforward*) (Haykin, 2009).

2. Πολυεπίπεδα δίκτυα πρόσθιας τροφοδότησης



Σχήμα 14: Παράδειγμα δικτύου πρόσθιας τροφοδότησης με ένα κρυφό επίπεδο (Haykin, 2009).

3. Αναδρομικά δίκτυα (*recurrent neural networks*)



Σχήμα 15: Παράδειγμα αναδρομικού δικτύου (Haykin, 2009).

Λειτουργίες

Τα τεχνητά νευρωνικά δίκτυα έχουν τις εξής δύο βασικές λειτουργίες:

1. Μάθηση (*learning*) ή Εκπαίδευση (*training*): η τροποποίηση των τιμών των βαρών του δικτύου, ώστε δοθέντος μιας εισόδου να παραχθεί μια συγκεκριμένη έξοδος. Όπως αναφέραμε και στην παράγραφο 2.1.1, η μάθηση – εκπαίδευση ενός μοντέλου ταξινόμησης γίνεται με χρήση ενός training set. Το ίδιο συμβαίνει και όταν χρησιμοποιούμαι τα τεχνητά νευρωνικά δίκτυα, δηλαδή δοθέντος του training set τα συναπτικά βάρη μεταβάλλονται κατάλληλα, ώστε να δώσουν την προβλεπόμενη από το training set έξοδο. Για την εκπαίδευση ενός τεχνητού νευρωνικού δικτύου υπάρχουν πολλά είδη μάθησης, καθένα από τα οποία αλλάζει με διαφορετικό τρόπο τις παραμέτρους. Συγκεκριμένα: α) η μάθηση με επίβλεψη (*supervised learning*), β) η μάθηση χωρίς επίβλεψη (*unsupervised learning*) και γ) η βαθμολογημένη ή διαβαθμισμένη μάθηση (*graded learning*).
2. Ανάκληση (*recall*): ο υπολογισμός μιας εξόδου λαμβάνοντας συγκεκριμένες τιμές εισόδου και βαρών.

Πλεονεκτήματα

Τα τεχνητά νευρωνικά δίκτυα χρησιμοποιούνται ευρέως τα τελευταία χρόνια σε διάφορους τομείς λόγω της υπολογιστικής τους ταχύτητας, της δυνατότητας αντιμετώπισης πολύπλοκων μη γραμμικών λειτουργιών και της ικανότητας τους να αναγνωρίζουν τις σχέσεις μεταξύ ποσοτήτων, που είναι γενικά δύσκολο να μοντελοποιηθούν. Επίσης, πολύ βασική και σημαντική τους ικανότητα είναι αυτή της γενίκευσης, δηλαδή της επιτυχούς εκτίμησης των κλάσεων νέων εισόδων και όχι μόνο αυτών που χρησιμοποιήθηκαν για την εκπαίδευση του τεχνητού νευρωνικού δικτύου.

2.2.3 Λογιστική Παλινδρόμηση (Logistic Regression)

Η λογιστική παλινδρόμηση είναι μια ευρέως χρησιμοποιούμενη τακτική που άρχισε να γίνεται πιο γνωστή κατά τη δεκαετία του 50 μέσω της εφαρμογής της σε θέματα βιοστατιστικής. Τα μοντέλα της λογιστικής παλινδρόμησης έχουν αρκετά κοινά με αυτά της γραμμικής παλινδρόμησης, είναι ευέλικτα, εύκολα στην ερμηνεία και συχνά αρκετά ακριβή. Είναι χρήσιμα σε καταστάσεις όπου επιθυμείται η πρόβλεψη της ύπαρξης ή απουσίας ενός χαρακτηριστικού ή ενός συμβάντος.

Συγκεκριμένα, μέσω της λογιστικής παλινδρόμησης κατασκευάζεται ένα μη γραμμικό μοντέλο, το οποίο χρησιμοποιεί ένα σύνολο ανεξάρτητων μεταβλητών $x_1, x_2, \dots, x_n$, όπου $n \in \mathbb{N}$, για να περιγράψει την μεταβολή μιας κατηγορικής (*categorical*) εξαρτημένης μεταβλητής y ανάλογα με τις τιμές των συντελεστών $\beta_1, \beta_2, \dots, \beta_n$ των $x_1, x_2, \dots, x_n$ αντίστοιχα.

- **Απλό Λογιστικό Μοντέλο:** Θα αναφερθούμε στην περίπτωση όπου η εξαρτημένη μεταβλητή y παίρνει μόνο δύο τιμές, συνήθως 0 και 1, καθεμία από τις οποίες αντιστοιχεί σε μια κατηγορία. Η περίπτωση όπου η μεταβλητή y έχει περισσότερες από δύο κατηγορίες είναι εκτός των σκοπών της παρούσας εργασίας και δεν θα εξεταστεί. Ορίζουμε, λοιπόν, την τιμή $y = 1$ ως «επιτυχία» και την τιμή $y = 0$ ως «αποτυχία». Οπότε έχουμε ότι:

$$y \sim B(p)$$

Δηλαδή, η μεταβλητή y είναι της κατανομής Bernoulli με πιθανότητα επιτυχίας $p = P[y = 1]$, πιθανότητα αποτυχίας $1 - p = P[y = 0]$, αναμενόμενη τιμή $E[y] = p$ και διασπορά $V[y] = p(1 - p)$. Προφανώς ισχύει $0 \leq E[y] \leq 1$.
Μέσω του απλού γραμμικού μοντέλου προκύπτει ότι:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

όπου ε είναι το τυχαίο σφάλμα, για το οποίο ισχύει $\varepsilon \sim N(0, \sigma^2)$. Άρα:

$$E[y] = E[\beta_0 + \beta_1 x + \varepsilon] = \beta_0 + \beta_1 x + E[\varepsilon] = \beta_0 + \beta_1 x$$

Όμως, η αναμενόμενη τιμή $E[y]$ που δίνει η παραπάνω σχέση κυμαίνεται σε όλο το σύνολο των πραγματικών αριθμών. Οπότε, για να εξασφαλίσουμε τον περιορισμό $0 \leq E[y] \leq 1$, που επιβάλλει η δυαδικότητα της εξαρτημένης μεταβλητής y, χρησιμοποιούμε τον logit μετασχηματισμό της πιθανότητας p:

$$p' = \ln\left(\frac{p}{1-p}\right)$$

όπου το πηλίκο $\frac{p}{1-p}$ ονομάζεται odds, όπως αναφέραμε και στο κεφάλαιο 1.

Επομένως, έχουμε:

$$\boldsymbol{\beta_0 + \beta_1 x = \ln\left(\frac{p}{1-p}\right)} \quad \Leftrightarrow \quad \frac{p}{1-p} = e^{\beta_0 + \beta_1 x} \quad \Leftrightarrow \quad p = e^{\beta_0 + \beta_1 x}(1-p)$$

$$\Leftrightarrow \quad p = e^{\beta_0 + \beta_1 x} - p\, e^{\beta_0 + \beta_1 x} \Leftrightarrow p + p e^{\beta_0 + \beta_1 x} = e^{\beta_0 + \beta_1 x}$$

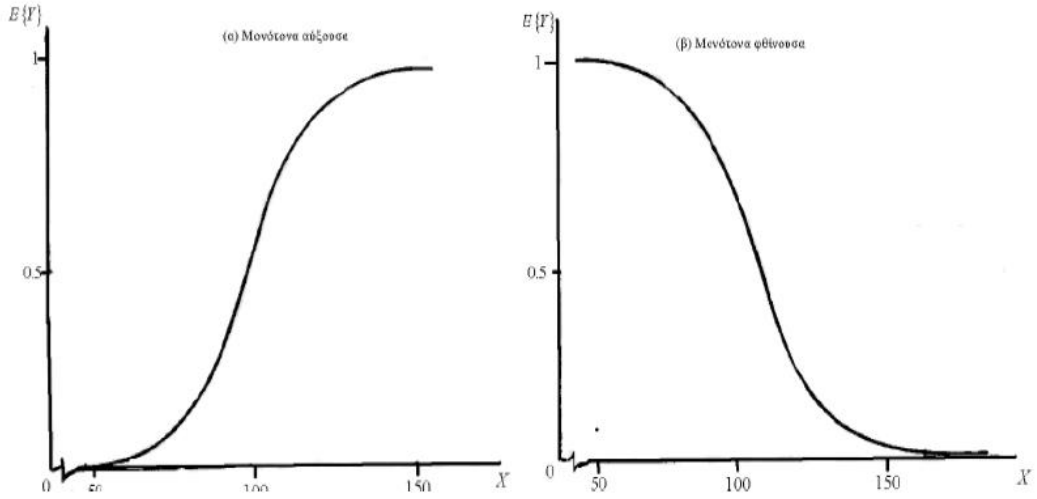
$$\Leftrightarrow \quad p\left(1 + e^{\beta_0 + \beta_1 x}\right) = e^{\beta_0 + \beta_1 x} \quad \Leftrightarrow \quad \boldsymbol{p = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}}$$

Το λογιστικό μοντέλο ορίζεται λοιπόν από τη σχέση:

$$y = E[y] + \varepsilon$$

όπου

$$E[y] = p(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$



Σχήμα 16: Απεικόνιση αναμενόμενης τιμής απλού λογιστικού μοντέλου.

Για να εκτιμήσουμε τις παραμέτρους του μοντέλου, δηλαδή τα β_0 και β_1, κάνουμε χρήση της μεθόδου της μέγιστης πιθανοφάνειας (*maximum likelihood method*).

- **Πολλαπλή Λογιστική Παλινδρόμηση:** Όταν έχουμε παραπάνω από μια ανεξάρτητες μεταβλητές x_i, επεκτείνουμε το απλό λογιστικό μοντέλο ως εξής:

$$y_i = E[y_i] + \varepsilon_i$$

όπου

$$E[y_i] = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n}} = \frac{e^{\beta' x_i}}{1 + e^{\beta' x_i}}$$

με

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{pmatrix}, \quad x = \begin{pmatrix} 1 \\ x_1 \\ \vdots \\ x_n \end{pmatrix} \quad \text{και} \quad x_i = \begin{pmatrix} 1 \\ x_{i1} \\ \vdots \\ x_{in} \end{pmatrix}$$

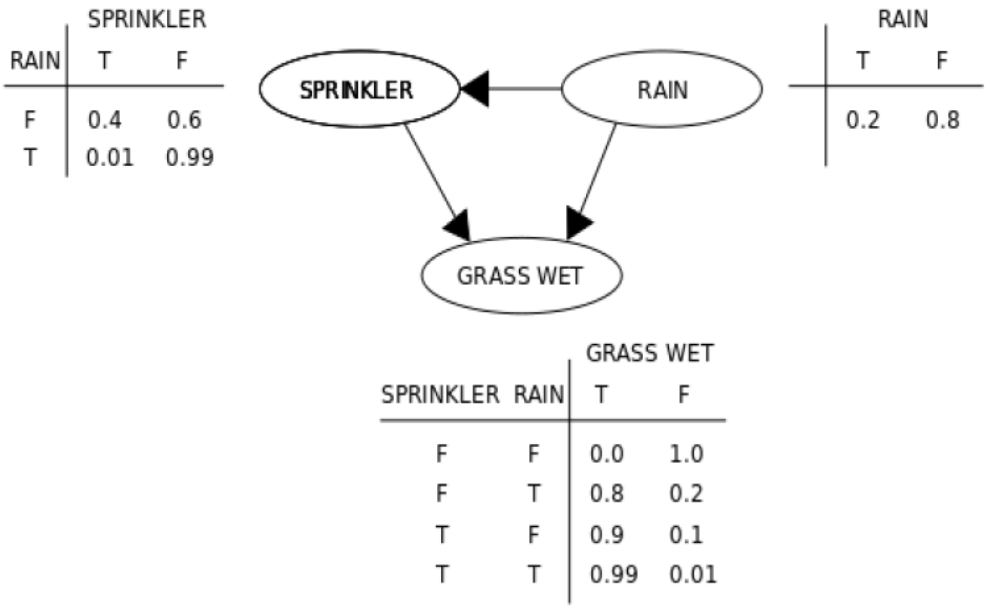
Για την εκτίμηση των παραμέτρων του μοντέλου, δηλαδή του πίνακα β, κάνουμε χρήση της μεθόδου της μέγιστης πιθανοφάνειας (*maximum likelihood method*), όπως στο απλό λογιστικό μοντέλο.

2.2.4 Μπεϋζιανά μοντέλα δικτύου (Bayesian network models)

Ένα Μπεϋζιανό δίκτυο (*Bayesian network*) είναι ένα κατευθυνόμενο άκυκλο γράφημα (DAG), μέσω του οποίου περιγράφεται η κοινή κατανομή πιθανότητας ενός δεδομένου συνόλου τυχαίων μεταβλητών. Οι κόμβοι του άκυκλου γραφήματος αποτελούν τις τυχαίες μεταβλητές στην Μπεϋζιανή λογική (παρατηρήσιμες ποσότητες, λανθάνουσες μεταβλητές, άγνωστες παράμετροι ή υποθέσεις), ενώ οι ακμές εκφράζουν την αλληλεξάρτηση των τυχαίων μεταβλητών. Κάθε κόμβος θεωρείται ανεξάρτητος από όλους αυτούς που δεν

είναι απόγονοι του και αντιπροσωπεύεται από μια συνάρτηση πιθανότητας, η οποία υπολογίζει την πιθανότητα της τυχαίας μεταβλητής που αποτελεί ο αντίστοιχος κόμβος. Αυτό το επιτυγχάνει, λαμβάνοντας ως είσοδο ένα συγκεκριμένο σύνολο τιμών για τις μεταβλητές, οι οποίες αποτελούν τους κόμβους γονείς του κόμβου που η συνάρτηση αντιπροσωπεύει. Για παράδειγμα, αν οι γονείς ενός κόμβου k αποτελούν m τυχαίες μεταβλητές, η συνάρτηση πιθανότητας που αντιπροσωπεύει τον κόμβο k θα μπορούσε να είναι ένας πίνακας με 2^m στοιχεία, δηλαδή ένα στοιχείο για κάθε έναν από τους 2^m δυνατούς συνδυασμούς των γονέων του.

Ένα μοντέλο Μπεϋζιανού δικτύου (*Bayesian network model*) πρόκειται για έναν κατηγοριοποιητή που κάνει αποτίμηση πιθανοτήτων βασιζόμενος στη θεωρία του Bayes ($\mathrm{P(B|A)} = \frac{\mathrm{P(A|B)} \cdot \mathrm{P(B)}}{\mathrm{P(A)}}$) και όχι στις προβλέψεις. Αποτελείται από ένα Μπεϋζιανό δίκτυο, δηλαδή από ένα κατευθυνόμενο άκυκλο γράφημα, έστω $\mathrm{G} = (\mathrm{V}, \mathrm{E})$ με V το σύνολο των κόμβων και E το σύνολο των ακμών, μαζί με έναν πίνακα υπό συνθήκη πιθανοτήτων (*conditional probability table*) για κάθε ένα κόμβο δεδομένου των κόμβων γονέων του (βλέπε Σχήμα 17). Με τον όρο υπό συνθήκη πιθανότητα εννοούμε την τιμή $\mathrm{P(x|p_1, p_2, p_3, \dots, p_n)}$ που εκφράζει την πιθανότητα του να βρίσκεται μια μεταβλητή X στην κατάσταση x, δεδομένου ότι οι γονείς τις $\mathrm{P_1, P_2, P_3, \dots, P_n}$ βρίσκονται στην κατάσταση $\mathrm{p_1, p_2, p_3, \dots, p_n}$ αντίστοιχα. Η κοινή κατανομή πιθανότητας των τυχαίων μεταβλητών V υπολογίζεται ως το γινόμενο των υπό συνθήκη πιθανοτήτων για όλους τους κόμβους δεδομένων των γονέων τους.



Σχήμα 17: Παράδειγμα Μπεϋζιανού μοντέλου δικτύου (βλ. ιστότοπο Wikipedia).

Σύνθεση

Για την δημιουργία του κατάλληλου μοντέλου Μπεϋζιανού δικτύου δεδομένου ενός συνόλου τυχαίων μεταβλητών $\mathbb{V}$ και των αντίστοιχων δεδομένων δείγματος, απαραίτητες είναι δύο διαδικασίες:

1. Η δομή μάθησης: καθορισμός των κατάλληλων ακμών στο γράφημα G
2. Η παράμετρος μάθησης: εκτίμηση του πίνακα των υπό συνθήκη πιθανοτήτων για κάθε κόμβο δεδομένων των κόμβων γονέων του.

2.2.5 Μηχανές Διανυσμάτων Υποστήριξης (SVM)

Οι μηχανές διανυσμάτων υποστήριξης είναι μια από τις πιο παλιές τεχνικές στον τομέα της στατιστικής και της μηχανικής μάθησης. Η ιδέα των SVM προτάθηκε αρχικά το 1963 από τον Vapnik, ο οποίος σε συνεργασία με την Cortes το 1995 σχεδίασε την σημερινή τελική μορφή της μεθόδου (Cortes & Vapnik, 1995). Ο αλγόριθμος των SVM αποτελεί ένα μοντέλο, που μπορεί να χρησιμοποιηθεί για την ταξινόμηση, την παλινδρόμηση ή άλλες εργασίες. Κύριος σκοπός του είναι η εύρεση ενός βέλτιστου διαχωριστικού υπερεπιπέδου (*separating hyperplane*) ανάμεσα στα δεδομένα, μεγιστοποιώντας την απόσταση αυτού από τις κοντινότερες παρατηρήσεις του συνόλου εκπαίδευσης, η εύρεση δηλαδή του λεγόμενου υπερεπιπέδου μεγίστου περιθωρίου (*max margin hyperplane*). Στις δύο διαστάσεις το υπερεπίπεδο καλείται ευθεία, στις τρεις επίπεδο, ενώ σε παραπάνω διαστάσεις υπερεπίπεδο. Έτσι, μπορούμε να πούμε ότι ένα μοντέλο SVM είναι μια αναπαράσταση του training set ως σημεία στο χώρο, τα οποία χαρτογραφούνται ώστε να χωρίζονται από ένα όσο το δυνατόν σαφέστερο και ευρύτερο κενό. Επομένως, τα νέα δεδομένα προβλέπεται σε ποια κατηγορία ανήκουν ανάλογα με την πλευρά του περιθωρίου (*margin*) που θα πέσουν. Παρατηρούμε λοιπόν ότι η ανάπτυξη της SVM μεθόδου είναι εντελώς διαφορετική από τους συνήθεις αλγορίθμους μηχανικής μάθησης.

Κάποιες βασικές έννοιες της SVM μεθόδου είναι το διαχωριστικό υπερεπίπεδο (*separating hyperplane*), το περιθώριο (*margin*), το «μαλακό» περιθώριο (*soft margin*) και το βέλτιστο υπερεπίπεδο (*max margin hyperplane*). Αναλυτική περιγραφή αυτών, καθώς και της λειτουργίας των μηχανών διανυσμάτων υποστήριξης θα γίνει στο κεφάλαιο 3.

Εφαρμογή

Οι SVM έχουν εφαρμογές σε πολλά πεδία. Για παράδειγμα χρησιμεύουν στην:

i. Κατηγοριοποίηση κειμένου (πχ. email filtering) (Tong & Koller, 2000 και Joachims, 1999).
ii. Ταξινόμηση των εικόνων (*relevance feedback*) (Tong & Chang, 2001).
iii. Ιατρική επιστήμη.
iv. Αναγνώριση χειρόγραφων χαρακτήρων.

Πλεονεκτήματα

Τα βασικά πλεονεκτήματα των SVM είναι:

- Βασίζονται σε πολύ απλές και ξεκάθαρες ιδέες από τη θεωρία στατιστικής μάθησης (Cortes & Vapnik, 1995) και μπορούν να χρησιμοποιηθούν για την πρόβλεψη μελλοντικών δεδομένων.
- Έχουν ισχυρό θεωρητικό υπόβαθρο και έτσι μπορεί να ακολουθηθεί η διαδικασία βήμα προς βήμα.
- Η εκπαίδευση τους είναι σχετικά εύκολη. Κλιμακώνεται σε σχετικά καλές υψηλές διαστάσεις των δεδομένων και η εξισορρόπηση μεταξύ της ταξινόμησης, της πολυπλοκότητας και του λάθους μπορεί να ελεγχθεί καλά. Το μόνο που απαιτείται είναι η καλή λειτουργία του πυρήνα (βλ. παράγραφο 3.1.3).
- Με την εφαρμογή του πυρήνα (βλ. παράγραφο 3.1.3) αποκτούν ευελιξία όσον αφορά την επιλογή της μορφής του διαχωριστικού υπερεπιπέδου που διαχωρίζει τις κλάσεις, άσχετα από το αν αυτές είναι ή όχι γραμμικά διαχωριζόμενες και αν έχουν την ίδια συνάρτηση για όλα τα δεδομένα.
- Δεν απαιτούν γνώση της στατιστικής κατανομής των δεδομένων.
- Εκπαιδεύονται από την επίλυση ενός περιορισμένου τετραγωνικού προβλήματος βελτιστοποίησης και έτσι προσφέρουν μια μοναδική, βέλτιστη και ολική λύση για κάθε επιλογή των SVM παραμέτρων w και b (βλέπε κεφάλαιο 3).
- Μπορούν να χρησιμοποιηθούν για μια ποικιλία από αναπαραστάσεις, όπως τα νευρωνικά δίκτυα, splines, πολυωνυμικούς εκτιμητές κ.λ.π. Σε πληθώρα πραγματικών εφαρμογών έχουν επιδείξει ισάξια ή και καλύτερη επίδοση συγκριτικά με άλλες ανταγωνιστικές μεθόδους.
- Παρέχουν αρκετά καλή γενίκευση και εκτός δείγματος.
- Ξεπερνούν σε σημαντικό βαθμό το πρόβλημα υπερπροσαρμογής στα δεδομένα (*overfitting*).
- Το πλήθος των παραμέτρων που απαιτούν ρύθμιση είναι σημαντικά μικρότερο από το αντίστοιχο άλλων μεθοδολογιών.
- Μπορούν να παράγουν περίπλοκα μη γραμμικά μοντέλα, που έχουν συγκεκριμένη συναρτησιακή διατύπωση.

ΚΕΦΑΛΑΙΟ 3 : ΜΗΧΑΝΕΣ ΔΙΑΝΥΣΜΑΤΩΝ ΥΠΟΣΤΗΡΙΞΗΣ

3.1 Εισαγωγή στις Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines - SVM)

Οι SVM λειτουργούν ως μια από τις καλύτερες προσεγγίσεις για την μοντελοποίηση των δεδομένων. Αποτελούν μοντέλα μάθησης με επίβλεψη[1] και χρησιμοποιούνται τόσο στην ταξινόμηση δεδομένων όσο και στην πρόβλεψη άγνωστων παραμέτρων.

3.1.1 Δυαδική Ταξινόμηση (binary classification) με χρήση SVM

Τα τελευταία δέκα χρόνια οι SVM αποτελούν σημείο αναφοράς για πολλά προβλήματα ταξινόμησης, λόγω της ευελιξίας τους, της υπολογιστικής αποδοτικότητας τους και της ικανότητας διαχείρισης δεδομένων υψηλής διάστασης. Προφανώς, στόχος των μηχανών διανυσμάτων υποστήριξης ως μέθοδος ταξινόμησης είναι, αφού εκπαιδευτούν με ένα training set, να δημιουργήσουν ένα μοντέλο για την πρόβλεψη της κλάσης νέων set δεδομένων. Στην περίπτωση της δυαδικής ταξινόμησης, ο διαχωρισμός των αντικειμένων γίνεται σε δύο κατηγορίες, όπου η μια συμβολίζεται με -1, ενώ η άλλη με +1. Η περίπτωση όπου έχουμε περισσότερες των δύο κατηγορίες (*multiclass SVM*) είναι εκτός των σκοπών της παρούσας εργασίας και δεν θα εξεταστεί.

Για την δυαδική ταξινόμηση με χρήση SVM διακρίνουμε δύο μορφές δεδομένων: τα γραμμικά διαχωριζόμενα δεδομένα και τα μη γραμμικά διαχωριζόμενα δεδομένα. Βασιζόμενοι στο βιβλίο του Hastie et al. (2001) ορίζουμε τα παρακάτω:

[1] **Μάθηση με επίβλεψη (*supervised learning*)**: ο αλγόριθμος εκπαιδεύεται μέσω ενός training set, τα στοιχεία του οποίου συνοδεύονται από ετικέτες που δείχνουν την κλάση τους. Τα νέα δεδομένα κατηγοριοποιούνται με βάση το σύνολο εκπαίδευσης.

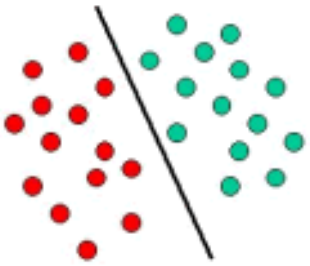
- **Γραμμικά διαχωριζόμενα δεδομένα (*linearly separable data*)**

Έστω ότι το σύνολο εκπαίδευσης περιέχει n εγγραφές $t_1, t_2, \dots, t_n$. Θεωρούμε ότι τα t_i έχουν τη μορφή ζεύγους $t_i = (x_i, y_i)$, όπου $x_i \in \mathbb{R}^p$ η τιμή που παίρνει η t_i εγγραφή σε κάθε χαρακτηριστικό, p ο αριθμός των χαρακτηριστικών του προβλήματος ταξινόμησης, y_i η κατηγορία της t_i (δηλαδή $y_i = +1$ ή $y_i = -1$) και $i = 1, 2, \dots, n$. Δηλαδή το σύνολο εκπαίδευσης είναι της μορφής:

$$T = \{t_1, t_2, \dots, t_n\} = \left\{(x_i, y_i) \middle| x_i \in \mathbb{R}^p, y_i \in \{-1,1\}\right\}_{i=1}^{n}$$

Ορισμός

Λέμε ότι τα δεδομένα είναι γραμμικά διαχωριζόμενα όταν μπορούμε να δημιουργήσουμε ένα υπερεπίπεδο στο γράφημα των $t_1, t_2, \dots, t_n$ που να χωρίζει τις δύο κλάσεις $y_i = +1$ και $y_i = -1$. Παράδειγμα γραμμικά διαχωριζόμενων δεδομένων απεικονίζεται στο σχήμα που ακολουθεί (Σχήμα 18).



Σχήμα 18: Γραφική απεικόνιση γραμμικά διαχωριζόμενων δεδομένων. Παρατηρούμε ότι η γραμμή μεταξύ κόκκινων (έστω κατηγορία $\boldsymbol{y_i = -1}$) και πράσινων (έστω κατηγορία $\boldsymbol{y_i = +1}$) κουκίδων διαχωρίζει ξεκάθαρα τα δεδομένα των δύο κλάσεων (βλ. ιστότοπο statsoft).

Στόχος της SVM μεθόδου, όπως αναφέραμε και στο κεφάλαιο 2, είναι να βρει το βέλτιστο διαχωριστικό υπερεπίπεδο (*separating hyperplane*), δηλαδή το υπερεπίπεδο μεγίστου περιθωρίου (*max margin hyperplane*). Κάθε separating hyperplane μπορεί να γραφτεί ως το σύνολο των σημείων x που ικανοποιούν την εξίσωση:

$$w \cdot x + b = 0$$

όπου w είναι το κάθετο διάνυσμα σε όλο το υπερεπίπεδο, γνωστό και ως διάνυσμα βαρών, και $b/\|w\|$ η κάθετη απόσταση του υπερεπιπέδου από την αρχή των αξόνων, με την τιμή b να αποτελεί μια πόλωση (*bias*).

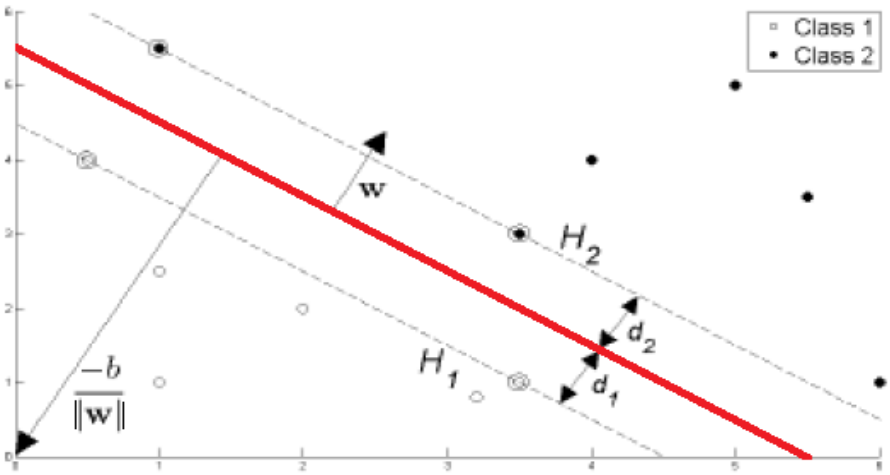
Έτσι, η υλοποίηση SVM στηρίζεται ουσιαστικά στην επιλογή των μεταβλητών w και b, ώστε εν τέλει να προσδιορίσει το υπερεπίπεδο, το οποίο θα διαχωρίζει τα σημεία - εγγραφές που έχουν $y_i = +1$ από αυτά που έχουν $y_i = -1$ και θα είναι προσανατολισμένο με τέτοιο τρόπο ώστε να βρίσκεται όσο το δυνατόν μακρύτερα από τα πλησιέστερα μέλη των δύο κλάσεων, τα αποκαλούμενα και διανύσματα υποστήριξης (*support vectors*). Τα support vectors βρίσκονται πάνω στα επίπεδα H_1 και H_2, τα οποία περιγράφονται αντιστοίχως από τις σχέσεις:

$$w \cdot x_i + b = +1$$

$$w \cdot x_i + b = -1$$

όπου x_i είναι το διάνυσμα υποστήριξης.

Ορίζοντας ως d_1 την απόσταση του H_1 από το διαχωριστικό υπερεπίπεδο και d_2 την απόσταση του H_2 από το διαχωριστικό υπερεπίπεδο, τότε για $d_1 = d_2 = d$ λέμε ότι τα H_1 και H_2 απέχουν ίση απόσταση από αυτό και η τιμή d ονομάζεται περιθώριο SVM (*margin*) (ουσιαστικά το περιθώριο είναι η απόσταση του υπερεπιπέδου από ένα διάνυσμα υποστήριξης x_i).



Σχήμα 19: Γραφική απεικόνιση του max margin hyperplane (κόκκινη γραμμή) στην περίπτωση γραμμικά διαχωριζόμενων δεδομένων. Στη γραφική αυτή παράσταση μπορούμε ξεκάθαρα να παρατηρήσουμε και τα

επίπεδα H_1 και H_2, τις αποστάσεις $\mathbf{d_1}$, $\mathbf{d_2}$, το κάθετο διάνυσμα $\mathbf{w}$ στο υπερεπίπεδο καθώς και την απόσταση $\mathbf{b}/\|\mathbf{w}\|$ αυτού από την αρχή των αξόνων.

Άρα, για να βρει η SVM μέθοδος τον κατάλληλο προσανατολισμό για το διαχωριστικό υπερεπίπεδο αρκεί να μεγιστοποιήσει το περιθώριο, το οποίο κάνοντας χρήση της γεωμετρίας προκύπτει ότι ισούται με $1/\|w\|$. Η μεγιστοποίηση της τιμής $1/\|w\|$ είναι προφανώς ισοδύναμη με την εύρεση του $\min\|w\|$. Για να αποφύγουμε κάποιο σημείο να πέσει στο περιθώριο οφείλουμε να θέσουμε τους εξής περιορισμούς:

$$\left.\begin{array}{l} w \cdot x_i + b \geq +1 \quad \text{για} \quad \underline{y_i = +1} \\ \\ w \cdot x_i + b \leq -1 \quad \text{για} \quad \underline{y_i = -1} \end{array}\right\} \Leftrightarrow y_i\,(w \cdot x_i + b) - 1 \geq 0 \quad \forall\, i = 1, 2, \dots, n$$

Η εύρεση του $\min\|w\|$ είναι ισοδύναμη με την εύρεση του $\min\frac{1}{2}\|w\|^2$, γεγονός που κάνει εφικτή την εκτέλεση της βελτιστοποίησης του τετραγωνικού προγραμματισμού (*Quadratic programming optimization*)[2]. Προκύπτει έτσι το αρχικό πρόβλημα βελτιστοποίησης[3]:

$$\min\frac{1}{2}\|w\|^2 \quad \text{υπό τον περιορισμό} \quad y_i\,(w \cdot x_i + b) - 1 \geq 0, \quad \text{όπου } i = 1, 2, \dots, n$$

Η $\frac{1}{2}\|w\|^2$ ονομάζεται αντικειμενική συνάρτηση του προβλήματος βελτιστοποίησης και εύκολα παρατηρεί κανείς ότι είναι κυρτή[4] (προφανώς, λόγω της τριγωνικής ανισότητας που ισχύει για τις νόρμες, $\forall\, w_1, w_2$ ισχύει ότι: $\frac{1}{2}\|\lambda w_1 + (1-\lambda)w_2\| \leq \frac{1}{2}\lambda\|w_1\| + \frac{1}{2}(1-\lambda)\|w_2\|$, $\forall\, 0 \leq \lambda \leq 1$).

---

[2]**Τετραγωνικός προγραμματισμός:** αλγόριθμος που επωφελείται από την κυρτότητα μιας συνάρτησης για να λύσει κυρτά προβλήματα βελτιστοποίησης.

[3]**Πρόβλημα βελτιστοποίησης:** είναι ένα πρόβλημα, στο οποίο θέλουμε να επιλέξουμε την καλύτερη λύση από έναν αριθμό εφικτών λύσεων. Οι εφικτές λύσεις ταξινομούνται από μια αντικειμενική συνάρτηση και στόχος είναι η εύρεση της λύσης που ελαχιστοποιεί την συνάρτηση αυτή. Επίσης, κάθε πρόβλημα βελτιστοποίησης διαθέτει ένα σύνολο περιορισμών, οι οποίοι θέτουν τα όρια ως προς το αν μια λύση είναι εφικτή ή όχι.

[4]**Κυρτή συνάρτηση:** μια συνάρτηση f(x) είναι κυρτή αν $\forall x_1, x_2$ ισχύει $f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda)f(x_2)$, $\forall\, 0 \leq \lambda \leq 1$.

Για να παράγουμε τις μηχανές διανυσμάτων υποστήριξης κάνουμε χρήση των πολλαπλασιαστών Lagrange $\vec{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_n)$, με $\alpha_i \geq 0 \; \forall i$, όπου κάθε α_i αντιστοιχεί σε έναν από τους περιορισμούς $y_i \; (w \cdot x_i + b) - 1 \geq 0$. Με αυτό τον τρόπο έχουμε μια νέα αντικειμενική συνάρτηση L_p, η οποία λέγεται Λαγκρανζιανή:

$$L_p = \frac{1}{2}\|w\|^2 - \alpha\,[\,y_i(w \cdot x_i + b) - 1\,] = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i [\,y_i\,(w \cdot x_i + b) - 1\,]$$

$$\Leftrightarrow \; L_p = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i\, y_i\,(w \cdot x_i + b) + \sum_{i=1}^{n} \alpha_i$$

Οι λύσεις στο νέο πρόβλημα βελτιστοποίησης Lagrange είναι τα σημεία που μεγιστοποιούν την L_p ως προς $\vec{\alpha}$ και ταυτόχρονα ελαχιστοποιούν τη συνάρτηση ως προς τις μεταβλητές w και b. Για την εύρεση τους αρκεί, αρχικά, να μηδενίσουμε τις μερικές παραγώγους της L_p ως προς w και b:

$$\frac{\partial L_p}{\partial w} = 0 \Rightarrow \mathbf{w = \sum_{i=1}^{n} \alpha_i y_i x_i} \qquad [1]$$

$$\frac{\partial L_p}{\partial b} = 0 \Rightarrow \mathbf{\sum_{i=1}^{n} \alpha_i y_i = 0} \qquad [2]$$

Αντικαθιστώντας στην L_p τις σχέσεις [1] και [2] έχουμε μια διαφορετική μορφή για αυτή, την L_D, η οποία ονομάζεται διπλή μορφή της πρωτοβάθμιας L_p :

$$L_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i x_j$$

υπό τις συνθήκες $\alpha_i \geq 0 \; \forall i$ και $\sum_{i=1}^{n} \alpha_i y_i = 0$.

$$L_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i H_{ij} \alpha_j$$

όπου $H_{ij} = y_i y_j x_i x_j$

$$L_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \alpha^T H \alpha$$

υπό τις συνθήκες $\alpha_i \geq 0 \; \forall i$ και $\sum_{i=1}^{n} \alpha_i y_i = 0$

Τέλος, πρέπει όπως αναφέραμε να βρούμε το max L_D ως προς $\vec{\alpha}$ υπό τις συνθήκες $\alpha_i \geq 0 \; \forall i$ και $\sum_{i=1}^{n} \alpha_i y_i = 0$. Δηλαδή, έχουμε να επιλύσουμε ένα νέο, μάλιστα κυρτό (η αντικειμενική συνάρτηση είναι κυρτή, ενώ οι περιορισμοί γραμμικοί) και τετραγωνικό πρόβλημα βελτιστοποίησης:

$$\max L_D \quad \text{υπό τους περιορισμούς} \quad \alpha_i \geq 0 \; \forall i \text{ και } \sum_{i=1}^{n} \alpha_i y_i = 0$$

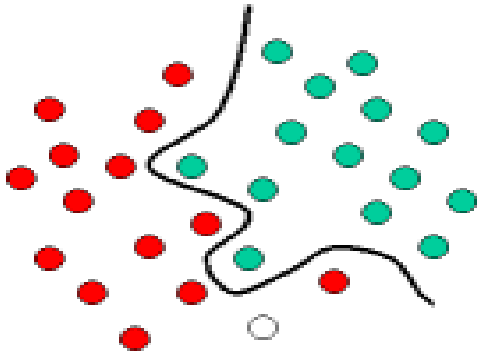
Διατρέχοντας μια QP επίλυση βρίσκουμε το $\vec{\alpha}$ και άρα μέσω της σχέσης [1] το w. Όσον αφορά την τιμή b, αντικαθιστώντας στην [1] την σχέση [2] και χρησιμοποιώντας τον αρχικό περιορισμό $y_i (w \cdot x_i + b) - 1 \geq 0$, υπολογίζουμε τον μέσο όρο όλων των x_i και έτσι υπολογίζουμε το b.

Έχοντας υπολογίσει τώρα τα w και b, μπορούμε να ορίσουμε το υπερεπίπεδο μεγίστου περιθωρίου και ως εκ τούτου την SVM.

- **Μη γραμμικά διαχωριζόμενα δεδομένα (*nonlinearly separable data*)**

Ορισμός

Προφανώς, λέμε ότι τα δεδομένα είναι μη γραμμικά διαχωριζόμενα όταν δεν μπορούμε να δημιουργήσουμε ένα υπερεπίπεδο στο γράφημα των $t_1, t_2, \ldots, t_n$ που να χωρίζει τις δύο κλάσεις $y_i = +1$ και $y_i = -1$. Δηλαδή στις διαχωριζόμενες περιοχές υπάρχουν εγγραφές και των δύο κλάσεων ταυτοχρόνως. Παράδειγμα μη γραμμικά διαχωριζόμενων δεδομένων απεικονίζεται στο σχήμα που ακολουθεί (Σχήμα 20).



Σχήμα 20: Γραφική απεικόνιση μη γραμμικά διαχωριζόμενων δεδομένων. Παρατηρούμε ότι οι δύο κλάσεις εμφανίζονται ανακατεμένες με αποτέλεσμα να μην μπορούμε να διαχωρίσουμε με μια ευθεία γραμμή τα δεδομένα (βλ. ιστότοπο statsoft).

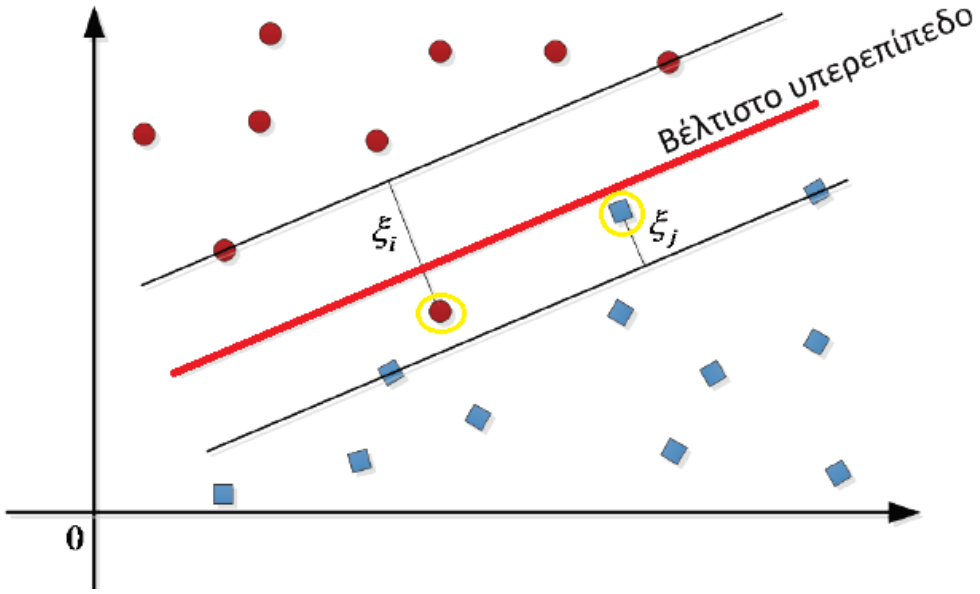
Προκειμένου να επεκταθεί η χρήση των SVM και σε μη γραμμικά διαχωριζόμενα δεδομένα πραγματοποιούμε δύο ενέργειες:

1. Προβάλλουμε τα x_i του συνόλου εκπαίδευσης μέσω μιας απεικόνισης Φ σε έναν χώρο υψηλότερης διάστασης, στον οποίο είναι πιο πιθανό τα δεδομένα να είναι γραμμικά διαχωριζόμενα. Έτσι, κάθε x_i χαρτογραφείται ως ένα διάνυσμα $\Phi(x_i) = [\Phi(x_{i1}), \Phi(x_{i2}), \dots]$.
2. Χαλαρώνουμε λίγο τους περιορισμούς που θέσαμε για τα γραμμικά διαχωριζόμενα δεδομένα. Για τον σκοπό αυτό εισάγουμε μια ποινή, μια χαλαρή μεταβλητή $\xi_i \geq 0,\ i = 1, 2, \dots, n$, όπου το άθροισμα $\sum_i^n \xi_i$ αποτελεί το πλήθος των λανθασμένα ταξινομημένων εγγραφών.

Έτσι, έχουμε:

$$\left.\begin{array}{ll} w \cdot \Phi(x_i) + b \geq +1 - \xi_i & \text{για } \underline{y_i = +1} \\ w \cdot \Phi(x_i) + b \leq -1 + \xi_i & \text{για } \underline{y_i = -1} \end{array}\right\} \Leftrightarrow y_i\,(w \cdot \Phi(x_i) + b) - 1 + \xi_i \geq 0 \quad \forall\, i = 1, 2, \dots, n$$

Στην περίπτωση αυτή, η έννοια του περιθωρίου αντικαθιστάται από αυτήν του «μαλακού» περιθωρίου (*soft margin*). Με αυτόν τον τρόπο επιτρέπεται στα δεδομένα εκπαίδευσης να παραβιάζουν το διαχωριστικό υπερεπίπεδο, όπως φαίνεται στο παρακάτω σχήμα (Σχήμα 21).



Σχήμα 21: Γραφική απεικόνιση του max margin hyperplane (κόκκινη γραμμή) στην περίπτωση μη γραμμικά διαχωριζόμενων δεδομένων. Παρατηρούμε ότι οι κυκλωμένες με κίτρινο κουκίδες είναι λάθος ταξινομημένες με βάση το διαχωριστικό επίπεδο που έχει οριστεί στο σχήμα.

Όμοια με την περίπτωση των γραμμικά διαχωριζόμενων δεδομένων, στόχος μας είναι η μεγιστοποίηση του περιθωρίου, εδώ του «μαλακού» περιθωρίου, δηλαδή η εύρεση του $\min\|w\|$. Καθώς το άθροισμα των χαλαρών μεταβλητών $\sum_i^n \xi_i$ αποτελεί, όπως προαναφέραμε, το πλήθος των λανθασμένα ταξινομημένων εγγραφών, οφείλουμε να προσθέσουμε το ανάλογο του (έστω $C \cdot \sum_i^n \xi_i$, όπου C θετική μη μηδενική σταθερά) στην αντικειμενική συνάρτηση. Η τιμή της σταθεράς C δίνεται από τον χρήστη και ουσιαστικά αποτελεί το βάρος του κόστους των λανθασμένων ταξινομήσεων. Γενικά, ισχύει ότι όσο πιο μεγάλο είναι το C, τόσο μικρότερο είναι το περιθώριο, άρα και η πιθανότητα λανθασμένης ταξινόμησης. Αντίστροφα, όσο πιο μικρό είναι το C, τόσο μεγαλύτερο είναι το περιθώριο και άρα αυξάνεται η πιθανότητα λανθασμένης ταξινόμησης. Έχουμε λοιπόν το εξής αρχικό πρόβλημα βελτιστοποίησης:

$$\min \frac{1}{2}\|w\|^2 + C \cdot \sum_i^n \xi_i \text{ υπό τον περιορισμό } y_i\,(w \cdot \Phi(x_i) + b) - 1 + \xi_i \geq 0$$

όπου $i = 1, 2, \dots, n$

Χρησιμοποιώντας και πάλι τους πολλαπλασιαστές Lagrange $\vec{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_n)$ για κάθε περιορισμό $y_i\,(w \cdot \Phi(x_i) + b) - 1 + \xi_i \geq 0$ και $\vec{\mu} = (\mu_1, \mu_2, \dots, \mu_n)$ για κάθε χαλαρή μεταβλητή ξ_i, με $\alpha_i \geq 0$ και $\mu_i \geq 0$ $\forall i$, έχουμε:

$$L_p = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i[\, y_i\,(w\cdot\Phi(x_i) + b) - 1 + \xi_i\,] - \sum_{i=1}^{n}\mu_i\xi_i$$

Παραγωγίζοντας ως προς w, b και ξ_i την L_p προκύπτει:

$$\frac{\partial L_p}{\partial w} = 0 \Rightarrow \mathbf{w} = \sum_{\mathbf{i=1}}^{\mathbf{n}} \boldsymbol{\alpha_i y_i \Phi(xi)} \qquad [3]$$

$$\frac{\partial L_p}{\partial b} = 0 \Rightarrow \sum_{\mathbf{i=1}}^{\mathbf{n}} \boldsymbol{\alpha_i y_i} = \mathbf{0} \qquad [4]$$

$$\frac{\partial L_p}{\partial \xi_i} = 0 \Rightarrow \mathbf{C} = \boldsymbol{\alpha_i} + \boldsymbol{\mu_i} \qquad [5]$$

Αντικαθιστώντας στην L_p τις σχέσεις [3], [4], [5] καταλήγουμε πάλι στην έκφραση:

$$L_D = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\,\alpha^T H\,\alpha$$

Επειδή $\mu_i \geq 0 \quad \forall i$ και $C = \alpha_i + \mu_i$ προκύπτει ότι $\alpha_i \leq C \ \forall i$.

Στόχος μας, όπως και στα γραμμικά διαχωριζόμενα δεδομένα, είναι η επίλυση του προβλήματος βελτιστοποίησης:

$$\max L_D \quad \text{υπό τους περιορισμούς} \quad 0 \leq \alpha_i \leq C \ \forall i \quad \text{και} \quad \sum_{i=1}^{n}\alpha_i y_i = 0$$

Με την επίλυση αυτού προκύπτουν εν τέλει οι μεταβλητές w και b και έτσι ορίζουμε το βέλτιστο διαχωριστικό υπερεπίπεδο και κατ' επέκταση την SVM.

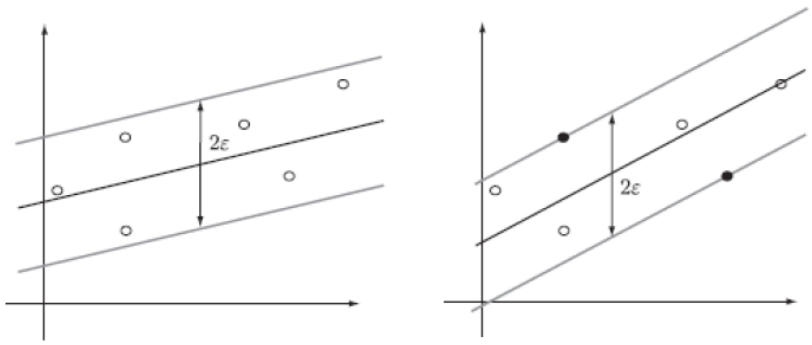
3.1.2 Παλινδρόμηση με χρήση SVM (Support Vector Regression - SVR)

Οι μηχανές διανυσμάτων υποστήριξης μπορούν να χρησιμοποιηθούν και σε προβλήματα παλινδρόμησης. Η παλινδρόμηση (*regression*) είναι μια στατιστική τεχνική που σχετίζεται με την μοντελοποίηση μιας εξαρτημένης μεταβλητής και μιας ή πολλών ανεξάρτητων μεταβλητών. Οπότε, αντί να προσπαθούμε να κατατάξουμε νέες άγνωστες μεταβλητές σε μια κατηγορία, όπως κάναμε στην ταξινόμηση, επιθυμούμε να εκτιμήσουμε μια πραγματική τιμή εξόδου $\widehat{y_i}$. Θεωρούμε ότι το σύνολο εκπαίδευσης της παλινδρόμησης είναι λοιπόν της μορφής:

$$T = \{t_1, t_2, \dots, t_n\} = \left\{(x_i, \widehat{y_i}) \middle| x_i \in \mathbb{R}^p, \widehat{y_i} \in \mathbb{R}\right\}_{i=1}^{n}$$

όπου $\widehat{y_i} = wx_i + b$.

Σύμφωνα με τον Hamel (2009), στην SVM παλινδρόμηση τα δεδομένα προσαρμόζονται σε ένα υπερσωλήνα πλάτους 2ε (ε – *insensitive tube*), όπου $\varepsilon > 0$. Αυτός ουσιαστικά αποτελεί ένα μοντέλο παλινδρόμησης και ακριβώς στο κέντρο του είναι τοποθετημένο ένα υπερεπίπεδο που μοντελοποιεί τις παρατηρήσεις (βλέπε Σχήμα 22). Υπάρχουν πολλοί τρόποι προσανατολισμού του υπερ-σωλήνα αυτού, ώστε τα δεδομένα να βρίσκονται εντός του. Ωστόσο, υπάρχει ένας βέλτιστος προσανατολισμός, όπως υπήρχε και για το max margin hyperplane στην ταξινόμηση, κατά τον οποίο οι παρατηρήσεις ωθούνται πιο κοντά στα εξωτερικά «τοιχώματα» του υπερσωλήνα. Με άλλα λόγια, ο βέλτιστος αυτός προσανατολισμός επιτυγχάνεται όταν οι αποστάσεις των παρατηρήσεων από το υπερεπίπεδο που βρίσκεται στο κέντρο του σωλήνα μεγιστοποιούνται.



Σχήμα 22: Επίλυση προβλήματος παλινδρόμησης με χρήση SVM. Αριστερά παρατηρούμε έναν υπερσωλήνα μήκους **2ε**, όπου όλες οι παρατηρήσεις βρίσκονται εντός του και δεξιά απεικονίζεται ο βέλτιστα προσανατολισμένος υπερσωλήνας μεγίστου περιθωρίου.

Άρα, παρατηρούμε ότι στόχος μας για ακόμη μια φορά είναι η μεγιστοποίηση ενός περιθωρίου. Έτσι, όμοια με το πρόβλημα βελτιστοποίησης που ορίσαμε στην παράγραφο 3.1.1 στην περίπτωση των γραμμικά διαχωριζόμενων δεδομένων, επιθυμούμαι να βρούμε το $\min\frac{1}{2}\|w\|^2$, αυτή τη φορά όμως υπό άλλους περιορισμούς. Συγκεκριμένα, θέλουμε να εξασφαλίσουμε ότι οι παρατηρήσεις θα είναι όλες εντός του υπερσωλήνα. Οπότε θέτουμε τις εξής συνθήκες:

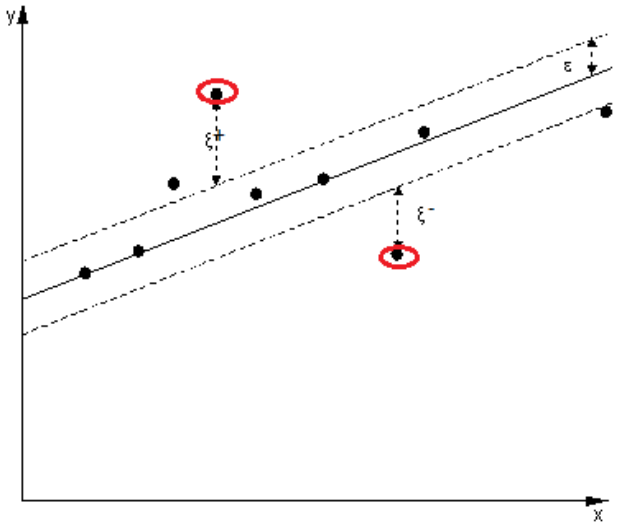
$$\left.\begin{matrix} y_i - \hat{y_i} \leq \varepsilon \\ \hat{y_i} - y_i \leq \varepsilon \end{matrix}\right\} \Leftrightarrow |y_i - \hat{y_i}| \leq \varepsilon \ \forall\, i = 1, 2, \dots, n$$

όπου y_i η πραγματική τιμή και $\hat{y_i} = wx_i + b$ η εκτιμώμενη από το υπερεπίπεδο τιμή.

Στην περίπτωση αυτή έχουμε κάνει την υπόθεση ότι είναι δυνατόν όλες οι παρατηρήσεις να χωρέσουν σε έναν υπερσωλήνα πλάτους 2ε. Όμως στην πραγματικότητα αυτό είναι δύσκολο να συμβεί. Τι γίνεται λοιπόν σε αυτή την περίπτωση; Όπως όταν είχαμε μη γραμμικά διαχωριζόμενα δεδομένα, εισάγουμε για κάθε μια από τις παρατηρήσεις $(x_i, \hat{y_i})$ που είναι εκτός του υπερσωλήνα, δηλαδή που δεν ικανοποιούν τον περιορισμό $|y_i - \hat{y_i}| \leq \varepsilon$ $\forall\, i = 1, 2, \dots, n$, μια θετική χαλαρή μεταβλητή ξ_i^{+} ή ξ_i^{-} ως ποινή, ανάλογα με το αν είναι πάνω (ξ_i^{+}) ή κάτω (ξ_i^{-}) από τον σωλήνα αντίστοιχα. Οι ξ_i^{+} και ξ_i^{-} μας πληροφορούν για το πόσο πρέπει να διορθώσουμε αυτές τις παρατηρήσεις, ώστε να μετακινηθούν στο εσωτερικό του σωλήνα. Πιο συγκεκριμένα ορίζουμε:

$$\xi_i^{+} = \begin{cases} 0 \text{ , αν } y_i - \hat{y_i} \leq \varepsilon \\ |y_i - \hat{y_i}| - \varepsilon \text{ , αλλιώς} \end{cases} \quad \forall\, i = 1, 2, \dots, n$$

$$\xi_i^{-} = \begin{cases} 0 \text{ , αν } \hat{y_i} - y_i \leq \varepsilon \\ |y_i - \hat{y_i}| - \varepsilon \text{ , αλλιώς} \end{cases} \quad \forall\, i = 1, 2, \dots, n$$



Σχήμα 23: Επίλυση προβλήματος παλινδρόμησης με χρήση SVM. Παρατηρούμε ότι οι κυκλωμένες με κόκκινο παρατηρήσεις βρίσκονται εκτός του **ε** - insensitive tube και άρα τους χορηγούνται οι ποινές $\boldsymbol{\xi}^{+}$ και $\boldsymbol{\xi}^{-}$.

Εφόσον κάναμε χρήση χαλαρών μεταβλητών για την SVM παλινδρόμηση, οφείλουμε να τις συμπεριλάβουμε στους περιορισμούς, καθώς και να προσθέσουμε το ανάλογο τους στην αντικειμενική συνάρτηση που προβλήματος βελτιστοποίησης. Επομένως έχουμε:

$$\min \frac{1}{2}\|w\|^2 + C \cdot \sum_i^n \left(\xi_i^{+} + \xi_i^{-}\right) \quad \text{υπό τους περιορισμούς} \begin{cases} y_i - \hat{y_i} \leq \xi_i^{+} + \varepsilon \\ \hat{y_i} - y_i \leq \xi_i^{-} + \varepsilon \\ \xi_i^{+} \geq 0, \ \xi_i^{-} \geq 0 \end{cases}$$

όπου $i = 1, 2, \dots, n$.

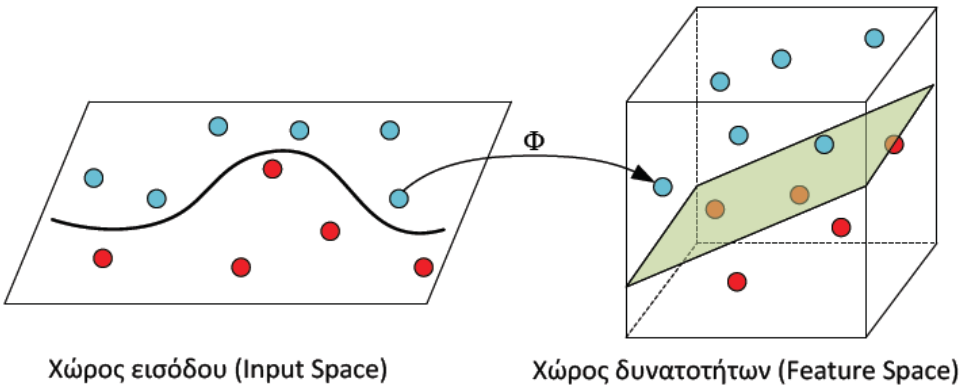
Η διαδικασία που ακολουθούμε στη συνέχεια είναι όμοια με αυτήν στην δυαδική ταξινόμηση. Εισάγουμε, δηλαδή, πολλαπλασιαστές Lagrange $\overrightarrow{\alpha^{+}} = (a_1^{+}, a_2^{+}, \dots, a_n^{+}), \overrightarrow{\alpha^{-}} = (a_1^{-}, a_2^{-}, \dots, a_n^{-})$ και $\overrightarrow{\mu^{+}} = (\mu_1^{+}, \mu_2^{+}, \dots, \mu_n^{+})$, $\overrightarrow{\mu^{-}} = (\mu_1^{-}, \mu_2^{-}, \dots, \mu_n^{-})$, με $\alpha_i^{+} \geq 0$, $\alpha_i^{-} \geq 0$ και $\mu_i^{+} \geq 0$, $\mu_i^{-} \geq 0 \ \forall i$. Ορίζουμε την Λαγκρανζιανή L_p, την παραγωγίζουμε ως προς w, b, ξ_i^{+} και ξ_i^{-} θέτοντας τις παραγώγους ίσες με το μηδέν, αντικαθιστούμε τις σχέσεις που προκύπτουν και ορίζουμε την L_D, την οποία και τέλος μεγιστοποιούμε ως προς τα α_i^{+} και α_i^{-} υπό την συνθήκη $\alpha_i^{+} \geq 0$ και $\alpha_i^{-} \geq 0$. Με αυτό τον τρόπο βρίσκουμε τις παραμέτρους w και b που χρειαζόμαστε για να οριστεί η SVM.

3.1.3 Συναρτήσεις κελύφους - πυρήνα (kernel functions)

Στον πραγματικό κόσμο τα δεδομένα δεν είναι σχεδόν ποτέ γραμμικά διαχωριζόμενα. Για την αντιμετώπιση αυτής της κατάστασης χρησιμοποιούνται οι kernel functions, που αποτελούν πολύ δημοφιλή περιοχή της μηχανικής μάθησης. Στόχος μας είναι μέσω μιας συνάρτησης $k(x, y)$, η οποία αποτελεί την συνάρτηση πυρήνα (*kernel function*), να προβάλουμε τα δεδομένα σε έναν χώρο μεγαλύτερης διάστασης, ώστε να πετύχουμε έναν γραμμικό διαχωρισμό για αυτά με όσο το δυνατόν λιγότερα σφάλματα. Για να γίνει αυτό πηγαίνουμε αρχικά από τον χώρο εισόδου (*input space*), που περιέχει τα x_i του training set, σε έναν μετασχηματισμένο χώρο χαρακτηριστικών (*feature space*) υψηλότερης διάστασης μέσω μιας μη γραμμικής απεικόνισης $\Phi(x)$. Η συνάρτηση πυρήνα ορίζεται τότε ως:

$$k(x_i, x_j) = \Phi(x_i)^T \Phi(x_j)$$

Δηλαδή είναι ένα εσωτερικό γινόμενο μεταξύ δύο διανυσμάτων $\Phi(x_i)$ και $\Phi(x_j)$ του καινούριου χώρου χαρακτηριστικών, χωρίς να απαιτείται σε κάποιο σημείο ο αναλυτικός υπολογισμός της απεικόνισης $\Phi(x)$.



Σχήμα 24: Απεικόνιση συνάρτησης πυρήνα. Από τον χώρο εισόδου, ο οποίος περιέχει όλες τις εγγραφές $\mathbf{x_i}$, πηγαίνουμε μέσω της **Φ** στον χώρο δυνατοτήτων, που αποτελείται από τα $\mathbf{\Phi(x_i)}$.

Όπως είδαμε, κατά την εφαρμογή της SVM μεθόδου δημιουργήσαμε έναν πίνακα $H_{ij} = y_i y_j k(x_i, x_j) = y_i y_j x_i x_j$. Χρησιμοποιήσαμε δηλαδή μια γραμμική συνάρτηση πυρήνα $k(x_i, x_j) = x_i x_j = x_i^T x_j$.

Πολύ σημαντικό είναι εδώ ότι υπάρχουν συγκεκριμένες απαιτήσεις για μια συνάρτηση, ώστε να μπορεί να χρησιμοποιηθεί ως συνάρτηση πυρήνα. Οι

απαιτήσεις αυτές είναι εκτός των σκοπών της παρούσας εργασίας και δεν θα εξεταστούν.

Τύποι kernel function

Υπάρχουν πολλοί τύποι συναρτήσεων πυρήνα. Μερικοί από αυτούς είναι:

- Γραμμικός (*linear*) $\rightarrow k(x_i, x_j) = x_i x_j = {x_i}^T x_j$
- Πολυωνυμικός (*polynomial*) $\rightarrow k(x_i, x_j) = (x_i \cdot x_j + a)^b$
- Σιγμοειδής (*sigmoid*) $\rightarrow k(x_i, x_j) = \tanh(a \cdot x_i \cdot x_j - b)$
- Ακτινική Βάση Πυρήνα (*Radial Basis Function - RBF*) $\rightarrow k(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}}$

Μειονεκτήματα χρήσης των συναρτήσεων πυρήνα

Σύμφωνα με τον Noble (2006), μερικά μειονεκτήματα των συναρτήσεων πυρήνα είναι τα ακόλουθα:

i. Καθιστούν το μοντέλο ευάλωτο σε υπερπροσαρμογή (*overfitting*[5]).
ii. Προβάλλοντας τα δεδομένα σε πολλές διαστάσεις μπορεί να καταλήξουμε σε υπερβολική εξειδίκευση.
iii. Σε οποιοδήποτε πρόβλημα είναι δύσκολο να βρεθεί ποια είναι η καταλληλότερη συνάρτηση πυρήνα. Ο μελετητής θα πρέπει να δοκιμάσει όλες τις πιθανές συναρτήσεις και στην συνέχεια να αξιολογήσει ποια του προσφέρει τα καλύτερα αποτελέσματα.

3.1.4 Μέθοδοι επιλογής παραμέτρων – επιλογής μοντέλου για SVM

Η απόδοση των μηχανών διανυσμάτων υποστήριξης επηρεάζεται σημαντικά από την επιλογή των παραμέτρων, που απαιτούνται για τον ορισμό τους. Αυτοί είναι οι πυρήνες, η σταθερά κόστους C και η παράμετρος ε στην περίπτωση της παλινδρόμησης. Δηλαδή, ενώ γενικά η SVM τεχνική θεωρείται αρκετά ακριβής μέθοδος, για ορισμένα σύνολα δεδομένων η απόδοση της είναι πολύ ευαίσθητη ως προς το πώς επιλέγονται οι παραπάνω παράμετροι. Ως εκ τούτου, ο χρήστης πρέπει κανονικά να διεξάγει εκτεταμένες διαδικασίες cross validation[6],

[5] **Overfitting**: υπερμοντελοποίηση των δεδομένων εκπαίδευσης. Ο αλγόριθμος ταξινομεί μόνο τα δεδομένα εκπαίδευσης, ενώ σε άγνωστα δεδομένα παρουσιάζει μεγάλο σφάλμα.

[6] **Cross validation:** μέθοδος αξιολόγησης ενός μοντέλου (βλ. κεφάλαιο 4).

προκειμένου να υπολογίσει την βέλτιστη ρύθμιση παραμέτρων. Η διαδικασία αυτή αναφέρεται συνήθως ως επιλογή μοντέλου. Μια κοινώς χρησιμοποιούμενη μέθοδος επιλογής παραμέτρων SVM είναι το πλέγμα αναζήτησης (*grid search – GS*), η οποία όμως είναι αρκετά χρονοβόρα. Για την αντιμετώπιση αυτού του ζητήματος έχουν διεξαχθεί τα τελευταία χρόνια πολλές μελέτες. Για παράδειγμα, στην εργασία των Lebrun et. al (2006) προτείνεται μια νέα μέθοδο μάθησης για την κατασκευή μιας δίτιμης συνάρτησης αποφάσεων (*Binary Decision function - BDF*) στις μηχανές διανυσμάτων υποστήριξης μειώνοντας την πολυπλοκότητα και καθιστώντας αποτελεσματική τη γενίκευση, με στόχο την κατασκευή ενός γρήγορου και αποτελεσματικού SVM ταξινομητή. Λεπτομερέστερα, ορίζεται ένα κριτήριο για την αξιολόγηση της ποιότητας της συνάρτησης αποφάσεων (*Decision function Quality - DFQ*), η οποία λαμβάνει υπ' όψιν το ποσοστό αναγνώρισης και την πολυπλοκότητα της BDF. Για την απλοποίηση του συνόλου εκπαίδευσης χρησιμοποιείται Vector Quantization (*VQ*). Η επιλογή μοντέλου γίνεται με βάση την επιλογή του απλούστερου επιπέδου, ενός υποσυνόλου χαρακτηριστικών και των παραμέτρων του SVM (hyperparameters) και εκτελείται για την βελτιστοποίηση της DFQ. Ο χώρος όπου γίνεται η αναζήτηση για την επιλογή του καλύτερου μοντέλου είναι τεράστιος, έτσι χρησιμοποιείται ο Tabu Search (*TS*) για να βρεθεί ένα καλό υπο-βέλτιστο μοντέλο σε ευάγωγες περιπτώσεις.

3.2 Μέθοδος SVM – RFE (Support Vector Machines - Recursive Feature Elimination) / Αναδρομική εξάλειψη χαρακτηριστικών με χρήση ταξινομητών SVM

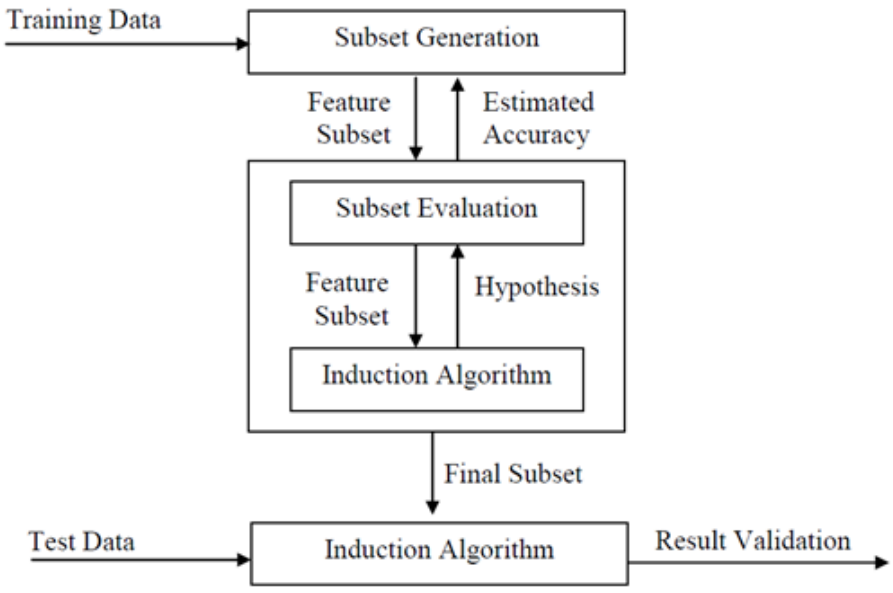
Καθώς ο αλγόριθμος SVM – RFE αποτελεί μια μέθοδο επιλογής χαρακτηριστικών (*feature selection*), θα κάνουμε πρώτα μια εισαγωγή στην feature selection, ώστε στη συνέχεια να γίνουν πιο κατανοητά από τον αναγνώστη ο σκοπός και η λειτουργία του αλγορίθμου.

3.2.1 Εισαγωγή στην Επιλογή Χαρακτηριστικών (Feature selection)

Εξαιτίας του τεράστιου όγκου δεδομένων που παράγονται καθημερινά, καθίσταται πλέον πολύ δύσκολο από έναν ερευνητή να επιλέξει ποιά από όλα αυτά είναι χρήσιμα για την επίλυση ενός συγκεκριμένου προβλήματος. Για να κατανοήσουμε πιο εύκολα την κατάσταση θα παρουσιάσουμε ένα απλό πρόβλημα επιλογής χαρακτηριστικών. Έστω, λοιπόν, ότι ένας ασθενής

χρειάζεται να κάνει κάποιες εξετάσεις, ώστε να διαπιστώσει την πιθανότητα να εμφανίσει στο μέλλον καρδιακό νόσημα. Στην περίπτωση αυτή, ο γιατρός που θα κάνει την διάγνωση θα ανατρέξει πρώτα στο ιστορικό του ασθενή, δηλαδή σε μια βάση δεδομένων, για να εξάγει κάποια συμπεράσματα. Τον ενδιαφέρει να μελετήσει χαρακτηριστικά, όπως το βάρος, η ηλικία, η αρτηριακή πίεση, τα επίπεδα χοληστερίνης κ.α. Παρ' όλα αυτά, στο ιστορικό ενός ασθενή αναγράφονται και άλλα χαρακτηριστικά, για παράδειγμα το επάγγελμα, ο τόπος καταγωγής και η οικογενειακή κατάσταση, τα οποία είναι άχρηστα για το πρόβλημα που μελετάται, καθώς δεν προσφέρουν κάποια πληροφορία σχετικά με το αν ο ασθενής πρόκειται να προσβληθεί από κάποιο καρδιακό νόσημα. Έτσι, ο γιατρός θα τα αγνοήσει, αφού όμως πρώτα ξοδέψει κάποιο χρόνο στο να απομονώσει τα χαρακτηριστικά που είναι εν τέλει απαραίτητο να μελετήσει.

Μέσω του παραπάνω παραδείγματος κατανοούμε τη βασική ιδέα γύρω από το feature selection. Ότι δηλαδή για να σχεδιάσουμε ένα μοντέλο ταξινόμησης, αντί να χρησιμοποιήσουμε όλα τα διαθέσιμα χαρακτηριστικά, επιλέγουμε ένα υποσύνολο χαρακτηριστικών, συγκεκριμένα το υποσύνολο χρήσιμων χαρακτηριστικών, αποκλείοντας όσα είναι περιττά (περιέχουν πληροφορία που δίνεται και από άλλα χαρακτηριστικά) ή άσχετα με το πρόβλημα. Έτσι, μπορούμε να πούμε ότι το feature selection είναι η διαδικασία, στην οποία δοσμένου ενός συνόλου $F = \{x_i : i = 1, 2, \dots, p\}$ p χαρακτηριστικών και μιας μεταβλητής Y, στόχος είναι να επιλεγεί από το διάστημα F ένα υποσύνολο, το οποίο θα αποτελείται από $m << p$ χαρακτηριστικά και θα χαρακτηρίζει με βέλτιστο τρόπο την μεταβλητή Y, χωρίς να έχει προηγηθεί ο οποιοσδήποτε μετασχηματισμός στα δεδομένα.



Σχήμα 25: Σχηματική απεικόνιση της διαδικασίας επιλογής χαρακτηριστικών.

Προφανώς, η αξιολόγηση ενός χαρακτηριστικού ως χρήσιμο ή μη δεν είναι εύκολη διαδικασία. Αυτό είναι και το κύριο αντικείμενο μελέτης στο πρόβλημα της επιλογής χαρακτηριστικών. Η δυσκολία επιλογής του υποσυνόλου χρήσιμων χαρακτηριστικών οφείλεται στους εξής παράγοντες:

i. Υποθέτοντας ότι διαθέτουμε p χαρακτηριστικά, το πλήθος των υποσυνόλων χαρακτηριστικών που μπορούν να επιλεγούν ως χρήσιμα είναι 2^p. Έχουμε δηλαδή έναν τεράστιο αριθμό πιθανών επιλογών.
ii. Η ποιότητα ενός υποσυνόλου χαρακτηριστικών εξαρτάται από πολλές παραμέτρους, γεγονός που κάνει δύσκολο τον ορισμό ενός αντικειμενικού κριτηρίου για την αξιολόγηση τους.

Στα προβλήματα ταξινόμησης, ανάλογα με το πώς και το πότε αξιολογείται η χρησιμότητα των υποψήφιων προς επιλογή υποσυνόλων χαρακτηριστικών, διακρίνουμε τις παρακάτω μεθόδους feature selection, που χωρίζονται σε τρεις κατηγορίες:

i. Κατηγορία των φίλτρων (*filters*): δεν βασίζονται σε κάποιο ταξινομητή για να εκτιμήσουν την ποιότητα ενός υποσυνόλου χαρακτηριστικών, αλλά χρησιμοποιούν στατιστικά μέτρα με στόχο να εντοπίσουν συναφή χαρακτηριστικά. Στην κατηγορία αυτή εντάσσονται οι:
   - Μονοπαραγοντικές μέθοδοι (*univariate methods*)
   - Πολυπαραγοντικές μέθοδοι (*multivariate methods*)

ii. Κατηγορία των συσκευαστών (*wrappers*): χρησιμοποιούν την ακρίβεια ταξινόμησης ως κριτήριο αξιολόγησης των υποσυνόλων, γεγονός που απαιτεί την κατασκευή ενός ταξινομητή για κάθε σύνολο χαρακτηριστικών που εξετάζεται, με αρνητικό επακόλουθο το αυξημένο υπολογιστικό κόστος. Στην κατηγορία αυτή εντάσσονται οι:
   - Μέθοδος της προς τα εμπρός επιλογής (*forward selection*)
   - Μέθοδος της προς τα πίσω επιλογής (*backward selection*)

iii. Κατηγορία embedded μεθόδων: έχουν παρόμοια φιλοσοφία με τους wrappers, με τη διαφορά ότι αξιολογούν τα υποσύνολα χαρακτηριστικών με βάση το πώς επηρεάζονται κάποιοι παράμετροι, οι οποίοι εμπλέκονται στην διαδικασία εκπαίδευσης του ταξινομητή. Στην κατηγορία αυτή εντάσσονται:
   - Τα δέντρα απόφασης (*decision trees*)
   - Η μέθοδος της ελάχιστης απόλυτης συστολής και επιλογής φορέα (*least absolute shrinkage and selection operator – LASSO*)
   - Η μέθοδος του τυχαίου πολυωνυμικού λογαρίθμου (*random multinomial logit – RMNL*)

- Η μέθοδος αναδρομικής εξάλειψης χαρακτηριστικών με χρήση ταξινομητών SVM (*SVM - RFE*)

Πλεονεκτήματα

Μέσω της επιλογής χαρακτηριστικών:

- Μειώνεται η υπολογιστική πολυπλοκότητα.
- Διευκολύνεται η οπτικοποίηση και η κατανόηση των δεδομένων. Γενικά, διακρίνοντας ποια χαρακτηριστικά είναι πιο σημαντικά για το αποτέλεσμα μιας διαδικασίας, μπορεί να αποκτηθεί μια πιο σαφής αίσθηση του πραγματικού προβλήματος, επιτρέποντας έτσι στους ειδικούς του τομέα να το αντιμετωπίσουν αποτελεσματικότερα. Αυτή η πτυχή του feature selection είναι ιδιαίτερα σημαντική για προβλήματα βιοπληροφορικής, καθώς συμβάλλει στην αναγνώριση γονιδίων που σχετίζονται με διάφορες νόσους.
- Μειώνεται η ποσότητα των δεδομένων, που απαιτούνται για την εκπαίδευση και την βελτίωση της προγνωστικής ακρίβειας των αλγορίθμων, οι οποίοι χρησιμοποιούνται στην ταξινόμηση.

3.2.2 Περιγραφή μεθόδου SVM – RFE

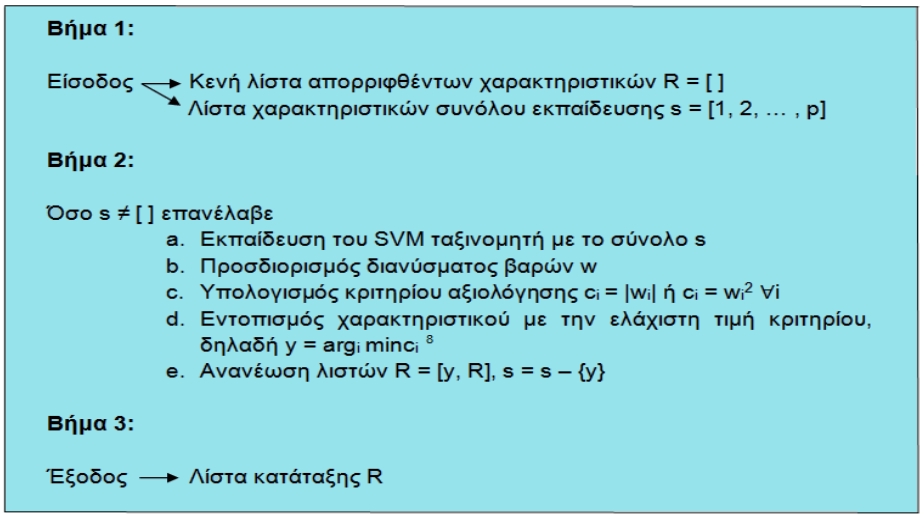
Η μέθοδος SVM – RFE προτάθηκε από τους Guyon, Weston, Barnhill και Vapnik (Guyon et al., 2002). Λόγω της επιτυχίας της στην επιλογή αντιπροσωπευτικών γονιδίων για την ταξινόμηση μορφών καρκίνου, η SVM – RFE απέκτησε μεγάλη δημοτικότητα και θεωρείται μια από τις πιο γνωστές και αποτελεσματικές μεθόδους επιλογής χαρακτηριστικών.

Με λίγα λόγια, σύμφωνα με τον Guyon et al. (2002), ο αλγόριθμος SVM – RFE είναι μια embedded μέθοδος επιλογής χαρακτηριστικών, που εκτελεί μια διαδικασία backward elimination[7], σε κάθε βήμα της οποίας διαγράφεται το λιγότερο σημαντικό χαρακτηριστικό. Λιγότερο σημαντικό θεωρείται το χαρακτηριστικό, του οποίου η διαγραφή θα προκαλούσε την μικρότερη μείωση του περιθωρίου SVM.

---

[7]**Backward elimination (προς τα πίσω επιλογή):** μέθοδος feature selection. Το αρχικό υποσύνολο περιέχει όλα τα χαρακτηριστικά. Σε κάθε επανάληψη εξετάζονται όλα τα υποσύνολα που προκύπτουν από την διαγραφή ενός χαρακτηριστικού από το τρέχον υποσύνολο. Τελικά, διαγράφεται αυτό του οποίου η απουσία οδηγεί στη μεγαλύτερη απόδοση ως προς το κριτήριο αξιολόγησης.

Πιο συγκεκριμένα, ο αλγόριθμος ξεκινά εκπαιδεύοντας τον ταξινομητή SVM, με το training set να περιέχει όλα τα διαθέσιμα χαρακτηριστικά του προβλήματος. Υπολογίζεται έτσι το διάνυσμα βαρών w και η πόλωση (*bias*) b, με τον τρόπο που περιγράψαμε αναλυτικά στην παράγραφο 3.1.1. Υποθέτοντας ότι χρησιμοποιήθηκε γραμμικός πυρήνας κατά την εκπαίδευση, ως κριτήριο αξιολόγησης των χαρακτηριστικών λαμβάνεται είτε η απόλυτη τιμή των συνιστωσών του διανύσματος βαρών w, $|w_i|$, είτε το τετράγωνο των συνιστωσών του διανύσματος βαρών w, ${w_i}^2$, όπου $i = 1, 2, \dots, p$, με το p να αποτελεί τον αριθμό των χαρακτηριστικών. Έτσι, το χαρακτηριστικό που προκαλεί την μικρότερη μείωση του περιθωρίου SVM είναι αυτό που αντιστοιχεί στο ελάχιστο $|w_i|$ ή στο ελάχιστο ${w_i}^2$. Δηλαδή, αν για παράδειγμα προκύψει ότι $\{\min|w_i|\}_{i=1}^{p} = |w_2|$, όμοια $\{\min {w_i}^2\}_{i=1}^{p} = {w_2}^2$, τότε το χαρακτηριστικό 2 απορρίπτεται και ο SVM ταξινομητής εκπαιδεύεται εκ νέου με τα εναπομείναντα χαρακτηριστικά $1, 3, 4, \dots, p$. Ακολουθώντας επαναληπτικά την παραπάνω διαδικασία, ο αλγόριθμος τερματίζει όταν δεν απομείνει κανένα χαρακτηριστικό. Παρατηρούμε ότι η μέθοδος SVM – RFE απαιτεί μια εκπαίδευση του ταξινομητή SVM σε κάθε backward βήμα προκειμένου να αποφασίσει ποιό χαρακτηριστικό θα διαγραφεί. Χρειάζονται δηλαδή p εκπαιδεύσεις υποθέτοντας ότι πρέπει να διαγραφούν p χαρακτηριστικά. Στην περίπτωση που δεν χρησιμοποιείται γραμμικός πυρήνας, η εύρεση του χαρακτηριστικού, του οποίου η αφαίρεση οδηγεί στη μικρότερη μείωση περιθωρίου SVM, απαιτεί πιο πολύπλοκους υπολογισμούς και είναι εκτός των σκοπών της συγκεκριμένης διπλωματικής εργασίας. Παρακάτω παρουσιάζουμε έναν ψευδοκώδικα για τον αλγόριθμο SVM – RFE, για μια πιο σύντομη, αλλά ξεκάθαρη, περιγραφή της μεθόδου.



Σχήμα 26: Ψευδοκώδικας για τον αλγόριθμο SVM – RFE.

[8] **$\text{arg}_i \min c_i$:** η τιμή i που δίνει το $\min c_i$, δηλαδή την ελάχιστη τιμή κριτηρίου.

3.2.3 Εφαρμογή της μεθόδου SVM – RFE

Για την εφαρμογή της μεθόδου SVM – RFE θα χρησιμοποιήσουμε ένα σετ δεδομένων (*data set*) από τον ιστότοπο UCI Machine Learning Repository. Συγκεκριμένα, κατεβάζουμε το “Statlog (Heart) Data Set”, το οποίο αφορά την καρδιοπάθεια. Στόχος είναι η εξέταση της ύπαρξης καρδιακής πάθησης σε ένα σύνολο ασθενών, βάσει ορισμένων χαρακτηριστικών.

Περιγραφή συνόλου δεδομένων

Το “Statlog (Heart) Data Set” αποτελεί μια ελαφρώς διαφορετική μορφή του “Heart Disease Data Set” από τον ίδιο ιστότοπο. Ενώ το “Heart Disease Data Set” περιέχει συνολικά 76 χαρακτηριστικά (*features / attributes*), όλες οι δημοσιευμένες μελέτες, στις οποίες χρησιμοποιήθηκε, παραπέμπουν στην χρήση ενός υποσυνόλου 14 χαρακτηριστικών από τα διαθέσιμα 76. Το “Heart Disease Data Set” αποτελείται από 4 σύνολα δεδομένων: ένα από το Cleveland, ένα από την Ουγγαρία, ένα από την Ελβετία και ένα από την VA Long Beach.

Το “Statlog (Heart) Data Set”, που θα χρησιμοποιήσουμε εδώ, αποτελείται από 270 παρατηρήσεις – περιπτώσεις ασθενών (*instances*) και 14 χαρακτηριστικά (13 features + 1 class). Αυτά αναλυτικά είναι:

1. Η ηλικία (*age*)
2. Το φύλο (*sex*): $1 = \text{άντρας}\ (male)$, $0 = \text{γυναίκα}\ (female)$
3. Το είδος πόνου στο στήθος (*chest pain type → cp*):
   $1 = \text{τυπική στηθάγχη}\ (typical\ angina)$,
   $2 = \text{άτυπη στηθάγχη}\ (atypical\ angina)$, $3 = \text{μη στηθαγχικός πόνος}\ (non-anginal\ pain)$, $4 = \text{ασυμπτωματικός}\ (asymptomatic)$
4. Η πίεση του αίματος σε κατάσταση ηρεμίας μετρημένη σε mm Hg κατά την εισαγωγή στο νοσοκομείο (*resting blood pressure in mm Hg on admission to the hospital → trestbps*)
5. Η χοληστερίνη ορού σε mg/dl (*serum cholestoral in mg/dl → chol*)
6. Το σάκχαρο στο αίμα σε κατάσταση νηστείας (*fasting blood sugar → fbs*)
7. Το αποτελέσματα ηλεκτροκαρδιογραφήματος σε κατάσταση ηρεμίας (*resting electrocardiographic results → restecg)*
8. Ο μέγιστος αριθμός καρδιακών παλμών (*maximum heart rate achieved → thalach*)
9. Η στηθάγχη που προκαλείται από άσκηση (*exercise induced angina → exang*): $1 = \text{ναι}\ (yes)$, $0 = \text{όχι}\ (no)$
10. Ένα από τα πορίσματα του ηλεκτροκαρδιογραφήματος, το λεγόμενο oldpeak (*ST depression induced by exercise relative to rest → oldpeak*)

11. Η κλήση του τμήματος ST στο ηλεκτροκαρδιογράφημα (*the slope of the peak exercise ST segment → slope*): $1 = \text{κλίση προς τα πάνω } (upsloping)$, $2 = \text{επίπεδη } (flat)$, $3 = \text{κλίση προς τα κάτω } (downsloping)$
12. Ο αριθμός των κύριων αιμοφόρων αγγείων (0 – 3) που χρωματίστηκαν με ακτινοσκόπηση (*number of major vessels (0 – 3) colored by flourosopy → ca*)
13. Η συγκέντρωση θαλλίου στην περιοχή της καρδιάς (*thal*): $3 = \text{φυσιολογική } (normal)$, $6 = \text{σταθερή βλάβη } (fixed\ defect)$, $7 = \text{αναστρέψιμη βλάβη } (reversible\ defect)$
14. Η παρουσία ή απουσία καρδιοπάθειας (*absence or presence of heart disease*) – η κλάση: $1 = \text{απουσία } (absence)$, $2 = \text{παρουσία } (presence)$

Ουσιαστικά, έχουμε μια μεταβλητή απόκρισης Υ, η οποία μέσω 13 επεξηγηματικών μεταβλητών $\mathrm{x}_1, \mathrm{x}_2, \dots, \mathrm{x}_{13}$ αποφαίνεται για το εάν ένας ασθενής πάσχει (τιμή 2) ή όχι (τιμή 1) από καρδιοπάθεια.

Εισαγωγή και επεξεργασία των δεδομένων στην R

Κατεβάζουμε λοιπόν το “Statlog (Heart) Data Set” και εισάγουμε τα δεδομένα σε ένα πίνακα στην R μέσω των εντολών *read.table()* και *as.matrix()*. Στη συνέχεια, τα διασπούμε σε διανύσματα και μετατρέπουμε σε παράγοντες (εντολή *as.factor()*) όσα χαρακτηριστικά είναι κατηγορικές μεταβλητές. Μέσω της εντολής *cbind()* ενώνουμε πάλι τα δεδομένα σε ένα πίνακα, που περιέχει και τις 14 μεταβλητές, $\mathrm{x}_1, \mathrm{x}_2, \dots, \mathrm{x}_{13}$ και Υ. Τώρα, για να μπορέσουμε να χρησιμοποιήσουμε την μέθοδο SVM, κατ’ επέκταση την συνάρτηση *svm()* της R, κατεβάζουμε το πακέτο “e1071” από την βιβλιοθήκη της R (εντολή *install.packages(“e1071”)*) και το φορτώνουμε (εντολή *library(e1071)*). Σειρά έχει η διάσπαση των δεδομένων σε training και test set. Όπως είναι και το πιο σύνηθες, θα χρησιμοποιήσουμε το 75% των δεδομένων για την εκπαίδευση του ταξινομητή SVM και το υπόλοιπο 25% για την επικύρωση. Έτσι, μέσω της εντολής *sample()* διαλέγουμε τυχαία το 25% των 270 παρατηρήσεων (67 instances), και συνθέτουμε το test set. Οι υπόλοιπες 203 παρατηρήσεις αποτελούν το training set. Αναλυτικά οι εντολές για όλα τα παραπάνω βρίσκονται στο παράρτημα 1 (1, 2, 3).

Κατασκευή κώδικα για τον αλγόριθμο SVM – RFE

Με βάση το ψευδοκώδικα που σχεδιάσαμε στην προηγούμενη παράγραφο, κατασκευάζουμε στην R τον κώδικα της μεθόδου SVM – RFE. Δημιουργούμε δηλαδή μια συνάρτηση (*function(x,y)*), η οποία λαμβάνει ως είσοδο έναν πίνακα X και ένα διάνυσμα y. Ο πίνακας X περιέχει τις τιμές των χαρακτηριστικών μόνο

για τις παρατηρήσεις του training set, δηλαδή οι στήλες του είναι τα $x_1, x_2, \dots, x_{13}$, όπου κάθε $x_i \in R^{203}$, και το διάνυσμα y αποτελεί την μεταβλητή απόκρισης, περιέχει δηλαδή σε ποια κλάση ανήκει κάθε ασθενής του συνόλου εκπαίδευσης με κάθε $y_j \in \{1, 2\}$, όπου $j = 1, 2, \dots, 203$. Στη συνέχεια, δημιουργούμε τα διανύσματα $s = [1, 2, \dots, 13]$ και R, όπως τα ονομάσαμε στον ψευδοκώδικα. Το R είναι και αυτό που θα λάβουμε ως έξοδο, καθώς περιέχει τα χαρακτηριστικά σε αύξουσα σειρά ανάλογα με την σημαντικότητα τους, με το τελευταίο χαρακτηριστικό να εκτιμάται από την μέθοδο ως το πιο σημαντικό. Ο κώδικας συνεχίζει και όσο το μήκος του s είναι μεγαλύτερο του μηδενός, δηλαδή διάφορο του κενού, εκπαιδεύει τον ταξινομητή SVM με χρήση του σιγμοειδή πυρήνα (εντολή *svm()*), υπολογίζει το διάνυσμα βαρών w (εντολή *crossprod()*) και το κριτήριο αξιολόγησης $c = w^2$, ταξινομεί τις συνιστώσες του c (εντολή *sort()*) και βρίσκει την μικρότερη, και τέλος ανανεώνει τα διανύσματα R και s. Όταν η επανάληψη λήξει, ο κώδικας μέσω της εντολής *return()* επιστρέφει το διάνυσμα απορριφθέντων χαρακτηριστικών R. Αναλυτικά ο κώδικας βρίσκεται στο παράρτημα 1 (4).

Εφαρμογή του αλγορίθμου SVM – RFE στο σετ δεδομένων "Statlog (Heart) Data Set" και αποτελέσματα

Δίνοντας ως είσοδο $x = \text{features}$ και $y = \text{response}$ (features: ο πίνακας χαρακτηριστικών του training set του σετ δεδομένων μας "Statlog (Heart) Data Set" χωρίς την τελευταία στήλη που αποτελεί την κλάση, responce: η τελευταία στήλη του πίνακα χαρακτηριστικών του training set του σετ δεδομένων μας "Statlog (Heart) Data Set") στον κώδικα που κατασκευάσαμε, λαμβάνουμε το παρακάτω διάνυσμα απορριφθέντων χαρακτηριστικών:

6 2 11 7 9 12 3 10 13 1 4 5 8

Συμπεραίνουμε λοιπόν ότι το χαρακτηριστικό 6, δηλαδή η ποσότητα του σακχάρου στο αίμα σε κατάσταση νηστείας (*fasting blood sugar → fbs*), είναι το λιγότερο σημαντικό από τα 14 χαρακτηριστικά, καθώς προκαλεί την μικρότερη μείωση περιθωρίου SVM σε σχέση με τα υπόλοιπα. Αντίθετα, το χαρακτηριστικό 8, δηλαδή ο μέγιστος αριθμός καρδιακών παλμών (*maximum heart rate achieved → thalach*) εκτιμάται από την μέθοδο SVM – RFE ως το σημαντικότερο feature. Άρα μπορούμε να πούμε ότι η αξιολόγηση του αριθμού των καρδιακών παλμών συμβάλει σε μεγάλο ποσοστό στη ανίχνευση καρδιοπάθειας.

Παρατήρηση

Πριν κλείσουμε, οφείλουμε να καταγράψουμε τι συμβαίνει στην περίπτωση όπου η εκπαίδευση του SVM ταξινομητή γίνεται με διαφορετικό training set. Χρησιμοποιώντας για ακόμη μια φορά την εντολή *sample()*, θα προκύψει μια διαφορετική διάσπαση του συνόλου δεδομένων “Statlog (Heart) Data Set” σε 25% και 75% για επικύρωση και εκπαίδευση αντίστοιχα. Οπότε, ξανατρέχουμε τον κώδικα που φτιάξαμε στην R για την μέθοδο SVM –RFE με είσοδο αυτή τη φορά τον νέο πίνακα features και το νέο διάνυσμα responce, που προέκυψαν μέσω του νέου training set. Το αποτέλεσμα που παίρνουμε είναι το ακόλουθο:

6 2 9 11 7 3 12 10 13 1 4 5 8

Παρατηρούμε λοιπόν ότι συγκρίνοντας το παραπάνω αποτέλεσμα με το αρχικό, τα πρώτα και τελευταία στοιχεία του διανύσματος απορριφθέντων στοιχείων R εμφανίζονται ίδια και στις δύο περιπτώσεις. Δηλαδή, τα χαρακτηριστικά 6 και 2 που ήταν τα λιγότερο σημαντικά παραμένουν ως τα λιγότερο σημαντικά, όπως επίσης και τα χαρακτηριστικά 13, 1, 4, 5 και 8 που ήταν τα σημαντικότερα παραμένουν κυρίαρχα.

Επομένως, συμπεραίνουμε ότι πρέπει να υπάρχει μια συμβατότητα μεταξύ των αποτελεσμάτων της μεθόδου SVM – RFE όταν χρησιμοποιούνται διαφορετικά training sets για την εκπαίδευση του ταξινομητή SVM. Η έξοδος, δηλαδή, του αλγορίθμου δεν επηρεάζεται στα βασικά σημεία της από το ποιο σύνολο εκπαίδευσης θα χρησιμοποιήσουμε και έτσι μπορούμε να αποφανθούμε ανεξάρτητα από αυτό για το ποια χαρακτηριστικά είναι εν τέλει σημαντικά και ποια όχι.

ΚΕΦΑΛΑΙΟ 4 : ΜΕΘΟΔΟΙ ΑΞΙΟΛΟΓΗΣΗΣ

4.1 Αξιολόγηση μοντέλων ταξινόμησης

Όπως αναφέραμε και στο κεφάλαιο 2, σύμφωνα με τον Hastie et al. (2001) η ιδιότητα γενίκευσης που κατέχει ένα μοντέλο ταξινόμησης σχετίζεται με την ικανότητα πρόβλεψης αυτού σε δεδομένα ανεξάρτητα και διαφορετικά από αυτά που χρησιμοποιήθηκαν κατά την εκπαίδευση του. Εύκολα κανείς κατανοεί ότι η αξιολόγηση της γενίκευσης - η αξιολόγηση της απόδοσης - ενός μοντέλου αποτελεί διαδικασία ζωτικής σημασίας, καθώς κατευθύνει τον ερευνητή στην επιλογή του καταλληλότερου αλγορίθμου για την περιγραφή των δεδομένων του και επίσης δίνει ένα μέτρο ποιότητας για το τελικώς επιλεγμένο μοντέλο.

Όπως τονίζεται στο σύγγραμμα του Hastie et al. (2001), η αξιολόγηση ενός μοντέλου περιλαμβάνει δύο στάδια - σκοπούς:

i. Την επιλογή του καταλληλότερου μοντέλου → ο ερευνητής εκτιμά τις αποδόσεις – την ικανότητα γενίκευσης πολλών μοντέλων ώστε να επιλέξει ποιό από όλα περιγράφει καλύτερα τα δεδομένα.

ii. Την εκτίμηση του τελικά επιλεχθέντος μοντέλου → ο ερευνητής, έχοντας αποφασίσει ποιο μοντέλο είναι το καλύτερο, οφείλει να εκτιμήσει το σφάλμα πρόβλεψης (σφάλμα γενίκευσης) του μοντέλου χρησιμοποιώντας ένα νέο σύνολο δεδομένων (*test set*).

Στην ιδανική περίπτωση όπου το σύνολο των διαθέσιμων δεδομένων είναι αρκετά μεγάλο, η καλύτερη προσέγγιση για τα δύο παραπάνω στάδια (επιλογή & εκτίμηση μοντέλου) είναι η διαίρεση του συνόλου δεδομένων τυχαία σε τρία μέρη (Hastie et al., 2001):

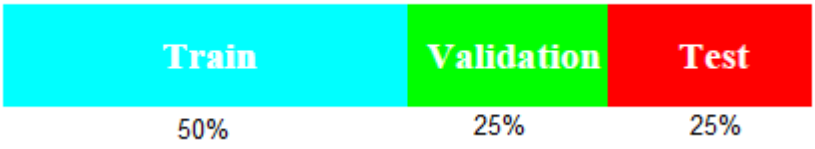
i. Ένα σύνολο εκπαίδευσης (*training set*) → χρησιμοποιείται προφανώς για την εκπαίδευση του μοντέλου

ii. Ένα σύνολο επικύρωσης (*validation set*) → χρησιμοποιείται για την εκτίμηση του σφάλματος πρόβλεψης (σφάλμα γενίκευσης) ενός μοντέλου. Με αυτόν τον τρόπο αξιολογείται η ικανότητα γενίκευσης κάθε μοντέλου, ώστε να αποφασιστεί ποιο είναι καταλληλότερο για τα δεδομένα.

iii. Ένα σύνολο δοκιμής (*test set*) → χρησιμοποιείται για τη εκτίμηση του σφάλματος γενίκευσης του τελικά επιλεχθέντος μοντέλου. Ιδανικά, το σύνολο

αυτό θα πρέπει να κρατείται απομονωμένο και να έρχεται στην επιφάνεια μόνο στο τελευταίο στάδιο της ανάλυσης δεδομένων, δηλαδή κατά την εκτίμηση του μοντέλου.

Γενικά, είναι δύσκολο να ορίσει κανείς έναν κανόνα για το πώς επιλέγεται ο αριθμός των παρατηρήσεων για κάθε ένα από τα τρία παραπάνω μέρη (*training, validation, test*), καθώς η επιλογή αυτή εξαρτάται από πολλές παραμέτρους. Μια τυπική διάσπαση των δεδομένων θα μπορούσε να είναι 50% αυτών για την εκπαίδευση, 25% για την επικύρωση και 25% για την δοκιμή (Hastie et al., 2001).



Σχήμα 27: Τυπική διάσπαση των δεδομένων σε σύνολα για εκπαίδευση (train), επικύρωση (validation)και δοκιμή (test) (Hastie et al., 2001).

Οι μέθοδοι που παρατίθενται σε αυτό το κεφάλαιο είναι σχεδιασμένες για καταστάσεις όπου το σύνολο των δεδομένων δεν είναι αρκετά μεγάλο, αλλά ανεπαρκές, γεγονός που καθιστά αδύνατο τον διαχωρισμό του σε τρία ανεξάρτητα μέρη.

4.1.1 Βασικοί τύποι σφαλμάτων

Θεωρούμε ότι έχουμε μια ποσοτική μεταβλητή απόκρισης - μια μεταβλητή στόχο Y, ένα διάνυσμα εισόδου X και ένα μοντέλο πρόβλεψης $\hat{f}(X)$, το οποίο έχει εκπαιδευτεί με χρήση ενός training set T. Τότε, κατά τον Hastie et al. (2001), η συνάρτηση που μετρά το σφάλμα μεταξύ της πραγματικής τιμής Y και της εκτιμώμενης από το μοντέλο τιμής $\hat{f}(X)$ συμβολίζεται με $L\left(Y,\hat{f}(X)\right)$, με τυπικές επιλογές τις παρακάτω:

$$L\left(Y, \hat{f}(X)\right) = \begin{cases} \left(Y - \hat{f}(X)\right)^2 \rightarrow \text{τετραγωνικό σφάλμα (}\textit{squared error}\text{)} \\ \left|Y - \hat{f}(X)\right| \rightarrow \text{απόλυτο σφάλμα (}\textit{absolute error}\text{)} \end{cases}$$

Μέσω της συνάρτησης $L\left(Y, \hat{f}(X)\right)$ θα ορίσουμε κάποια βασικά είδη σφαλμάτων, όπως αυτά καταγράφονται στο σύγγραμμα του Hastie et al. (2001), τα οποία είναι απαραίτητα για την κατανόηση των παρακάτω παραγράφων.

- **Σφάλμα δοκιμής (*test error*) ή σφάλμα γενίκευσης (*generalization error*):** είναι το σφάλμα πρόβλεψης πάνω σε ένα ανεξάρτητο σύνολο δοκιμών (*test set*). Ορίζεται ως η αναμενόμενη τιμή του τετραγώνου της διαφοράς μεταξύ της εκτιμήτριας και του ακριβούς στόχου (μέσο τετραγωνικό σφάλμα).

$$\mathrm{Err}_T = E\left[L\left(Y, \hat{f}(X)\right) \mid T\right]$$

Εδώ πρέπει να προσέξουμε ότι το σύνολο εκπαίδευσης είναι συγκεκριμένο, δηλαδή η τιμή που θα προκύψει για το σφάλμα δοκιμής θα αφορά αποκλειστικά και μόνο τις εγγραφές του training set T. Στόχος μας για την μελέτη της απόδοσης ενός μοντέλου είναι η εκτίμηση του Err_T.

- **Αναμενόμενο σφάλμα πρόβλεψης (*expected prediction error*) ή αναμενόμενο σφάλμα δοκιμής (*expected test error*):** είναι η μέση τιμή του σφάλματος γενίκευσης.

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

Το σφάλμα αυτό είναι πιο επιδεκτικό σε στατιστικές αναλύσεις και εκτιμάται αποτελεσματικά από τις περισσότερες μεθόδους αξιολόγησης μοντέλων. Γι' αυτόν τον λόγο, παρά το ότι όπως προαναφέραμε στόχος είναι η εκτίμηση του Err_T, εν τέλει υπολογίζεται και μελετάται το αναμενόμενο σφάλμα πρόβλεψης Err κατά την αξιολόγηση ενός μοντέλου.

- **Σφάλμα εκπαίδευσης (*training error*):** είναι η μέση απώλεια πάνω στο δείγμα εκπαίδευσης.

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

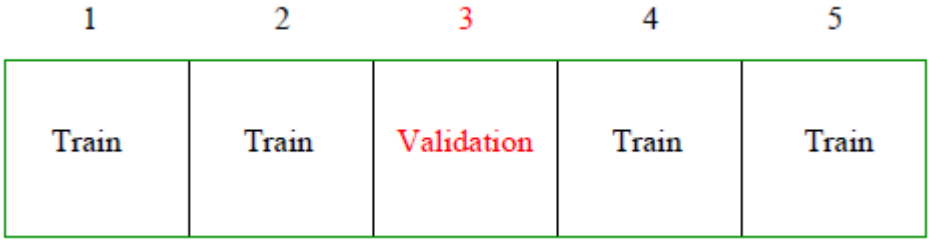
Το training error αποτελεί μια εκτιμήτρια για το σφάλμα γενίκευσης, αλλά δυστυχώς όχι καλή. Μειώνεται σταθερά ανάλογα με την πολυπλοκότητα του μοντέλου και μπορεί να πέσει μέχρι και την τιμή μηδέν σε περίπτωση μεγάλης αύξησης της. Ωστόσο, ένα μοντέλο με μηδενικό σφάλμα εκπαίδευσης θεωρείται ότι παρουσιάζει υπερπροσαρμογή (*overfit*) στα δεδομένα εκπαίδευσης και άρα δεν έχει καλή ικανότητα γενίκευσης.

4.1.2 Διασταυρωμένη επικύρωση (Cross Validation)

Η διασταυρωμένη επικύρωση (*cross validation*) είναι ίσως η πιο απλή και ευρέως χρησιμοποιούμενη στατιστική μέθοδος αξιολόγησης και σύγκρισης αλγορίθμων μάθησης (Hastie et al., 2001). Εκτιμά άμεσα το αναμενόμενο σφάλμα πρόβλεψης (*expected prediction error*) $\mathrm{Err} = \mathrm{E}\left[L\left(Y, \hat{f}(X)\right)\right]$ όταν το μοντέλο προσαρμόζεται σε ένα ανεξάρτητο σύνολο δοκιμής (*test set*). Η βασική μορφή διασταυρωμένης επικύρωσης είναι η k – fold cross validation. Οι περισσότερες από τις υπόλοιπες μορφές διασταυρωμένης επικύρωσης είναι ειδικές περιπτώσεις της k – fold ή περιλαμβάνουν επαναλαμβανόμενους γύρους της.

k – fold Διασταυρωμένη Επικύρωση (k – fold cross validation)

Σύμφωνα με τον Hastie et al. (2001), κατά την k – fold διασταυρωμένη επικύρωση τα διαθέσιμα δεδομένα χωρίζονται σε k, περίπου ισομεγέθη, υποσύνολα. Από αυτά, τα k – 1 χρησιμοποιούνται για την εκπαίδευση του μοντέλου και το ένα εναπομείναν υποσύνολο αποτελεί το validation set. Για παράδειγμα, για k = 5 έχουμε τον εξής πιθανό διαχωρισμό:



Σχήμα 28: Πιθανός διαχωρισμός των διαθέσιμων δεδομένων κατά την 5 – fold cross validation, όπως παρουσιάζει στο σύγγραμμα του Hastie et al. (2001).

Συγκεκριμένα, παρατηρούμε ότι 5 – 1 = 4 υποσύνολα (τα 1, 2, 4 και 5) συνθέτουν το training set και το υποσύνολο 3 χρησιμοποιείται για την επικύρωση του ταξινομητή, του οποίου θέλουμε να ελέγξουμε την απόδοση. Η διαδικασία αυτή, δηλαδή η τυχαία επιλογή k – 1 υποσυνόλων από τα k για εκπαίδευση και ενός από τα k για επικύρωση, πραγματοποιείται k φορές, έτσι ώστε κάθε υποσύνολο, που δημιουργήθηκε από τον διαχωρισμό των δεδομένων, να χρησιμοποιηθεί ακριβώς μια φορά ως validation set του μοντέλου. Αυτό είναι και το πλεονέκτημα της μεθόδου, ότι δηλαδή όλες οι παρατηρήσεις χρησιμοποιούνται τόσο στην εκπαίδευση όσο και στην επικύρωση.

Με αυτόν τον τρόπο παράγονται συνολικά k τιμές για το σφάλμα πρόβλεψης, των οποίων τέλος υπολογίζουμε την μέση τιμή ώστε να λάβουμε μια πιο ενιαία εκτίμηση του σφάλματος. Ακριβέστερα, έστω K : {1, 2, … , n} → {1, … , k} μια συνάρτηση «ευρετηρίου», η οποία υποδηλώνει σε ποιο υποσύνολο από τα 1, 2, … , k ανήκει η i – οστή παρατήρηση, με i = 1, 2, … , n, όπου n είναι το μέγεθος του συνόλου εκπαίδευσης (Hastie et al., 2001). Συμβολίζουμε με $\hat{f}^{-K}(x)$ την πρόβλεψη που έκανε το μοντέλο, του οποίου την απόδοση εξετάζουμε, χωρίς να περιλαμβάνεται το υποσύνολο K στο training set του. Τότε, η εκτιμήτρια διασταυρωμένης επικύρωσης για το σφάλμα πρόβλεψης (*cross validation estimate of prediction error*) του μοντέλου αυτού όπως την όρισε ο Hastie et al. (2001) είναι:

$$CV(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n} L\left(y_i, \hat{f}^{-K(i)}(x_i)\right)$$

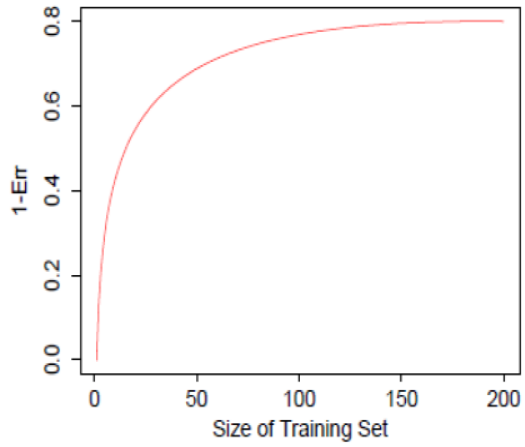
Στην περίπτωση που επιθυμούμαι να συγκρίνουμε διαφορετικά μοντέλα μεταξύ τους, η παραπάνω σχέση μετασχηματίζεται στην:

$$CV(\hat{f}, a) = \frac{1}{n}\sum_{i=1}^{n} L\left(y_i, \hat{f}^{-K(i)}(x_i, a)\right)$$

όπου η παράμετρος **a** αποτελεί μια ρυθμιστική παράμετρο για το κάθε μοντέλο και μεταβάλλεται ανάλογα με την πολυπλοκότητα του.

Η συνάρτηση $CV(\hat{f}, a)$ παρέχει μια εκτίμηση της καμπύλης του σφάλματος πρόβλεψης, γνωστή και ως καμπύλη μάθησης, και σκοπός είναι να βρούμε το $\hat{a}$ που την ελαχιστοποιεί. Το μοντέλο που τελικά επιλέγεται ως το καταλληλότερο για την περιγραφή των δεδομένων είναι το $f(x, \hat{a})$ (Hastie et al., 2001).

Τυπικές επιλογές για την τιμή k μιας k – fold διασταυρωμένης επικύρωσης είναι 5 ή 10. Η περίπτωση όπου $k = n$ είναι γνωστή ως leave – one – out cross validation. Στην leave – one – out cross validation ισχύει ότι $k(i) = i$ και η $\hat{f}^{-K}(x)$ υπολογίζεται αφαιρώντας μόνο την παρατήρηση i. Η εκτιμήτρια του σφάλματος πρόβλεψης που προκύπτει σε αυτή την περίπτωση είναι αμερόληπτη (*unbiased*), αλλά μεγάλης διακύμανσης (*variance*). Επιπλέον, απαιτούνται n εκπαιδεύσεις του μοντέλου, άρα γενικά μιλάμε για μια περίπτωση με μεγάλο υπολογιστικό κόστος (Hastie et al., 2001).



Σχήμα 29: Διάγραμμα απόδοσης (1 – Err ⇔ 1 – σφάλμα πρόβλεψης) σε σχέση με το μέγεθος του συνόλου εκπαίδευσης n (*Size of Training Set*), όπως απεικονίζεται στο σύγγραμμα των Hastie et al. (2001). Δοθέντος ενός συγκεκριμένου μοντέλου ταξινόμησης παρατηρούμε ότι η απόδοση του ταξινομητή βελτιώνεται όσο αυξάνει ο αριθμός των παρατηρήσεων στο training set και φτάνει μέχρι το 100. Ωστόσο, η περαιτέρω αύξηση του n στις 200 παρατηρήσεις δεν αποφέρει κάποιο ιδιαίτερο όφελος στην απόδοση του μοντέλου.

Τέλος, παραθέτουμε έναν ψευδοκώδικα για την k – fold διασταυρωμένη επικύρωση.

Βήμα 1:

Επέλεξε τιμή για το k.

Βήμα 2:

Διέσπασε το σύνολο των δεδομένων σε k, περίπου ισομεγέθη, υποσύνολα.

Βήμα 3:

Για i = 1 εώς i = k επανέλαβε:

a. Επέλεξε k - 1 από τα k υποσύνολα του βήματος 2 και εκπαίδευσε με αυτά το μοντέλο.

b. Χρησιμοποίησε το εναπομείναν υποσύνολο για επικύρωση του μοντέλου και υπολόγισε το σφάλμα πρόβλεψης. ΠΡΟΣΟΧΗ! Σε καμία άλλη από τις επόμενες επαναλήψεις δεν πρέπει να χρησιμοποιηθεί το ίδιο υποσύνολο για επικύρωση! Κάθε υποσύνολο χρησιμοποιείται μια και μοναδική φορά ως validation set.

c. Αύξησε το i κατά 1.

Βήμα 4:

Υπολόγισε τον μέσο όρο των k σφαλμάτων πρόβλεψης που προέκυψαν από το βήμα 3.

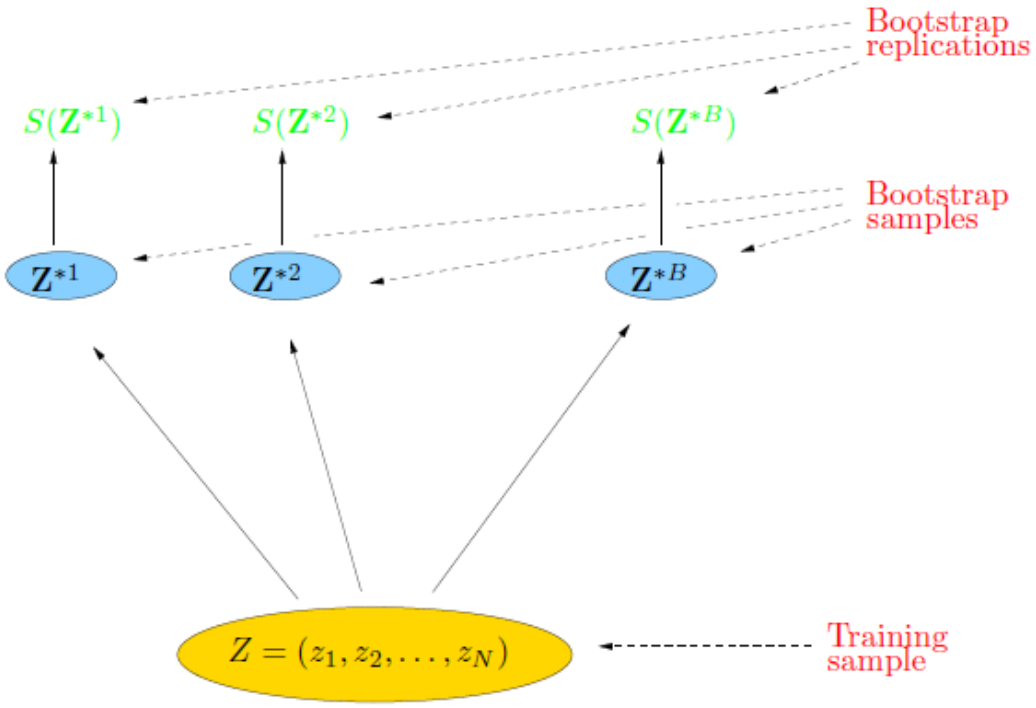
Σχήμα 30: Ψευδοκώδικας για την μέθοδο k – fold cross validation.

4.1.3 Bootstrap Μέθοδοι

Οι bootstrap μέθοδοι είναι ίσως ο πιο βέλτιστος τρόπος για τον υπολογισμό της απόδοσης όταν το σύνολο των δεδομένων είναι πολύ μικρό. Αποτελούν ένα γενικό εργαλείο για την αξιολόγηση της στατιστικής ακρίβειας. Στόχος της μεθόδου είναι ο υπολογισμός του σφάλματος δοκιμής $\mathrm{Err}_{\mathrm{T}}$, όμως αυτό που τελικά εκτιμάται από αυτήν είναι το αναμενόμενο σφάλμα πρόβλεψης Err, όπως συμβαίνει άλλωστε και στις περισσότερες μεθόδους αξιολόγησης μοντέλων (Hastie et al., 2001).

Υποθέτοντας ότι έχουμε εκπαιδεύσει ένα μοντέλο με χρήση ενός training set Z, η βασική ιδέα, σύμφωνα με τον Hastie et al. (2001), είναι να σχεδιάσουμε τυχαία σύνολα δεδομένων μέσω ανταλλαγής μεταξύ των δεδομένων που χρησιμοποιήθηκαν για την εκπαίδευση. Τα τυχαία αυτά σύνολα λέγονται

bootstrap datasets και έχουν το ίδιο μέγεθος με το αρχικό set εκπαίδευσης του μοντέλου. Στη συνέχεια, το μοντέλο επανεκπαιδεύεται κάθε φορά με ένα διαφορετικό bootstrap dataset και εξετάζεται η συμπεριφορά του.



Σχήμα 31: Σχηματική απεικόνιση της μεθόδου bootstrap, όπως παρουσιάζεται στο σύγγραμμα του Hastie et al. (2001). Με ανταλλαγή από το σύνολο εκπαίδευσης (training sample) έχουν παραχθεί B bootstrap datasets $\mathbf{Z}^{*\mathbf{b}}$, με $\mathbf{b} = \mathbf{1}, \mathbf{2}, \ldots, \mathbf{B}$, μεγέθους N, όπου N ο αριθμός των στοιχείων του Z. Στόχος είναι να αξιολογήσουμε την ακρίβεια μιας ποσότητας $\mathbf{S}(\mathbf{Z})$ (οποιαδήποτε ποσότητα υπολογίζεται με χρήση των δεδομένων του συνόλου Z) μέσω των ποσοτήτων $\mathbf{S}(\mathbf{Z}^{*\mathbf{1}}), \mathbf{S}(\mathbf{Z}^{*\mathbf{2}}), \ldots, \mathbf{S}(\mathbf{Z}^{*\mathbf{B}})$, οι οποίες εκτιμήθηκαν από το μοντέλο κατά την εκπαίδευση του με τα σύνολα $\mathbf{Z}^{*\mathbf{1}}, \mathbf{Z}^{*\mathbf{2}}, \ldots, \mathbf{Z}^{*\mathbf{B}}$ αντίστοιχα.

Πως εφαρμόζεται όμως η μέθοδος bootstrap για την εκτίμηση του σφάλματος πρόβλεψης; Μια διαδικασία είναι να εκπαιδεύσουμε το μοντέλο με ένα bootstrap dataset και στην συνέχεια να παρατηρήσουμε πόσο καλά προβλέπει το αρχικό «αυθεντικό» σύνολο εκπαίδευσης, να το χρησιμοποιήσουμε δηλαδή ως set επικύρωσης (Hastie et al., 2001). Αν συμβολίσουμε με $\hat{f}^{*b}(x_i)$ την προβλεπόμενη τιμή στο σημείο x_i από το μοντέλο που εκπαιδεύτηκε με το b – οστό bootstrap dataset, η bootstrap εκτιμήτρια για το σφάλμα πρόβλεψης, όπως την ορίζει ο Hastie et al. (2001) είναι:

$$\widehat{Err}_{boot} = \frac{1}{B}\frac{1}{n}\sum_{b=1}^{B}\sum_{i=1}^{n} L\left(y_i, \hat{f}^{*b}(x_i)\right)$$

όπου $\mathbf{n}$ το μέγεθος του αρχικού «αυθεντικού» συνόλου εκπαίδευσης Z.

Δυστυχώς η παραπάνω έκφραση δεν αποτελεί καλή εκτιμήτρια. Ο λόγος είναι ότι το bootstrap dataset, που χρησιμοποιείται ως training set, και το αρχικό «αυθεντικό» σύνολο εκπαίδευσης Z, που χρησιμοποιείται ως validation set, έχουν κοινά στοιχεία, γεγονός που οδηγεί σε υπερπροσαρμογή (overfitting). Μια καλύτερη εκτίμηση, που ξεπερνά αυτό το πρόβλημα, δίνει η leave – one – out bootstrap μέθοδος, κατά την οποία για κάθε παρατήρηση ενδιαφερόμαστε μόνο για τα bootstrap datasets που δεν την περιέχουν (Hastie et al., 2001). Προκύπτει λοιπόν η σχέση:

$$\widehat{\mathrm{Err}}_{\mathrm{boot}}^{\mathrm{loo}} = \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \frac{1}{|\mathrm{C}^{-\mathrm{i}}|} \sum_{\mathrm{b} \in \mathrm{C}^{-\mathrm{i}}} \mathrm{L}\left(\mathrm{y}_{\mathrm{i}}, \hat{\mathrm{f}}^{*\mathrm{b}}(\mathrm{x}_{\mathrm{i}})\right)$$

όπου με $\mathbf{C^{-i}}$ συμβολίζουμε το bootstrap dataset που δεν περιέχει την i – οστή παρατήρηση και με $\left|\mathbf{C^{-i}}\right|$ το πλήθος των συνόλων αυτών. Το $\mathbf{n}$ όπως και πριν αποτελεί το μέγεθος του αρχικού «αυθεντικού» συνόλου εκπαίδευσης Z.

Για να υπολογίσουμε το $\widehat{\mathrm{Err}}_{\mathrm{boot}}^{\mathrm{loo}}$ πρέπει να επιλέξουμε αρκετά μεγάλο B ώστε ∀i να ισχύει: $\left|\mathrm{C}^{-\mathrm{i}}\right| > 0$. Διαφορετικά, απορρίπτουμε τα i για τα οποία $\left|\mathrm{C}^{-\mathrm{i}}\right| = 0$ και δεν τα συμπεριλαμβάνουμε στον υπολογισμό της εκτίμησης του σφάλματος πρόβλεψης (Hastie et al., 2001).

4.1.4 Ακρίβεια, Πίνακας συνάφειας, Ευαισθησία, Ειδικότητα, Επιπολασμός

4.1.4.1 Ακρίβεια: Accuracy και Precision

Δεδομένου ενός συστήματος μέτρησης ορίζουμε τις παρακάτω έννοιες με βάση το σύγγραμμα των Hastie et al. (2001):

- **Ακρίβεια** (*accuracy*): είναι ο βαθμός εγγύτητας των μετρήσεων μιας ποσότητας σε σχέση με την πραγματική τιμή της ποσότητας αυτής.

- **Ακρίβεια** (*precision*): είναι ο βαθμός στον οποίο επαναλαμβανόμενες μετρήσεις υπό αμετάβλητες συνθήκες δίνουν τα ίδια αποτελέσματα.

- **Ορθότητα** (*trueness*): αφορά την εγγύτητα του μέσου όρου των αποτελεσμάτων των μετρήσεων και της πραγματικής τιμής.

- **Έγκυρο σύστημα μέτρησης**: λέγεται ένα σύστημα όταν είναι τόσο ακριβές όσο και σαφές, δηλαδή είναι ταυτόχρονα και accurate και precise.

Ένα σύστημα μέτρησης μπορεί να είναι ακριβές (*accurate*), αλλά όχι precise και το αντίστροφο. Για παράδειγμα, αν ένα πείραμα περιέχει ένα συστηματικό σφάλμα, τότε αυξάνοντας το μέγεθος του δείγματος αυξάνεται η precision αλλά όχι και η accuracy. Επιπλέον, η εξάλειψη του συστηματικού σφάλματος βελτιώνει την accuracy, όμως η precision παραμένει σταθερή. Στην ιδανική περίπτωση, το σύστημα μέτρησης είναι accurate και precise ταυτοχρόνως, με τις μετρήσεις να είναι κοντά και «σφικτά» συγκεντρωμένες γύρω από την πραγματική τιμή.

Σύμφωνα με το πρότυπο ISO 5725 – 1, οι όροι της ορθότητας και της precision ακρίβειας χρησιμοποιούνται για να περιγράψουν την accuracy ακρίβεια μιας μέτρησης. Μέσω αυτής μπορούμε επίσης να διακρίνουμε την διαφορά μεταξύ του μέσου όρου των μετρήσεων και της πραγματικής τιμής, δηλαδή την μεροληψία (*bias*) του συστήματος μετρήσεων, ο καθορισμός και η διόρθωση της οποίας είναι απαραίτητα για μια σωστή ταξινόμηση.

Η precision ακρίβεια είναι μερικές φορές στρωματοποιημένη σε επαναληψιμότητα και αναπαραγωγικότητα. Συγκεκριμένα:

- Η επαναληψιμότητα είναι η μεταβολή που προκύπτει όταν όλες οι προσπάθειες που καταβάλλονται για να κρατηθούν σταθερές οι συνθήκες, χρησιμοποιώντας το ίδιο όργανο και χειριστή, επαναλαμβάνονται σε σύντομο χρονικό διάστημα.
- Η αναπαραγωγικότητα είναι η μεταβολή που προκύπτει χρησιμοποιώντας την ίδια διαδικασία μέτρησης μεταξύ των διαφόρων μέσων επαναλαμβάνοντας σε μεγαλύτερες χρονικές περιόδους.

4.1.4.2 Πίνακας συνάφειας

Δεδομένου ενός ταξινομητή και ενός παραδείγματος υπάρχουν τέσσερα πιθανά αποτελέσματα:

TP: αν η περίπτωση είναι θετική και έχει ταξινομηθεί και ως θετική, υπολογίζεται τελικά ως μια αληθώς θετική περίπτωση (πχ. όταν ένας ασθενής έχει μια νόσο

και το διαγνωστικό test, που χρησιμοποιείται για την ανίχνευση της, βγαίνει πράγματι θετικό).
FN: αν η περίπτωση είναι θετική και έχει ταξινομηθεί ως αρνητική, υπολογίζεται τελικά ως μια ψευδώς αρνητική περίπτωση (πχ. όταν ένας ασθενής έχει μια νόσο, αλλά το διαγνωστικό test, που χρησιμοποιείται για την ανίχνευση της, βγαίνει αρνητικό).
TN: αν η περίπτωση είναι αρνητική και έχει ταξινομηθεί και ως αρνητική, υπολογίζεται τελικά ως μια αληθώς αρνητική περίπτωση (πχ. όταν ένας ασθενής δεν έχει μια νόσο και το διαγνωστικό test, που χρησιμοποιείται για την ανίχνευση της, βγαίνει πράγματι αρνητικό).
FP: αν η περίπτωση είναι αρνητική και έχει ταξινομηθεί ως θετική, υπολογίζεται τελικά ως μια ψευδώς θετική περίπτωση (πχ. όταν ένας ασθενής δεν έχει μια νόσο, αλλά το διαγνωστικό test, που χρησιμοποιείται για την ανίχνευση της, βγαίνει θετικό).

Οι διατάξεις του συνόλου των περιπτώσεων αντιπροσωπεύονται από έναν πίνακα 2x2, τον λεγόμενο πίνακα συνάφειας ή πίνακα έκτακτης ανάγκης, ο οποίος αποτελεί βάση για πολλές μετρήσεις. Οι αριθμοί κατά μήκος των κύριων διαγωνίων του αντιπροσωπεύουν τις σωστές αποφάσεις, ενώ οι αριθμοί εκτός αυτής αντιπροσωπεύουν τα λάθη – τη σύγχυση – μεταξύ των διάφορων κατηγοριών.

Ορίζουμε λοιπόν για έναν ταξινομητή επιπλέον τις παρακάτω έννοιες:

- **Αληθώς Θετικό Ποσοστό** (*True Positive Rate* → TPR) ή Ποσοστό Επιτυχίας και Ανάκλασης

$$\text{TPR} = \frac{\text{αληθώς θετικά}}{\text{σύνολο θετικών}} = \frac{\text{TP}}{\text{P}} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- **Ψευδώς Θετικό Ποσοστό** (*False Positive Rate* → FPR):

$$\text{FPR} = \frac{\text{ψευδώς θετικά}}{\text{σύνολο αρνητικών}} = \frac{\text{FP}}{\text{N}} = \frac{\text{FP}}{\text{TN} + \text{FP}}$$

Μέσω των TP, TN, FN και FP μπορούμε τώρα να ορίσουμε μαθηματικά την accuracy και precision ακρίβεια, καθώς και ένα ακόμη μέτρο για την εκτίμηση

του ποσοστού επιτυχίας μιας ταξινόμησης, το οποίο ορίζει στο paper του ο Fishel et al. (2007).

- **Ακρίβεια** (*accuracy*): είναι η αναλογία των πραγματικών αποτελεσμάτων (αληθώς θετικά (TP) και αληθώς αρνητικά (TN)) στον πληθυσμό.

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

100% accuracy σημαίνει ότι οι τιμές που εκτιμούνται από τον ταξινομητή είναι ακριβώς ίδιες με τις αληθείς τιμές.

- **Ακρίβεια** (*precision*): είναι το ποσοστό των αληθώς θετικών έναντι όλων των θετικών αποτελεσμάτων (τόσο αληθώς θετικά όσο και ψευδώς θετικά).

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

- **Ποσοστό Επιτυχίας Ταξινόμησης** (classification success rate → CSR) (Fishel et al., 2007): είναι ο σταθμισμένος μέσος των αληθώς θετικών (TP) και των αληθώς αρνητικών (TN).

$$\text{CSR} = \frac{1}{2}\left(\frac{\text{TP}}{\text{TP} + \text{FP}} + \frac{\text{TN}}{\text{TN} + \text{FN}}\right)$$

4.1.4.3 Ευαισθησία (sensitivity) και Ειδικότητα (specificity)

Η ευαισθησία και η ειδικότητα είναι στατιστικά μέτρα της απόδοσης ενός test δυαδικής ταξινόμησης και είναι γνωστές στη στατιστική ως συναρτήσεις ταξινόμησης. Αποτελούν τις συχνότερα χρησιμοποιούμενες συνιστώσες διαγνωστικής ποιότητας μιας δοκιμασίας, καθώς καθορίζουν την διακριτική της ικανότητα. Επίσης, συνδέονται στενά με τις έννοιες των σφαλμάτων τύπου

I [9] και τύπου II [10] των στατιστικών ελέγχων υποθέσεων. Ένας τέλειος εκτιμητής – ένα καλό μοντέλο ταξινόμησης οφείλει να έχει 100% ευαισθησία και 100% ειδικότητα.

- **Ευαισθησία** (*sensitivity*) ή Ποσοστό των Αληθώς Θετικών Αποτελεσμάτων (TPR): είναι το ποσοστό των θετικών ενδείξεων στον πληθυσμό της δοκιμασίας, δηλαδή η πιθανότητα το test να είναι θετικό δεδομένου ότι κάποιος έχει το χαρακτηριστικό (πχ. την νόσο) που εξετάζουμε.

$$\mathrm{SE} = \mathrm{TPR} = \frac{\text{αληθώς θετικά}}{\text{σύνολο θετικών}} = \frac{\mathrm{TP}}{\mathrm{P}} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

Η ευαισθησία σχετίζεται με την ικανότητα του test να προσδιορίζει θετικά αποτελέσματα. Μια δοκιμή με υψηλή ευαισθησία έχει χαμηλό ποσοστό σφάλματος τύπου II.

- **Ειδικότητα** (*specificity*) ή Ποσοστό Αληθώς Αρνητικών Αποτελεσμάτων (TNR): είναι το ποσοστό των αρνητικών ενδείξεων στον πληθυσμό, δηλαδή η πιθανότητα το test να είναι αρνητικό δεδομένου ότι κάποιος δεν έχει το χαρακτηριστικό (πχ. την νόσο) που εξετάζουμε.

$$\mathrm{SPC} = \mathrm{TNR} = \frac{\text{αληθώς αρνητικά}}{\text{σύνολο αρνητικών}} = \frac{\mathrm{TN}}{\mathrm{N}} = \frac{\mathrm{TN}}{\mathrm{FP} + \mathrm{TN}}$$

Η ειδικότητα σχετίζεται με την ικανότητα του test να εντοπίζει αρνητικά αποτελέσματα. Μια δοκιμή με υψηλή εξειδίκευση έχει χαμηλό ποσοστό σφάλματος τύπου I.

Οι τιμές της ευαισθησίας και της ειδικότητας, δηλαδή τα ποσοστά TPR και TNR, καθώς και τα συμπληρωματικά τους (ποσοστό ψευδώς αρνητικών (FNR) και ψευδώς θετικών αποτελεσμάτων (FPR) αντίστοιχα) ονομάζονται πιθανοφάνειες (*likelihoods*) ή αλλιώς λειτουργικά χαρακτηριστικά (*operating characteristics*) της διαγνωστικής δοκιμασίας. Ισχύει η παρακάτω σχέση:

---

[9] **Σφάλμα τύπου I:** όταν σε ένα στατιστικό έλεγχο υποθέσεων η H_0 (αρχική υπόθεση) απορρίπτεται, με μια πιθανότητα έστω α, ενώ είναι αληθής.

[10] **Σφάλμα τύπου II:** όταν σε ένα στατιστικό έλεγχο υποθέσεων η H_0 (αρχική υπόθεση) γίνεται δεκτή, με μια πιθανότητα έστω β, ενώ η H_1 (εναλλακτική υπόθεση) είναι αληθής.

$$\mathrm{TPR} = 1 - \mathrm{FNR}$$

όπου

$$\mathrm{FNR} = \frac{\mathrm{FN}}{\mathrm{P}} = \frac{\mathrm{FN}}{\mathrm{TP} + \mathrm{FN}}$$

4.1.4.4 Επιπολασμός (prevalence)

Προτού ορίσουμε το τι είναι επιπολασμός, εισάγουμε τις παρακάτω έννοιες:

- **Θετική Προγνωστική Αξία** (*positive predictive value* → PPV): εκφράζει την πιθανότητα εμφάνισης θετικού περιστατικού μεταξύ όλων των θετικών προβλέψεων. Για παράδειγμα, την πιθανότητα κάποιος να είναι όντως ασθενής όταν ο διαγνωστικός έλεγχος έχει βρεθεί θετικός.

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

- **Αρνητική Προγνωστική Αξία** (*negative predictive value* → NPV): εκφράζει την πιθανότητα εμφάνισης αρνητικού περιστατικού μεταξύ όλων των αρνητικών προβλέψεων. Για παράδειγμα, την πιθανότητα κάποιος όντως να μην είναι ασθενής όταν ο διαγνωστικός έλεγχος έχει βρεθεί αρνητικός.

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

Όταν και οι δύο προγνωστικές αξίες είναι υψηλές και κοντά στο 1 (ή στο 100% αν μιλάμε για ποσοστά), ο διαγνωστικός έλεγχος θεωρείται καλός. Στην πράξη όμως τυχαίνει να είναι υψηλή μια μόνο από τις προγνωστικές αξίες.

- **Επιπολασμός** (*prevalence* → PRV): είναι το σύνολο των θετικών περιστατικών προς το σύνολο ολόκληρου του πληθυσμού.

$$\mathrm{PRV} = \frac{\mathrm{TP} + \mathrm{FN}}{\mathrm{P} + \mathrm{N}}$$

Οι διαγνωστικές έννοιες PPV και NPV λειτουργούν συμπληρωματικά με τον επιπολασμό, ο οποίος εκφράζει την πιθανότητα προ δοκιμασίας (*pretest probability*).

Τέλος, εφόσον είναι γνωστός ο επιπολασμός, μπορούμε μέσω της ευαισθησίας και της ειδικότητας να προσδιορίσουμε την accuracy ακρίβεια. Συγκεκριμένα:

$$\mathrm{accuracy} = (\mathrm{sensitivity})(\mathrm{prevalence}) + (\mathrm{specificity})(1 - \mathrm{prevalence})$$

Συνοψίζοντας όλα τα παραπάνω μέτρα προκύπτει ο ακόλουθος πίνακας συνάφειας (Πίνακας 4).

Πίνακας 4: Πίνακας συνάφειας και μέτρα.

		ΑΠΟΤΕΛΕΣΜΑ ΤΟΥ ΤΕΣΤ ΠΡΟΒΛΕΨΗΣ			
		Θετικό (**P**)	Αρνητικό (**N**)	Επιπολασμός (**PRV**)	
ΠΡΑΓΜΑΤΙΚΗ ΤΙΜΗ	Αληθές (**T**)	Αληθώς Θετικό (**TP**)	Ψευδώς Αρνητικό **(TN)**	Ευαισθησία (**TPR** / sensitivity)	Ποσοστό Ψευδώς Αρνητικών (**FNR**)
	Ψευδές (**F**)	Ψευδώς Θετικό **(FP)**	Αληθώς Αρνητικό **(FN)**	Ειδικότητα (**TNR** / specificity)	Ποσοστό Ψευδώς Θετικών (**FPR**)
	Ακρίβεια (accuracy)	Θετική Προγνωστική Αξία (**PPV** / precision)	Αρνητική Προγνωστική Αξία (**NPV**)		

4.2 Αξιολόγηση μετα–ανάλυσης

Όπως αναφέραμε και στο κεφάλαιο 1, κάποιες επιπλέον διαδικασίες, όπως η subgroup analysis, η sensitivity analysis και η μελέτη της publication bias, είναι απαραίτητο να πραγματοποιηθούν από τον συγγραφέα μιας μετα–ανάλυσης, ώστε να μπορεί μετά μέσα από τα αποτελέσματα αυτών να αξιολογηθεί και η συνολική αξιοπιστία της. Επιπλέον, μέτρα όπως η precision ακρίβεια, και η

ανάκτηση (*recall*), μπορούν να χρησιμοποιηθούν από τον μετα–αναλυτή για την αξιολόγηση της επιτυχίας της βιβλιογραφικής έρευνας, και κατ' επέκταση της μετα–ανάλυσης.

4.2.1 Ακρίβεια (precision) και ανάκτηση (recall)

Όπως αναφέρεται στην μελέτη της Normand (1998), αν και ο μετα–αναλυτής επιθυμεί να πραγματοποιήσει μια όσο το δυνατόν πιο ολοκληρωμένη έρευνα, δεν είναι προφανώς εφικτό να εξασφαλίσει κάθε κομμάτι της βιβλιογραφίας που σχετίζεται με το αντικείμενο ενδιαφέροντος. Για να αξιολογηθεί λοιπόν η επιτυχία της έρευνας, η οποία αποτελεί πολύ σημαντικό κομμάτι της μετα–ανάλυσης, μπορούν να χρησιμοποιηθούν δυο μέτρα: η ακρίβεια (*precision*) και η ανάκτηση (*recall*). Όπως ορίζει στο paper της η Normand (1998), η ακρίβεια και η ανάκτηση δίνονται από τις σχέσεις:

$$\text{recall} = \frac{\text{Number of relevant documents retrieved}}{\text{Total number that should be retrieved}} \times 100$$

$$\text{precision} = \frac{\text{Number retrieved and relevant}}{\text{Number retrieved}} \times 100$$

Παρατηρούμε λοιπόν ότι η ανάκτηση εκφράζει την αναλογία μεταξύ του αριθμού των μελετών που ανακτήθηκαν σε σχέση με τον αριθμό των μελετών που θα έπρεπε να έχουν ανακτηθεί κατά την βιβλιογραφική έρευνα, ενώ η ακρίβεια, όπως προαναφέραμε και στην παράγραφο 4.2, είναι ο λόγος του αριθμού των μελετών που ανακτήθηκαν και ήταν σχετικές με το αντικείμενο της έρευνας προς το σύνολο των μελετών που ανακτήθηκαν κατά την βιβλιογραφική έρευνα.

Πρέπει να προσέξουμε εδώ ότι ο παρανομαστής στη σχέση που δίνει το recall είναι ουσιαστικά άγνωστος, καθώς ο μετα–αναλυτής δεν μπορεί να γνωρίζει από πριν αν το σύνολο των μελετών που έχει ανακτήσει είναι αντιπροσωπευτικό όλων των διαθέσιμων μελετών σχετικών με το αντικείμενο της έρευνας. Για την αντιμετώπιση αυτού του προβλήματος προτείνεται η χρήση μεθόδων (πχ. "capture – recapture" → βλ. κεφάλαιο 1, βήμα 3) για την εκτίμηση του μεγέθους του πληθυσμού, δηλαδή του αριθμού των διαθέσιμων μελετών πάνω σε ένα συγκεκριμένο ερευνητικό θέμα (Normand, 1998).

4.2.2 Ανάλυση υποσυνόλων (subgroup analysis)

Όπως αναφέραμε και στο κεφάλαιο 1, η ανάλυση αυτή ρίχνει φως στις αιτίες ύπαρξης ετερογένειας (Whitehead & Whitehead, 1991), η σωστή αντιμετώπιση και αξιολόγηση της οποίας προσθέτει μεγάλη αξιοπιστία στη μετα–ανάλυση. Οι μελέτες χωρίζονται σε υποσύνολα με όσο το δυνατόν πιο ομοιογενή αποτελέσματα βάσει των χαρακτηριστικών τους, όπως για παράδειγμα τον σχεδιασμό τους, την χώρα προέλευσης τους, τον τρόπο διεξαγωγής τους, τα χαρακτηριστικά των ατόμων που συμμετέχουν σε αυτές κ.λ.π, ώστε να υπολογιστεί ένα συνολικό μέτρο επίδρασης για κάθε υποσύνολο ξεχωριστά. Με τον τρόπο αυτό εντοπίζονται επιδράσεις που μπορεί να μην ήταν ορατές ή σαφείς στην ανάλυση που περιείχε όλες τις μελέτες.

4.2.3 Ανάλυση ευαισθησίας (sensitivity analysis)

Όπως αναφέραμε και στο κεφάλαιο 1, η sensitivity analysis αποτιμά την “ευαισθησία” των αποτελεσμάτων μιας μετα–ανάλυσης. Χρησιμεύει, δηλαδή, στην εκτίμηση της σταθερότητας των εκτιμώμενων μέτρων αποτελέσματος όταν προκύπτουν αλλαγές στα δεδομένα (Normand, 1998). Έστω ότι στην μετα-ανάλυση συμπεριλαμβάνονται n μελέτες. Η ανάλυση ευαισθησίας πραγματοποιείται αφαιρώντας κάθε φορά μια διαφορετική μελέτη από την μετα-ανάλυση που περιέχει και τις n μελέτες και επαναλαμβάνοντας την με τις εναπομείναντες n-1 μελέτες. Πραγματοποιούμαι δηλαδή n επιπλέον αναλύσεις. Στη συνέχεια συγκρίνονται τα αποτελέσματα των αναλύσεων αυτών μεταξύ τους, καθώς και με τα αποτελέσματα της μετα–ανάλυσης των n μελετών και παρατηρείται αν η αφαίρεση κάποιας από αυτές προκάλεσε στατιστικά σημαντικές μεταβολές στα αποτελέσματα της αρχικής μετα–ανάλυσης (των n μελετών) (Normand, 1998).

4.2.4 Σφάλμα δημοσίευσης (publication bias)

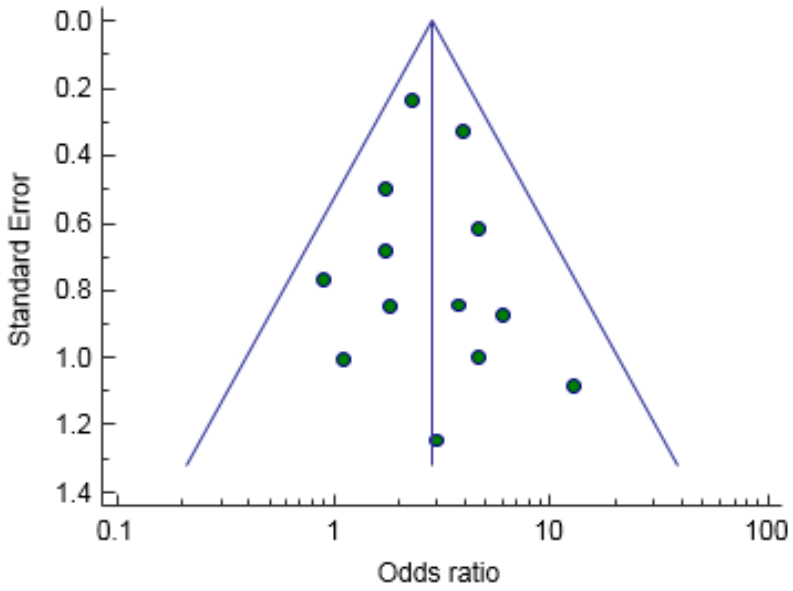
Η μελέτη της ύπαρξης ή μη publication bias σε μια μελέτη, καθώς και η αντιμετώπιση της αποτελεί πολύ σημαντικό βήμα στην σύνταξη μιας μετα–ανάλυσης και αυτό διότι το ποσό του σφάλματος δημοσίευσης καθορίζει σημαντικά την αξιοπιστία του αποτελέσματος μιας δημοσίευσης.

Το σφάλμα δημοσίευσης, με βάση τους Τσιάρα et al. (2011) και Gioacchino (2005), προκύπτει εξαιτίας του ότι:

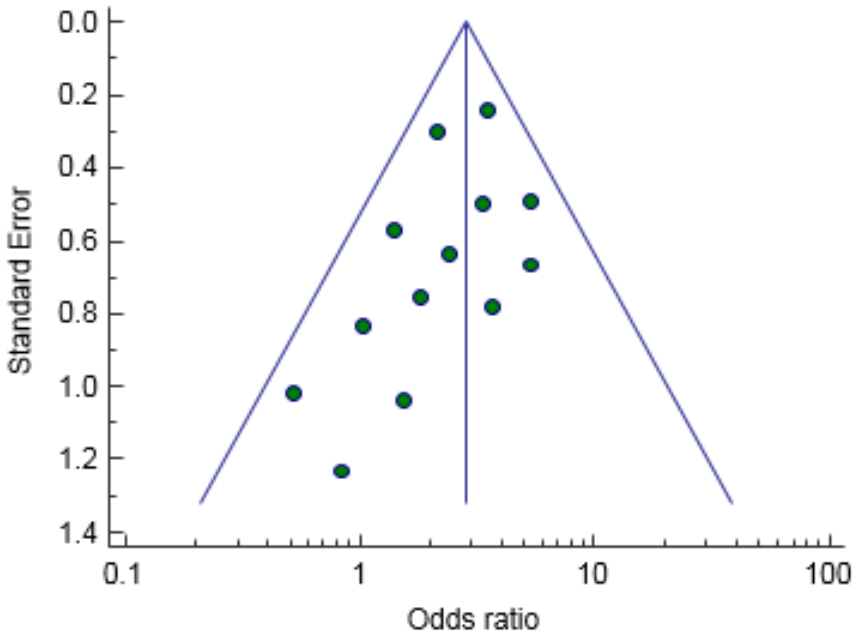
- ✓ Τα περισσότερα περιοδικά προτιμούν να δημοσιεύουν στατιστικά σημαντικά αποτελέσματα απ' ότι μη στατιστικά σημαντικά. Οπότε και οι συγγραφείς έχουν την τάση να υποβάλλουν προς δημοσίευση τέτοιου είδους μελέτες.
- ✓ Οι συγγραφείς όταν πραγματοποιούν την μετα–ανάλυση τους στηρίζονται μόνο σε δημοσιευμένες μελέτες, οι οποίες μπορεί να μην είναι αντιπροσωπευτικές του συνόλου της έρευνας στο μελετώμενο αντικείμενο.
- ✓ Τα περιοδικά πολλές φορές δεν αποδέχονται αποτελέσματα, τα οποία δεν είναι σύμφωνα με τα ήδη δημοσιευμένα, ή αντίστροφα αποδέχονται τέτοιου είδους αποτελέσματα, διότι τα θεωρούν καινοτόμα και όχι πλεονάζοντα.
- ✓ Τα περιοδικά τις περισσότερες φορές δεν δέχονται να δημοσιεύσουν αρνητικά αποτελέσματα (πχ. ότι ένα νέο φάρμακο δεν είναι εν τέλει τόσο αποτελεσματικό συγκρινόμενο με ένα παλιό) και κατ' επέκταση οι συγγραφείς προτιμούν να υποβάλουν προς δημοσίευση μελέτες που εμφανίζουν θετικά αποτελέσματα.

Διάγραμμα χωνί

Για να διερευνηθεί η ύπαρξη σφάλματος δημοσίευσης χρησιμοποιείται σχεδόν πάντα το διάγραμμα " χωνί " (*funnel plot*). Το funnel plot είναι ένα διάγραμμα διασποράς (*scatter plot*) με οριζόντιο άξονα τα εκτιμώμενα μέτρα αποτελέσματος (*estimated effect measures*) της κάθε μελέτης της μετα–ανάλυσης (πχ. τα OR) και κατακόρυφο άξονα το μέγεθος του δείγματος (*sample size*) ή κάποιο άλλο μέτρο ακρίβειας της κάθε μελέτης, πχ. το τυπικό σφάλμα του εκτιμώμενου μέτρου αποτελέσματος (*standard error*) (Normand, 1998). Στην περίπτωση που δεν υπάρχει σφάλμα δημοσίευσης στη μετα–ανάλυση, το διάγραμμα έχει σχήμα ανεστραμμένου χωνιού, όπως φαίνεται στο Σχήμα 32. Τότε οι μελέτες με μεγαλύτερο μέγεθος δείγματος, οι οποίες προσεγγίζουν με μεγαλύτερη ακρίβεια τον εκτιμητή του συνολικού αποτελέσματος, βρίσκονται στην κορυφή του ανεστραμμένου χωνιού, με τις τιμές του εκτιμώμενου μέτρου αποτελέσματος να μην παρουσιάζουν μεγάλη διασπορά, ενώ οι μελέτες με μικρό αριθμό συμμετεχόντων βρίσκονται στο κάτω μέρος του ανεστραμμένου χωνιού, με τις τιμές του εκτιμώμενου μέτρου αποτελέσματος να παρουσιάζουν αρκετά μεγαλύτερη διασπορά (Γαλάνης, 2009). Αντίθετα, όταν στο διάγραμμα "χωνί'' υπάρχει ασυμμετρία συμπεραίνουμε ότι υπάρχει σφάλμα δημοσίευσης στη μετα–ανάλυση, όπως φαίνεται στο Σχήμα 33. Επειδή η εκτίμηση που κάνουμε μέσω του funnel plot είναι καθαρά οπτική, η χρήση του συνίσταται στην περίπτωση που η μετα–ανάλυση περιέχει μεγάλο αριθμό μελετών (Gioacchino, 2005).



Σχήμα 32: Διάγραμμα " χωνί " στην περίπτωση απουσίας σφάλματος δημοσίευσης (βλ. ιστότοπο MedCalc).



Σχήμα 33: Διάγραμμα " χωνί " στην περίπτωση απουσίας σφάλματος δημοσίευσης (βλ. ιστότοπο MedCalc).

Μέθοδος του Klein

Αυτή η μέθοδος δεν απαντά ακριβώς στην ερώτηση αν υπάρχει ή όχι σφάλμα δημοσίευσης. Αντιθέτως θεωρείται ότι αποτελεί ένα test αξιοπιστίας της μετα–ανάλυσης, η οποία αντιμετωπίζει προβλήματα λόγω της ύπαρξης publication

bias (Gioacchino, 2005). Πιο αναλυτικά: υποθέτοντας ότι οι αδημοσίευτες μελέτες έχουν τα ίδια χαρακτηριστικά μεταξύ τους, υπολογίζουμε τον αριθμό AP των αδημοσίευτων μελετών, με αρνητικά ή μη στατιστικά σημαντικά αποτελέσματα, που χρειάζονται για να επηρεάσουν τα αποτελέσματα της μετα – ανάλυσης.

$$\mathrm{AP} = \left(\frac{\mathrm{k} \cdot \ln \mathrm{OR}}{1{,}96}\right)^2 \cdot \bar{\mathrm{w}} - \mathrm{k}$$

όπου $\bar{\mathrm{w}}$ η μέση τιμή των βαρών ($\mathrm{w_i} = \frac{1}{{\mathrm{s_i}}^2}$,με $\mathrm{i} = 1, 2, \ldots, \mathrm{k}$) k μελετών.

ΚΕΦΑΛΑΙΟ 5: ΣΥΝΔΙΑΣΜΟΣ ΜΕΤΑ–ΑΝΑΛΥΣΗΣ ΚΑΙ SVM ΣΕ ΔΕΔΟΜΕΝΑ ΓΟΝΙΔΙΑΚΗΣ ΕΚΦΡΑΣΗΣ

5.1 Βασικές έννοιες

Προτού αναφερθούμε αναλυτικά στο θέμα αυτού του κεφαλαίου, θα ορίσουμε και θα εξηγήσουμε κάποιες βασικές έννοιες, η γνώση των οποίων είναι απαραίτητη για την κατανόηση των παρακάτω παραγράφων.

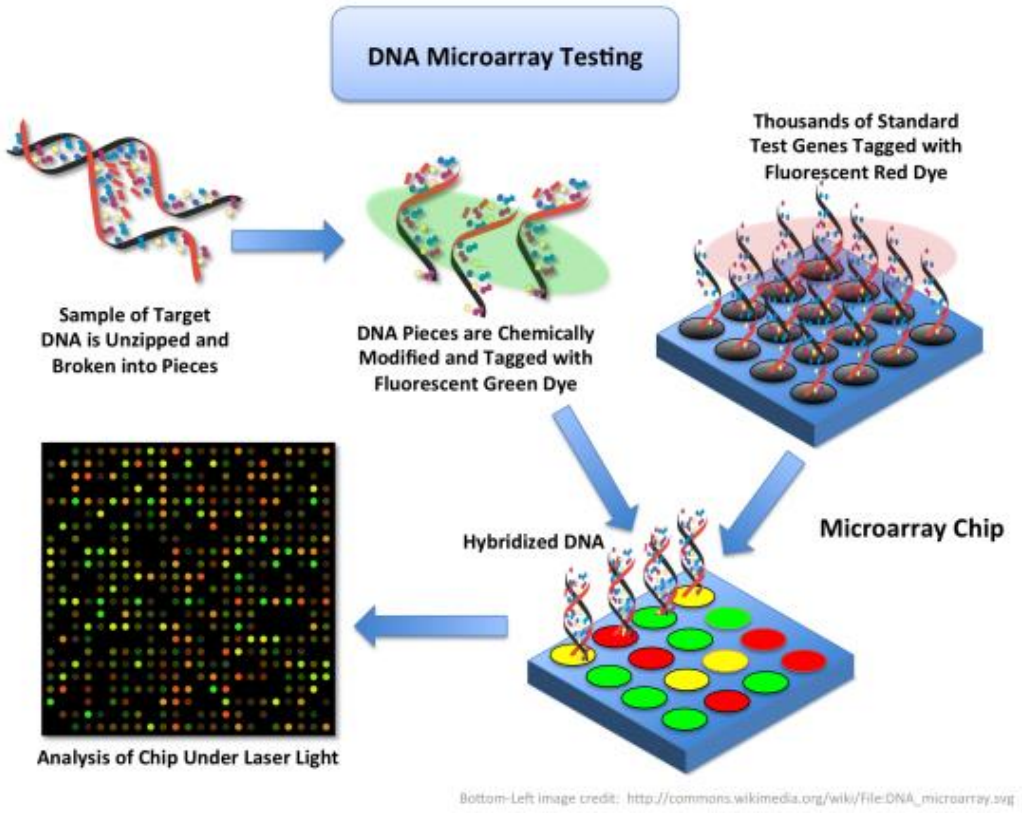
- **Γονίδιο (*gene*):** ένα κληρονομικό τμήμα DNA, το οποίο είναι απαραίτητο για την παραγωγή ενός λειτουργικού προϊόντος (*functional product*) (RNA ή πρωτεΐνη). Το ανθρώπινο γονιδίωμα (το σύνολο του γενετικού υλικού που φέρεται από έναν άνθρωπο) περιέχει περίπου 25.000 γονίδια, που δουλεύουν για την παραγωγή περίπου 1.000.000 ξεχωριστών πρωτεϊνών.

- **Υβριδισμός (*hybridization*):** είναι το φαινόμενο – η χημική διαδικασία κατά την οποία ένα μονόκλωνο μόριο DNA ή RNA αναδιατάσσεται σε συμπληρωματικό DNA ή RNA. Πιο συγκεκριμένα, αν και μια ακολουθία δίκλωνου DNA είναι γενικά σταθερή κάτω από φυσιολογικές συνθήκες, αν αλλάξουν εργαστηριακά αυτές οι συνθήκες (αυξάνοντας ουσιαστικά την θερμοκρασία του περιβάλλοντος) προκαλείται διαχωρισμός των μορίων σε μονούς κλώνους, οι οποίοι είναι συμπληρωματικοί μεταξύ τους, αλλά και με άλλες ακολουθίες του περιβάλλοντος τους. Έτσι, χαμηλώνοντας αυτή τη φορά την θερμοκρασία του περιβάλλοντος επιτρέπεται στα μονόκλωνα μόρια να αναδιαταχθούν, δηλαδή να «υβριδιστούν» μεταξύ τους.

- **Ανιχνευτές (*probes*):** ανιχνεύουν το ποσό mRNA ενός συγκεκριμένου γονιδίου. Γενικά, είναι ένας όρος για την περιγραφή ενός αντιδραστηρίου (ουσία που εμπλέκεται σε μια χημική αντίδραση), το οποίο χρησιμοποιείται για μια μοναδική μέτρηση σε ένα πείραμα γονιδιακής έκφρασης.

- **DNA microarray:** ένα DNA microarray ή DNA τσιπ (*chip*) ή βιοτσίπ (*biochip*) είναι μια συλλογή, ένα array, από χιλιάδες μικροσκοπικές κουκίδες DNA (*DNA spots*), οι οποίες καλούνται χαρακτηριστικά (*features*) και είναι

προσκολλημένες σε μια στερεή επιφάνεια. Κάθε τέτοια κουκίδα περιέχει 10^{-12} γραμμομόρια (*picomoles*) μιας συγκεκριμένης ακολουθίας DNA, γνωστά και ως ανιχνευτές (*probes*). Αυτοί οι ανιχνευτές μπορεί να είναι είτε ένα μικρό τμήμα ενός συγκεκριμένου γονιδίου, είτε άλλα στοιχεία DNA, και χρησιμοποιούνται για τον κάτω από υψηλή ένταση υβριδισμό ενός cDNA ή cRNA δείγματος, αποκαλούμενο και ως στόχο (*target*). Σε κάθε κουκίδα, αν ο υβριδισμός ανιχνευτή και δείγματος γίνει επιτυχώς, η κουκίδα θα φαίνεται να έχει το χρώμα κίτρινο υπό τον φωτισμό με λέιζερ. Αντίθετα, στην περίπτωση που το δείγμα φέρει κάποια μετάλλαξη, ο υβριδισμός δεν θα είναι επιτυχής σε όλα τα σημεία και άρα η κουκίδα θα φέρει πράσινο ή κόκκινο χρώμα υπό τον φωτισμό ενός λέιζερ. Στο παρακάτω σχήμα (Σχήμα 34) φαίνονται περιληπτικά οι διαδικασίες ενός πειράματος, κατά το οποίο χρησιμοποιείται ένα DNA microarray.



Σχήμα 34: Διαδικασίες κατά την πραγματοποίηση ενός DNA microarray πειράματος (βλ. ιστότοπο Gene Factor).

Τέλος, τα DNA microarrays έχουν εφαρμογές:

i. Στην αναγνώριση νέων γονιδίων και στην μελέτη της λειτουργίας αυτών και των επιπέδων έκφρασης τους κάτω από διάφορες συνθήκες.

ii. Στην διάγνωση ασθενειών, όπως η καρδιοπάθεια, οι ψυχικές διαταραχές, τα λοιμώδη νοσήματα και πάνω απ' όλα ο καρκίνος.
iii. Στην φαρμακογονιδιωματική (*pharmacogenomics*), όπου μελετάται η συσχέτιση μεταξύ των αποκρίσεων διαφόρων θεραπειών και των γενετικών προφίλ των ασθενών. Συγκεκριμένα, η συγκριτική ανάλυση των γονιδίων από νοσούντα και φυσιολογικά κύτταρα βοηθά στην αναγνώριση των βιοχημικών συστάσεων των πρωτεϊνών, οι οποίες συντίθενται από τα νοσούντα κύτταρα. Αυτή η πληροφορία είναι πολύ χρήσιμη για την κατασκευή φαρμάκων, όπου θα μάχονται αυτές τις πρωτεΐνες με στόχο να μειώσουν την επιρροή τους.
iv. Στην τοξινογονιδιωματική (*toxinogenomics*), όπου διαπιστώνονται συσχετίσεις μεταξύ των αποκρίσεων τοξικών ουσιών και των αλλαγών στα γενετικά προφίλ των κυττάρων, τα οποία εκτίθενται σε αυτές. Εξερευνούνται δηλαδή οι επιπτώσεις των τοξινών στα κύτταρα και στους απογόνους τους.

- **Microarray technology:** αναφέρεται ουσιαστικά στην χρήση των DNA microarrays. Είναι μια αποδοτική τεχνολογία ευρέως χρησιμοποιούμενη στην βιοϊατρική, η οποία δίνει την δυνατότητα στους ερευνητές να παρακολουθήσουν τα επίπεδα των εκφράσεων πολλών γονιδίων ταυτοχρόνως (Hongfang et al., 2014 και Fishel et al., 2007). Στόχος στη χρήση DNA microarrays είναι η διερεύνηση και η μελέτη των γενετικών αιτιών των ανωμαλιών που συμβαίνουν στις λειτουργίες του ανθρώπινου οργανισμού, αλλά και η αντιμετώπιση πειραματικών διαταραχών (Hongfang et al., 2014).

- **Γονιδιακή έκφραση (*gene expression*):** είναι η διαδικασία κατά την οποία πληροφορίες από ένα γονίδιο χρησιμοποιούνται για την σύνθεση ενός λειτουργικού προϊόντος (*functional product*). Στην γενετική, η γονιδιακή έκφραση είναι το θεμελιώδες επίπεδο στο οποίο ο γονότυπος (*genotype*) (το σύνολο των γονιδίων ενός οργανισμού) προκαλεί τον φαινότυπο (*phenotype*) (τα παρατηρήσιμα χαρακτηριστικά). Η ποσοτικοποίηση αυτού του επιπέδου αποτελεί ουσιαστικά και την μέτρηση της γονιδιακής έκφρασης. Μέσω αυτής μπορούν να αναγνωριστούν ιογενείς μολύνσεις στα κύτταρα, να προσδιοριστεί η προδιάθεση κάποιου στην εμφάνιση καρκίνου κ.α. Ιδανικά η μέτρηση της γονιδιακής έκφρασης ολοκληρώνεται με την ανίχνευση του τελικού λειτουργικού προϊόντος, που για τα περισσότερα γονίδια είναι η πρωτεΐνη. Συχνά όμως είναι ευκολότερη η ανίχνευση ενός προδρόμου mRNA του προϊόντος αυτού, μέσω της μέτρησης του οποίου συμπεραίνονται τα επίπεδα γονιδιακής έκφρασης (*gene expression levels*). Υπάρχουν πολλές μέθοδοι για την ποσοτικοποίηση των επιπέδων, όπως η northen blotting, η RT – qPCR (*Reverse transcription followed by quantitative PCR*), η SAGE (*Serial analysis of gene expression*) και άλλες.

- **Επίπεδα γονιδιακής έκφρασης (*gene expression level*):** αποτελούν το ποσό του mRNA στο γονιδιακό δείγμα.

- **Προφίλ γονιδιακής έκφρασης (*gene expression profile*):** συγκεντρωτική πληροφορία σχετικά με όλα τα mRNAs, που παράγονται στα ποικίλα κύτταρα. Ένα προφίλ γονιδιακής έκφρασης μπορεί να χρησιμοποιηθεί για την διάγνωση μιας ασθένειας ή για την παρατήρηση του πως αντιδρά το σώμα ενός ασθενή σε μια θεραπεία.

- **Υπογραφή γονιδίου (*gene signature*):** ένα σύνολο γονιδίων ενός κυττάρου, των οποίων ο συνδυασμός των εκφράσεων αποτελεί μοναδικό χαρακτηριστικό ενός βιολογικού φαινοτύπου ή μιας ιατρικής κατάστασης.

5.2 Εισαγωγή στο πρόβλημα της χαμηλής επικάλυψης (overlap) μεταξύ των λιστών γονιδίων

Ας υποθέσουμε ότι ένας ερευνητής επιθυμεί να μελετήσει τα προφίλ κάποιων γονιδιακών εκφράσεων (*gene expression profiles*) με στόχο την πρόγνωση ή την διάγνωση του καρκίνου ή ακόμα την πρόβλεψη του αποτελέσματος μιας θεραπείας για την αντιμετώπιση αυτού (Fishel et al., 2007). Καθώς δεν φέρουν όλα τα γονίδια το ίδιο ποσό πληροφορίας για ένα θέμα, ο ερευνητής για να πραγματοποιήσει τις προβλέψεις του πρέπει πρώτα να ξεχωρίσει το βέλτιστο υποσύνολο των γονιδίων, τα οποία παρέχουν την περισσότερη πληροφορία σχετικά με το θέμα που εξετάζει και κατά συνέπεια συμβάλουν, αν χρειαστεί, στην καλύτερη και σωστότερη ταξινόμηση ενός καινούριου δείγματος DNA (πχ. το καινούριο δείγμα ανήκει στους φυσιολογικούς ή στους καρκινικούς ιστούς) (Fishel et al., 2007). Ο ερευνητής κάνει δηλαδή μια gene selection. Ουσιαστικά η gene selection γίνεται μέσω της βαθμονόμησης των γονιδίων σύμφωνα με κάποιο μέτρο σημαντικότητας (*importance measure*) και στη συνέχεια διαλέγοντας τα υψηλότερα βαθμονομημένα γονίδια (*top – ranked genes*), τα λεγόμενα και differentially expressed genes, για περαιτέρω ανάλυση (Golub et al., 1999 και Guyon et al., 2002). Με αυτόν τον τρόπο αρκετές μελέτες κατάφεραν με χρήση της microarray technology να παράγουν σετ υψηλόβαθμων γονιδίων (*differentially expressed gene sets*), των οποίων τα προφίλ έκφρασης κάνουν επιτυχή πρόγνωση ή διάγνωση του καρκίνου ή προβλέπουν αξιόπιστα το αποτέλεσμα μιας θεραπείας του. Παρ' όλα αυτά, το θέμα που απασχολεί τους ερευνητές είναι ότι η επικάλυψη (*overlap*) μεταξύ

αυτών των σετ γονιδίων των διαφόρων μελετών είναι σχεδόν μηδενική (Ein-Dor et al., 2005). Για παράδειγμα, στις μελέτες των van't Veer et al. (2002), Sorlie et al. (2001) και Ramaswamy et al. (2003) εξεταζόταν το χρονικό διάστημα εμφάνισης μετάστασης και γινόταν διαχωρισμός των δειγμάτων (των ασθενών) σε ασθενείς με καλή πρόγνωση (σε χρονικό διάστημα μεγαλύτερο των 5 ετών δεν εμφανίζεται μετάσταση) και σε ασθενείς με φτωχή πρόγνωση (προτού περάσουν τα 5 έτη υπάρχει πιθανότητα εμφάνισης μετάστασης). Μεταξύ των 456 γονιδίων της λίστα που απέφερε η μελέτη της Sorlie et al. (2001) και των 231 γονιδίων της λίστας γονιδίων του van't Veer et al. (2002) εμφανίζονται μόνο 17 κοινά γονίδια, ενώ μεταξύ της λίστα της Sorlie et al. (2001) και του Ramaswamy et al. (2003) μόλις δύο γονίδια είναι κοινά (Ein-Dor et al., 2005). Έτσι τίθεται το εξής ερώτημα: ποιές είναι οι πιθανές αιτίες αυτής της ασυμφωνίας μεταξύ των λιστών των differentially expressed genes που προβλέπουν οι διάφορες μελέτες και γενικά της αστάθειας στην βαθμονόμηση των γονιδίων;

Σύμφωνα με τους Kuo et al. (2002) και Warnat et al. (2005), πολλοί υποστηρίζουν ότι οι αιτίες της χαμηλής επικάλυψης είναι:

i. οι βιολογικές διαφορές μεταξύ των δειγμάτων των διαφόρων μελετών, όπως η ηλικία των ασθενών, το στάδιο της ασθένειας κ.α.
ii. η χρήση διαφορετικών πλατφόρμων microarray (πλατφόρμες που χρησιμοποιούν cDNA έναντι πλατφόρμων που χρησιμοποιούν ολιγονουκλεοτίδες)
iii. οι διαφορές στον εξοπλισμό και στα πρωτόκολλα για την μέτρηση των γονιδιακών εκφράσεων (πχ. στο σκανάρισμα (*scanning*), την ανάλυση εικόνας (*image analysis*) και αλλού)
iv. οι διαφορές στις χρησιμοποιούμενες μεθόδους ανάλυσης

Παρ' όλα αυτά ο Ein – Dor et al. (2005) υποστήριξε ότι ακόμα και αν εξαλειφθούν οι παραπάνω διαφορές, η ασυμφωνία μεταξύ των λιστών, που περιέχουν τα differentially expressed genes, των διαφόρων μελετών παραμένει. Ο ίδιος παρατήρησε ότι εμφανίζεται επίσης μεγάλη αστάθεια στην βαθμονόμηση των γονιδίων ακόμα και μέσα στο ίδιο microarray dataset, πόσο μάλλον μεταξύ των microarray datasets διαφορετικών μελετών. Κατά τον Ein – Dor et al. (2005) λοιπόν, όλη αυτή η αστάθεια και η ασυμφωνία στις λίστες γονιδίων συνεπάγεται έλλειψη αληθοφάνειας και ανικανότητα γενίκευσης των προβλέψεων που οφείλονται σε αυτές τις λίστες.

Μια άλλη απάντηση δόθηκε από τον Somorjai et al. (2003), ο οποίος υποστήριξε ότι ο συνδυασμός της λεγόμενης «κατάρας της αραιότητας των σετ δεδομένων» ("*curse of data set sparsity*"), δηλαδή του περιορισμένου αριθμού των δειγμάτων, και της «κατάρας της διάστασης» ("*curse of dimensionality*"), δηλαδή του τεράστιου αριθμού των γονιδίων, είναι υπεύθυνος για την χαμηλή αυτή επικάλυψη των λιστών των γονιδίων, καθώς τα microarray σετ δεδομένων

είναι ευαίσθητα και στις δύο «κατάρες» (περιέχουν λίγα δείγματα και χιλιάδες γονίδια).

Μια λύση κυρίως στο πρόβλημα της αστάθειας στη βαθμονόμηση των γονιδίων, αλλά και στην χαμηλή overlap μεταξύ των λιστών των μελετών, έρχεται να προτείνει η μέθοδος του Fishel et al. (2007), η οποία βασίζεται στον συνδυασμό της μετα–ανάλυσης και των μηχανών διανυσμάτων υποστήριξης (SVM). Αναλυτική περιγραφή αυτής γίνεται στην επόμενη παράγραφο. Γενικά ο Fishel et al. (2007) τονίζει ότι η χρήση της μετα–ανάλυσης στην συγκεκριμένη περίπτωση βοηθά στον εντοπισμό επαναλαμβανόμενων βιοδεικτών[11] (*biomarkers*), αλλά και στην αντιμετώπιση του μικρού μεγέθους δείγματος των microarray πειραμάτων, εξαλείφοντας έτσι μεροληψίες που εμφανίζονται στην κάθε μελέτη - πείραμα ξεχωριστά, γεγονός που οδηγεί σε πιο έγκυρα, αρμονικά και αληθή αποτελέσματα. Επίσης, οι διαδικασίες μηχανικής μάθησης με επίβλεψη (*supervised machine learning approaches*) αποτελούν ένα ισχυρό εργαλείο στην ανάλυση των γονιδιακών εκφράσεων και κατ' επέκταση στην πρόγνωση, διάγνωση και θεραπεία του καρκίνου.

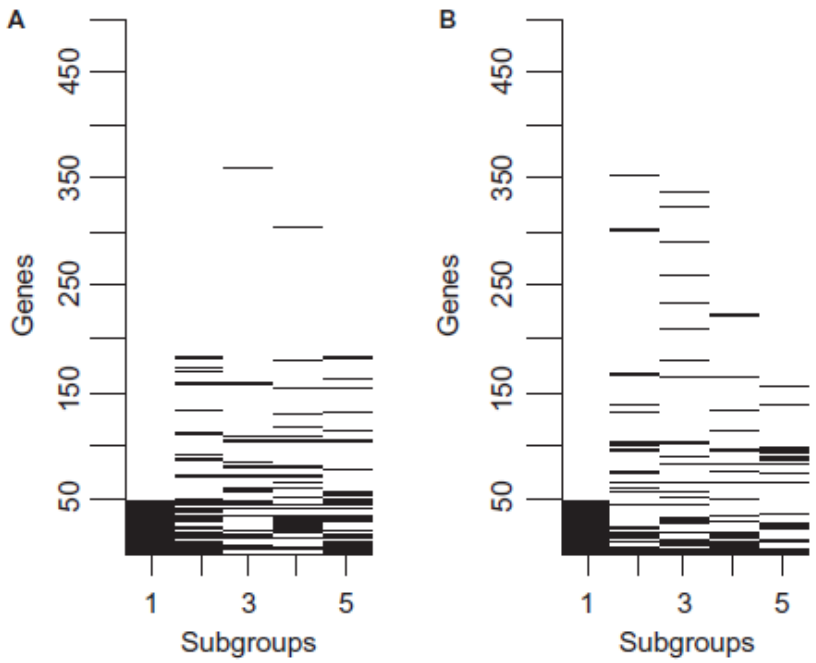
5.3 Εφαρμογή σε δεδομένα σχετικά με τον καρκίνο του πνεύμονα

5.3.1 Εισαγωγή

Σε αυτή την παράγραφο θα παρουσιάσουμε αναλυτικά την μέθοδο που προτείνει και ακολουθεί στην μελέτη του ο Fishel et al. (2007) για την παραγωγή όσο το δυνατόν πιο σταθερών λιστών βαθμονομημένων γονιδίων (RGLs στο paper), με στόχο τον σαφέστερο διαχωρισμό των πνευμονικών ιστών σε φυσιολογικούς και καρκινικούς (*normal versus tumor tissues*).

Αρχικά, ο Fishel et al. (2007) χρησιμοποιεί δύο μελέτες, του Beer et al. (2002) (Michigan) και του Bhattacharjee et al. (2001) (Harvard), και εφαρμόζει σε πέντε διαφορετικά subgroups ασθενών για κάθε μελέτη ξεχωριστά τον αλγόριθμο SVM – RFE για την βαθμονόμηση των γονιδιακών εκφράσεων. Στόχος είναι να τονίσει την εξάρτηση της βαθμονόμησης από το group των ασθενών και άρα την αστάθεια των ranked genes που παρέχει ο SVM – RFE. Τα αποτελέσματα αυτής της εφαρμογής φαίνονται στο παρακάτω σχήμα (Σχήμα 35).

[11] **Βιοδείκτης (*biomarker*):** βιολογικό μόριο που βρίσκεται στο αίμα ή σε άλλα υγρά του σώματος ή στους ιστούς και υποδεικνύει αν μια διαδικασία μέσα στον οργανικό ή εξέλιξη μιας ασθένειας μέσα στον ίδιο είναι φυσιολογική ή όχι. Ένας βιοδείκτης μπορεί να χρησιμοποιηθεί ώστε να μελετηθεί πως ο οργανισμός αντιδρά σε μια θεραπεία (βλ. NCI Dictionary of Cancer Terms).



Σχήμα 35: Τα 50 κορυφαία γονίδια που παρέχονται με την εφαρμογή του αλγορίθμου SVM – RFE σε πέντε διαφορετικά subgroups ασθενών που επιλέχθηκαν τυχαία από τον συνολικό αριθμό των ασθενών του Harvard (A) και του Michigan (B). Κάθε subgroup ασθενών περιέχει το 90% του δείγματος, δηλαδή του συνόλου των ασθενών κάθε μελέτης.

Έτσι παρατηρούμε ότι γονίδια, τα οποία ήταν υψηλά στη βαθμονόμηση χρησιμοποιώντας ένα subgroup ασθενών, χρησιμοποιώντας ένα διαφορετικό subgroup ασθενών αυτό μπορεί να αλλάξει και τα ίδια γονίδια να βρεθούν πολύ χαμηλά στην βαθμονόμηση. Δηλαδή, όπως τονίζει στη μελέτη του και ο Ein-Dor et al. (2005), η βαθμονόμηση των γονιδίων εξαρτάται από το training set. Το φαινόμενο αυτό δεν περιορίζεται μόνο στην περίπτωση περίπλοκων υπολογιστικών προκλήσεων, όπως είναι η εύρεση της υπογραφής προγνωστικών γονιδίων, αλλά παρατηρείται και σε πιο απλά προβλήματα, όπως η δυαδική ταξινόμηση καρκινικών έναντι φυσιολογικών ιστών.

Για την αντιμετώπιση λοιπόν αυτής της αστάθειας, η μελέτη του Fishel et al. (2007) παρουσιάζει μια predictor – based μετα–ανάλυση δύο δημόσια διαθέσιμων microarray datasets σχετικά με τον καρκίνο του πνεύμονα (Beer et al., 2002 και Bhattacharjee et al., 2001). Η μέθοδος που ακολουθείται υπολογίζει μέσω της από κοινού ανάλυσης των δύο σετ δεδομένων μια εύρωστη λίστα γονιδίων (στο paper αναφέρεται ως joint core of genes) και παράγει έναν ακριβή, από την άποψη της transferability[12], SVM ταξινομητή.

[12]**Transferability:** η ικανότητα μιας μεθόδου – ενός μοντέλου πρόβλεψης, που έχει παραχθεί βάση των δεδομένων μιας μελέτης, να διατηρεί την υψηλή αποδοτικότητα της όταν τεστάρετε στα δεδομένα μιας άλλης μελέτης.

Στη συνέχεια, τεστάρεται και αποδεικνύεται μέσω ενός τρίτου συνόλου δεδομένων (Garber et al., 2001) η transferability του joint core of genes. Με αυτόν τον τρόπο αντιμετωπίζεται εκτός από την αστάθεια στην βαθμονόμηση, η έλλειψη transferability στα αποτελέσματα.

5.3.2 Παρουσίαση δεδομένων

Στην μελέτη του Fishel et al. (2007) συμπεριλαμβάνονται τρείς μελέτες: του Michigan (Beer et al., 2002), του Harvard (Bhattacharjee et al., 2001) και του Stanford (Garber et al., 2001), όπως φαίνεται και στον πίνακα που ακολουθεί. Τα δεδομένα του Harvard και του Michigan συνθέτουν την predictor – based μετα – ανάλυση και τα δεδομένα του Stanford χρησιμοποιούνται για την επικύρωση της μεθόδου.

Πίνακας 5: Τα σετ δεδομένων που χρησιμοποιούνται στην ανάλυση στην μελέτη του Fishel et al. (2007).

Data set	Microarray platform	Probe sets	Cancer samples	Normal samples
Michigan (Beer *et al.*, 2002)	Affymetrix (Hu6800)	7127	86	10
Harvard (Bhattacharjee *et al.*, 2001)	Affymetrix (HG_U95Av2)	12600	139	17
Stanford (Garber *et al.*, 2001)	Spotted cDNA	24000	41	5

Και τα τρία σετ δεδομένων λήφθηκαν από δημόσια διαθέσιμους ιστότοπους. Στην ανάλυση συμπεριλαμβάνονται μόνο οι όγκοι αδενοκαρκινώματος (*adenocarcinoma tumors*) και τα φυσιολογικά δείγματα πνεύμονα (*normal lung samples*).

5.3.3 Αναλυτική περιγραφή της μεθόδου

Η μέθοδος που εισάγει ο Fishel et al. (2007) αποτελείται από δύο στάδια. Πιο αναλυτικά:

- **Στάδιο 1:** Δημιουργία σταθερών βαθμονομημένων λιστών (*stable ranked gene lists*), των λεγόμενων RGLs, για κάθε σετ δεδομένων ξεχωριστά.

Σε αυτό το στάδιο λοιπόν ο Fishel et al. (2007) ακολουθεί τα παρακάτω βήματα:

1^{ο} Βήμα → Διάσπαση των δεδομένων σε 80% working set και 20% validation set. Η αναλογία φυσιολογικών και καρκινικών ιστών στο working και στο validation set είναι προσαρμοσμένη σύμφωνα με την αναλογία φυσιολογικών και καρκινικών ιστών σε όλο το δείγμα ($10/86$ για το Michigan και $17/139$ για το Harvard). Το working set συμβάλλει στην κατασκευή των συνόλων προβλεπτικών γονιδίων (*predictive gene sets*) (βλ. 2^{ο} βήμα), βάση των οποίων στη συνέχεια κατασκευάζεται ο predictor, δηλαδή ο SVM ταξινομητής που διαχωρίζει τους καρκινικούς πνευμονικούς ιστούς από τους φυσιολογικούς. Η επικύρωση του SVM γίνεται μέσω του validation set που ορίζεται σε αυτό το βήμα και η απόδοση του υπολογίζεται χρησιμοποιώντας το classification success rate (CSR), που ορίσαμε στο κεφάλαιο 4.

2^{ο} Βήμα → Δημιουργία συνόλου προβλεπτικών γονιδίων (*predictive gene set*). Τα predictive gene sets είναι σύνολα που περιέχουν ένα συγκεκριμένο αριθμό υψηλά βαθμονομημένων γονιδίων. Συγκεκριμένα, στο βήμα αυτό αρχικά προσδιορίζεται ο βέλτιστος αριθμός των γονιδίων, που είναι απαραίτητα για την κατασκευή του SVM predictor και στη συνέχεια επιλέγονται ποια συγκεκριμένα θα είναι τα γονίδια αυτά. Για τον υπολογισμό του βέλτιστου αριθμού των γονιδίων λοιπόν, ο Fishel et al. (2007) τρέχει μια 5 – fold cross validation χρησιμοποιώντας το working set. Σε κάθε fold της διασταυρωμένης επικύρωσης υπολογίζεται μέσω της μεθόδου SVM - RFE μια βαθμονομημένη λίστα των γονιδίων του training set (προέκυψε από την διάσπαση του working set σε 4 υποσύνολα για εκπαίδευση και 1 για επικύρωση) και έπειτα ένας SVM ταξινομητής με γραμμικό πυρήνα εκπαιδεύεται χρησιμοποιώντας κάθε φορά διαφορετικό αριθμό γονιδίων του training set: ξεκινάει από τα 5 πρώτα γονίδια της λίστας που έδωσε ο SVM – RFE και συνεχίζει μέχρι τα 100 γονίδια ανεβαίνοντας κάθε φορά ανά 5 (5, 10, 15, 20, …, 100). Έτσι, προκύπτουν 20 διαφορετικά σφάλματα πρόβλεψης σε κάθε fold της cross validation, και άρα 20 $\mathrm{CV}(\hat{\mathrm{f}})$. Ο αριθμός των γονιδίων για τον οποίο ελαχιστοποιείται η $\mathrm{CV}(\hat{\mathrm{f}})$ είναι και ο βέλτιστος αριθμός γονιδίων, έστω N. Τώρα, αφού υπολογίστηκε ο αριθμός N, με χρήση της μεθόδου SVM – RFE βαθμονομούμε όλα τα γονίδια του working set και επιλέγουμε τα N πιο υψηλόβαθμα γονίδια. Αυτά συνθέτουν και το predictive gene set.

3^{ο} Βήμα → Δημιουργία σταθερής βαθμονομημένης λίστας γονιδίων βάση της επαναληψιμότητας αυτών στα predictive gene sets (*repeatability – based gene list → RGL*). Πιο αναλυτικά, αφού επαναλάβει τα βήματα 1 και 2 1000 φορές και κατασκευάσει 1000 predictive gene sets, ο Fishel et al. (2007) σε αυτό το βήμα κατασκευάζει για κάθε data set μια RGL. Κάθε RGL περιέχει όλα τα γονίδια του συνόλου δεδομένων βαθμονομημένα ανάλογα με την επαναληψιμότητα τους στα predictive gene sets, δηλαδή ανάλογα με την συχνότητα εμφάνισης τους σε

αυτά. Έτσι, τα γονίδια που έχουν την μεγαλύτερη συχνότητα εμφάνισης στα predictive gene sets βρίσκονται στην κορυφή της RGL. Σε αυτό το σημείο, ο Fishel et al. (2007) τονίζει ότι οι RGLs που προκύπτουν για κάθε μελέτη ξεχωριστά μέσω της μεθόδου που περιγράψαμε και οι αντίστοιχες RGLs που προκύπτουν κάνοντας χρήση bootstrapping είναι σχεδόν όμοιες και συγκεκριμένα εμφανίζουν συσχέτιση 0,89 και 0,86 για το Michigan και το Harvard αντίστοιχα, τιμή που εκτιμάται κάνοντας χρήση της Spearman correlation (βλ. παράρτημα 2).

4ο Βήμα → Δημιουργία συνόλου γονιδίων πυρήνα (*gene – core set*). Το gene – core set κάθε μελέτης περιέχει όλα τα γονίδια που εμφανίζουν μη μηδενική επαναληψιμότητα στα predictive gene sets. Δηλαδή, αφαιρούμε ουσιαστικά από την RGL τα γονίδια που δεν εμφανίζονται καθόλου στα predictive gene sets και κατά συνέπεια εμφανίζονται τελευταία στην RGL, καθώς είναι χαμηλά βαθμονομημένα. Τώρα, η βαθμονόμηση των γονιδίων στο gene – core set συμπίπτει με την βαθμονόμηση που έχουν τα γονίδια στην RGL.

- **Στάδιο 2:** Δημιουργία του από κοινού πυρήνα γονιδίων (*joint core of genes*), δηλαδή της σταθερής λίστας που προκύπτει από την κοινή ανάλυση των data set του Michigan και του Harvard και περιλαμβάνει τα γονίδια τα οποία φέρουν την περισσότερη πληροφορία σχετικά με τον διαχωρισμό των πνευμονικών ιστών σε καρκινικούς και μη, τα λεγόμενα όπως προαναφέραμε και differentially expressed genes. Συγκεκριμένα ο joint core περιέχει τα γονίδια που εμφανίζονται κοινά στα gene – core sets του Michigan και του Harvard. Όσον αφορά την βαθμονόμηση των γονιδίων του από κοινού πυρήνα γίνεται ως εξής: αρχικά ταξινομούνται σε φθίνουσα σειρά σε μια λίστα οι βαθμονομήσεις που εμφανίζουν τα γονίδια στα gene – core sets, έτσι κάθε γονίδιο εμφανίζεται με δύο βαθμονομήσεις στη λίστα αυτή. Στη συνέχεια, για κάθε γονίδιο υπολογίζεται ο μέσος όρος των θέσεων που καταλαμβάνουν οι βαθμονομήσεις του στη λίστα, τιμή που αποτελεί και την τελική του βαθμονόμηση στον joint core (πχ. αν οι δύο βαθμονομήσεις ενός γονιδίου καταλαμβάνουν την $3^{η}$ και την $10^{η}$ θέση στη λίστα βαθμονόμησης, το γονίδιο αυτό βαθμονομείται με την τιμή $(3 + 10)/2 = 6{,}5$).

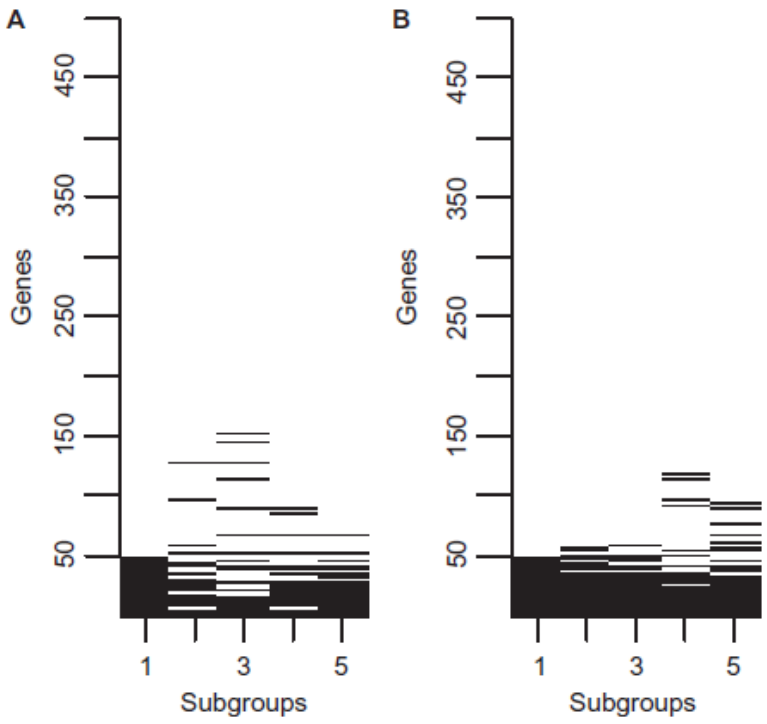
5.3.4 Αποτελέσματα

Αφού τελειώσαμε με την περιγραφή της μεθόδου και την κατασκευή των RGLs και του joint core of genes, σειρά έχει η εξέταση της συνοχής και της σταθερότητας των RGLs, η αξιολόγηση της απόδοσης του SVM predictor, η

αξιολόγηση της transferability του joint core of genes στο data set του Stanford (Garber et al., 2001), η σύγκριση των RGLs του Michigan και του Harvard, καθώς και η μελέτη της βιολογικής σημαντικότητας (*biological significance*) των 118 γονιδίων του joint core. Πιο αναλυτικά:

i. Εξέταση της συνοχής των παραγόμενων RGLs του Michigan και του Harvard: το στάδιο 1 της μεθόδου που περιγράψαμε επαναλαμβάνεται δύο φορές για κάθε σετ δεδομένων, παράγοντας έτσι 4000 predictive gene sets και κατ' επέκταση δύο RGLs για το Michigan και δύο RGLs για το Harvard. Στη συνέχεια, υπολογίζεται η Spearman correlation (βλ. παράρτημα 2) μεταξύ των δύο λιστών για κάθε data set. Συγκεκριμένα, οι δύο RGLs που προέκυψαν για το Michigan εμφανίζουν συσχέτιση μεταξύ τους 0,84 με τυπική απόκλιση 0,01 και οι δύο RGLs που προέκυψαν από τα δεδομένα του Harvard εμφανίζουν συσχέτιση μεταξύ τους 0,86 με τυπική απόκλιση 0,008. Γενικά, λοιπόν, παρατηρούμε ότι οι τιμές της Spearman correlation είναι αρκετά υψηλές, γεγονός που μας οδηγεί στο συμπέρασμα ότι η συνοχή των παραγόμενων RGLs του Michigan και του Harvard είναι μεγάλη.

ii. Εξέταση της σταθερότητας των παραγόμενων RGLs του Michigan και του Harvard: σε αυτό το σημείο ο Fishel et al. (2007) υπολογίζει για κάθε data set την επικάλυψη (overlap) μεταξύ των RGLs, που παράγονται βάσει πέντε διαφορετικών ομάδων ασθενών (κάθε ομάδα περιέχει περίπου το 90% του δείγματος) και την συγκρίνει με την overlap που προκύπτει χρησιμοποιώντας την μέθοδο SVM – RFE στα ίδια πέντε groups ασθενών. Αρχικά, παρατηρώντας το παρακάτω σχήμα (Σχήμα 36) εύκολα διαπιστώνει κανείς την φανερή επαναληψιμότητα των 50 top γονιδίων που δίνουν οι RGLs σε κάθε ομάδα ασθενών. Συγκεκριμένα, η μέση παρατηρούμενη επικάλυψη μεταξύ των 50 top βαθμονομημένων γονιδίων των πέντε διαφορετικών υποσυνόλων ασθενών είναι 37 στα 50, με τυπική απόκλιση 2,86, για το Harvard και 40,6 στα 50, με τυπική απόκλιση 3,23, για το Michigan. Οι τιμές αυτές είναι φανερά αρκετά υψηλές, ειδικά συγκρίνοντας τες με τις αντίστοιχες που προκύπτουν από την βαθμονόμηση με την SVM – RFE: 24,1 στα 50, με τυπική απόκλιση 8,34, για το Harvard και 26,8 στα 50, με τυπική απόκλιση 9,54, για το Michigan. Το πλεονέκτημα της μεθόδου του Fishel et al. (2007), δηλαδή των RGLs, έναντι της χρήσης της μεθόδου SVM – RFE για την βαθμονόμηση των γονιδίων, γίνεται επίσης ξεκάθαρο παρατηρώντας και συγκρίνοντας τα Σχήματα 35 και 36. Έτσι οδηγούμαστε στο συμπέρασμα ότι πράγματι οι RGLs είναι σταθερές εύρωστες λίστες βαθμονομημένων γονιδίων.

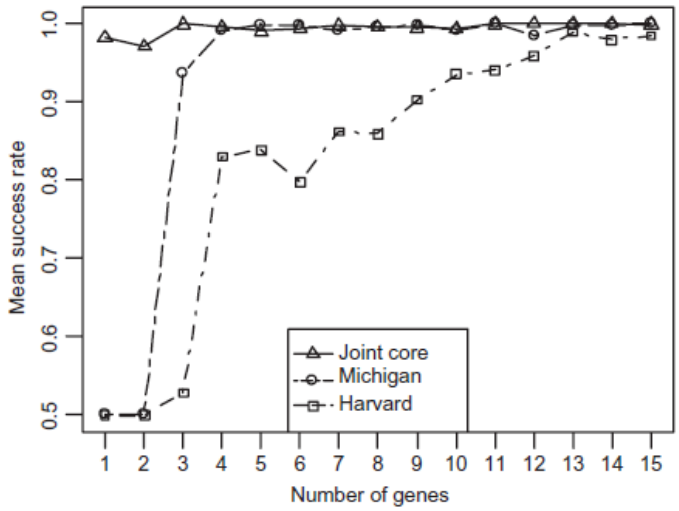


Σχήμα 36: Τα 50 κορυφαία γονίδια που παρέχονται από τις RGLs πέντε διαφορετικών subgroups ασθενών που επιλέχθηκαν τυχαία από τον συνολικό αριθμό των ασθενών του Harvard (A) και του Michigan (B). Κάθε subgroup ασθενών περιέχει το 90% του δείγματος, δηλαδή του συνόλου των ασθενών κάθε μελέτης.

iii. Αξιολόγηση της απόδοσης του SVM predictor: καταρχάς ο Fishel et al. (2007) αναφέρει ότι ο μέσος αριθμός των γονιδίων που συμμετέχουν στα predictive gene sets του Michigan και του Harvard αντίστοιχα είναι $\mathrm{mean(N)} = 15{,}8$ και $\mathrm{mean(N)} = 27{,}8$, με τυπικές αποκλίσεις $\mathrm{SD(N)} = 17{,}5$ και $\mathrm{SD(N)} = 24{,}3$. Τώρα, τα predictive gene sets, μέσω των οποίων προκύπτουν οι RGLs για το Michigan και το Harvard, εμφανίζουν μεγάλα ποσοστά απόδοσης, συγκεκριμένα $\mathrm{meanCSR} = 98{,}6\%$ και $\mathrm{meanCSR} = 90\%$ για το Michigan και το Harvard αντίστοιχα. Άρα, συμπεραίνουμε ότι ο SVM predictor που κατασκευάζεται σε αυτή τη μελέτη, αποτελεί ένα μοντέλο υψηλής απόδοσης για τον διαχωρισμό των πνευμονικών ιστών σε καρκινικούς και φυσιολογικούς.

iv. Αξιολόγηση της transferability του joint core of genes: παρέχει πράγματι ο joint core περισσότερη πληροφορία σε σχέση με τα gene – core sets του Michigan και του Harvard ξεχωριστά; Προτού περιγράψουμε την επικύρωση του joint core χρησιμοποιώντας το σετ δεδομένων του Stanford, δίνουμε τα μεγέθη των gene – core sets του Michigan και του Harvard, καθώς και του joint core of genes. Συγκεκριμένα, από τα 4579 γονίδια που περιέχουν τα δύο data sets μόνο 547 από αυτά συνθέτουν το gene – core set του Harvard και 411 το gene – core set του Michigan, γεγονός που φανερώνει ότι πολλά από τα γονίδια του δείγματος των δύο σετ δεδομένων δεν παρέχουν καμία πληροφορία σχετικά με τον

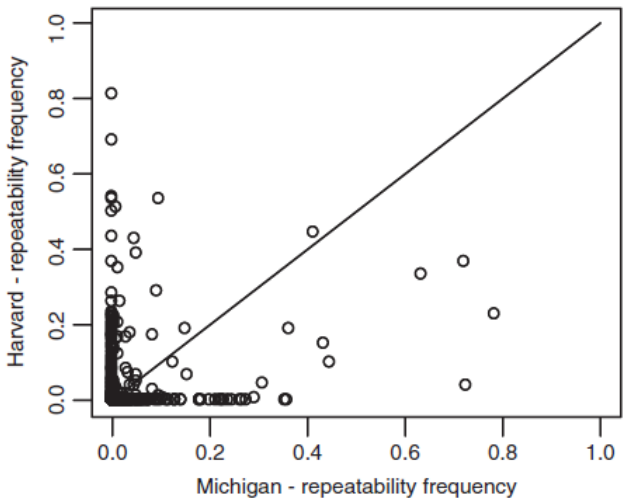
διαχωρισμό των πνευμονικών ιστών σε καρκινικούς και φυσιολογικούς. Επιπλέον, το μέγεθος του joint core είναι 118 γονίδια, και θεωρείται στατιστικά σημαντικό. Τώρα, όσον αφορά την transferability του joint core of genes σε σχέση με την transferability των δύο gene – core sets του Michigan και του Harvard ξεχωριστά, ο Fishel et al. (2007) για να την αξιολογήσει ακολουθεί την εξής διαδικασία 100 φορές για κάθε αριθμό επιλεγμένων γονιδίων: διασπά το data set του Stanford σε 80% training set και 20% test set για την εκπαίδευση και την επικύρωση ενός SVM ταξινομητή αντίστοιχα, αρχικά χρησιμοποιώντας μόνο τα γονίδια που περιέχει το gene – core set του Michigan ξεκινώντας με το πιο υψηλόβαθμο γονίδιο και πηγαίνοντας μέχρι και το 15ο γονίδιο, και στη συνέχεια κάνοντας την ίδια διαδικασία για το gene – core set του Harvard και τέλος για τον joint core. Έτσι, εκτιμώντας την απόδοση του SVM ταξινομητή μέσω του μέτρου CSR που ορίσαμε στο κεφάλαιο 4 (βλ. σελίδα 96), προκύπτουν τα αποτελέσματα που φαίνονται στο παρακάτω σχήμα (Σχήμα 37).



Σχήμα 37: Αξιολόγηση της transferability του joint core.

Παρατηρείται, λοιπόν, ότι ο joint core χρησιμοποιώντας μόλις λιγότερα από τέσσερα top βαθμονομημένα γονίδια για την ταξινόμηση, πετυχαίνει πολύ υψηλά ποσοστά απόδοσης. Συγκεκριμένα, το πιο υψηλόβαθμο γονίδιο του joint core (το RAGE) εμφανίζει $\mathrm{CSR} = 98\%$, απόδοση την οποία για να πετύχει το gene – core set του Michigan απαιτεί την χρήση τεσσάρων γονιδίων, ενώ το gene – core set του Harvard καταφέρνει να φτάσει σε απόδοση 98% χρησιμοποιώντας τα 13 top βαθμονομημένα γονίδια του. Άρα, συμπεραίνουμε ότι προφανώς ο joint core of genes φέρει περισσότερη πληροφορία, χρήσιμη για την ταξινόμηση των πνευμονικών ιστών σε καρκινικούς και φυσιολογικούς, σε σχέση με κάθε gene – core set ξεχωριστά.

v. Σύγκριση των RGLs του Michigan και του Harvard: εξαιτίας της σταθερότητας και την συνοχής που παρουσιάζει κάθε RGL ξεχωριστά, κανείς θα αναμένει ότι τα γονίδια που είναι υψηλά βαθμονομημένα στην RGL του Michigan θα είναι υψηλά βαθμονομημένα και στην RGL του Harvard, δηλαδή θα υπάρχει μια σχετική ταύτιση – μια σχετική συμφωνία μεταξύ των gene – core sets των δύο σετ δεδομένων. Αυτό δυστυχώς, όπως επισημαίνει ο Fishel et al. (2007), δεν συμβαίνει. Συγκεκριμένα, 6 από τα 10 πιο υψηλόβαθμα γονίδια του Harvard gene – core set δεν εμφανίζονται καθόλου στο gene – core set του Michigan, γεγονός που μπορεί να υποδηλώνει ότι τα γονίδια αυτά δεν είναι αντιπροσωπευτικά του προβλήματος, αλλά είναι “data set specific”, δηλαδή εμπεριέχουν μια δειγματική μεροληψία. Αντίθετα, 8 από τα 10 πιο υψηλόβαθμα γονίδια του gene – core set του Michigan εμπεριέχονται στο gene – core set του Harvard και εμφανίζονται και σε αυτό με σχετικά υψηλή βαθμονόμηση. Άρα, συμπεραίνουμε ότι τα γονίδια αυτά εμφανίζουν μεγαλύτερη αξιοπιστία και πράγματι είναι πιο πιθανό να συμβάλουν σημαντικά στον διαχωρισμό των καρκινικών και των φυσιολογικών πνευμονικών ιστών σε σχέση με άλλα. Το παρακάτω σχήμα (Σχήμα 38) και ο πίνακας (Πίνακας 6) επιβεβαιώνουν την ασυμφωνία μεταξύ των RGLs του Michigan και του Harvard.



Σχήμα 38: Σύγκριση των βαθμονομήσεων των γονιδίων που περιέχουν οι RGLs του Michigan και του Harvard. Κάθε σημείο αναπαριστά την βαθμονόμηση ενός γονιδίου σύμφωνα με τις RGLs του Michigan (x άξονας) και του Harvard (y άξονας). Τα σημεία που είναι κοντά στην διαγώνιο αναπαριστούν γονίδια τα οποία είναι ισοδύναμα βαθμονομημένα στις RGLs του Michigan και του Harvard.

Συγκεκριμένα, παρατηρούμε ότι ελάχιστα σημεία είναι κοντά στη διαγώνιο, πράγμα που σημαίνει ότι ελάχιστα γονίδια εμφανίζουν την ίδια βαθμονόμηση και στις δύο RGLs.

Πίνακας 6: Τα 10 πιο υψηλόβαθμα γονίδια του joint core of genes, του gene – core set του Michigan και του gene – core set του Harvard. Τα γονίδια που δεν εμφανίζονται στον joint core είναι γραμμένα σε bold.

	Joint core	Michigan core-set	Harvard core-set
1	RAGE	TNXB	**SMAD6**
2	TNA	CA4	**GRK5**
3	FABP4	RAGE	**HYAL2**
4	TNXB	FABP4	TEK
5	*COX7A1*	FGR	**CD34**
6	PHLDA2	PHLDA2	*S100A3*
7	FGR	TNA	**FKBP1A**
8	TEK	*COX7A1*	TNA
9	TACSTD1	**CEACAM5**	**TLK1**
10	MAP4	**CASP1**	EMP2

Τέλος, η ανομοιότητα μεταξύ των RGLs του Michigan και του Harvard επιβεβαιώνεται και από την χαμηλή Spearman correlation που εμφανίζουν μεταξύ τους, μόλις 0,173.

vi. Μελέτη της βιολογικής σημαντικότητας (*biological significance*) των 118 γονιδίων του joint core: οι Hanahan & Weinberg (2000) όρισαν τις λεγόμενες «σφραγίδες του καρκίνου» (“*hallmarks of cancer*”), οι οποίες βάσει ουσιωδών μεταβολών στην φυσιολογία των κυττάρων, υποδηλώνουν την γέννηση καρκίνου (*tumor genesis*). Στον joint core προκύπτουν εξαιτίας συγκεκριμένων γονιδίων αρκετά σημάδια αυτών των μεταβολών. Συγκεκριμένα, η αυτοεπάρκεια σε σήματα ανάπτυξης (*Self – Sufficiency in growth signals*) (γονίδιο ErbB3: με βαθμονόμηση στον joint core → 72), η αναισθησία σε σήματα αντιανάπτυξης (*Insensitivity to antigrowth signals*) (TGFBR3: 36), η απόφυση απόπτωσης (*Evading apoptosis*) (PHLDA2: 6, SPP1: 21, ZBTB16: 32, DNASE1L3: 38, CSF2RB: 60, PML: 80, IGFBP3: 81, TNFRSF25: 82), η παρατεταμένη αγγειογένεση (*Sustained angiogenesis*) (TEK ή TIE-2: 8, MDK: 15, EDNRB: 23, PECAM1 ή CD31: 24, ANG1: 35, CDH1: 65) και η εισβολή και μετάσταση των ιστών (*Tissue invasion and metastasis*) (RAGE:1, S100A4: 94, S100A3:18, S100G: 30, S100A8: 52, CAV1: 13, SPP1: 21, SPINT2: 58). Αν και είναι αρκετά ενθαρρυντικό το αποτέλεσμα ότι αρκετά γονίδια του joint core εμφανίζονται να εμπλέκονται στην πρόκληση καρκίνου, ο ρόλος αυτών στην ουσιαστική παθογένεση αδενοκαρκινόματος χρήζει περαιτέρω έρευνας.

5.3.5 Τελικά συμπεράσματα

Η μέθοδος ανάλυσης που προτείνει ο Fishel et al. (2007) στην μελέτη του αυξάνει την αξιοπιστία και την ισχύ των predictive gene sets, παράγει σταθερές και συνεκτικές λίστες βαθμονόμησης των γονιδίων (τις RGLs) και κατασκευάζει έναν joint core υψηλής transferability, που με μόλις τρία top βαθμονομημένα γονίδια πετυχαίνει 99,8% απόδοση του SVM predictor. Παρ' όλα αυτά η ασυμφωνία μεταξύ των RGLs των διαφόρων μελετών παραμένει, με την μεταξύ τους συσχέτιση να είναι μόνο 0,173, φαινόμενο το οποίο ο Fishel et al. (2007) αποδίδει σε παράγοντες, όπως οι βιολογικές διαφορές μεταξύ των δειγμάτων των μελετών και η χρήση διαφορετικών πλατφόρμων και πρωτόκολλων, παρά στην εσωτερική αστάθεια των δεδομένων. Έτσι, ο Fishel et al. (2007) καταλήγει λέγοντας ότι τα γονίδια του joint core μπορεί να αποτελούν στόχους θεραπείας και σημάδια διάγνωσης του καρκίνου. Επίσης, η χρήση της προτεινόμενης μεθόδου από άλλους ερευνητές και η ένταξη περισσότερων των δύο μελετών σε αυτή μπορεί να βοηθήσει στην καλύτερη κατανόηση προηγούμενων microarray μελετών, χωρίς να χρειάζεται η πραγματοποίηση επιπλέον πειραμάτων.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Beer D.G. et al. (2002). Gene - expression profiles predict survival of patients with lung adenocarcinoma. *Nature Medicine*, **8**: 816-824.

[2] Bhattacharjee A. et al. (2001). Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences (PNAS)*, **98**: 13790-13795.

[3] Bradley A. P. (1997). The Use of the Area Under a ROC Curve in the Evaluation of Machine Learning Algorithms. *Pattern Recognition*, **30(7)**: 1145-1159.

[4] Breiman et al. (1984). *Classification and Recognition Trees*. Monterey, CA: Wadsworth.

[5] Cohen J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates (LEA Publications).

[6] Corder G.W. & Foreman D.I. (2009). *Nonparametric Statistics for Non – Statisticians: a step – by – step approach*. Wiley.

[7] Cortes C. & Vapnik V. (1995). Support – Vector Networks. *Machine Learning*, **20**: 273-297.

[8] Ein-Dor I. et al. (2005). Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, **21**: 171-178.

[9] Fishel et al. (2007). Meta analysis of gene expression data: a predictor based approach. *Bioinformatics*, **23**: 1599-1606.

[10] Garber M.E. et al. (2001). Diversity of gene expression in adenocarcinoma of the lung. *Proceedings of the National Academy of Sciences (PNAS)*, **98**: 13784-13789.

[11] Gioacchino L. (2005). *Meta - analysis in Medical Research. The handbook for the understanding and practice of meta – analysis*. Blackwell Publishing, BMJ Books.

[12] Glass GE. (1976). Primary, Secondary and Meta – Analysis of Research. *Educational Research*, **5(10)**: 3-8.

[13] Golub T.R. et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**: 531-537.

[14] Guyon I. et al. (2002). Gene Selection for Cancer classification using Support Vector Machines. *Machine Learning Journal*, **46**: 389-422.

[15] Hamel L.H. (2009). *Knowledge Discovery with Support Vector Machines*. Wiley Series on Methods and Applications in Data Mining. Daniel T. Larose, Series Editor.

[16] Hanahan D. & Weinberg R.A. (2000). The hallmarks of cancer. *Cell*, **100**: 57-70.

[17] Hand D.J. (1981). *Discrimination and Classification*. John Wiley & Sons, Chichester — Brigbane — New York — Toronto 1981.

[18] Hand D.J. & Till R.J. (2001). A Simple Generalization of the Area Under the ROC Curve for Multiple Class Classification Problems. *Machine Learning*, **45(2)**: 171-186.

[19] Hanley JA. & McNeil BJ. (1982). The Meaning and Use of the Area Under a Receiver operating Characteristic (ROC) Curve. *Radiology*, **143(1)**: 29-36.

[20] Hastie D. et al. (2001). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. Spinger Series in Statistics, Spinger Verlag, New York.

[21] Haykin S. (2009). *Νευρωνικά Δίκτυα και Μηχανική Μάθηση*. Εκδόσεις Παπασωτηρίου.

[22] Hedges LV. & Olkin I. (1985). *Statistical Methods for Meta – Analysis*. Orlando, Florida, Academic Press.

[23] Higgins J. et al. (2003). *Cochrane Handbook for Systematic Reviews of Interventions*. Wiley, 2008.

[24] Hongfang L. et al. (2014). Microarrays probes and probe sets. *National Institutes of Health. Frontiers in Bioscience (Elite edition)*, **2**: 325-338.

[25] Joachims T. (1999). Transductive Inference for Text Classification using Support Vector Machines. *Proceedings of the Sixteenth International Conference on machine Learning*, p. 200-209, June 27-30, 1999.

[26] Kim J. (2011). Meta – analysis. *Statistical Seminar Fall 2011*, University of Utah, School of Medicine, Study Design and Biostatistics Center, Department of Family and Preventive Medicine.

[27] Kuo W.P. et al. (2002). Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics*, **18**: 405-412.

[28] Lebrun G. et al. (2006). A New Model Selection Method for SVM. E. Corchado et al. (Eds.): *Intelligent Data Engineering and Automated Learning - IDEAL 2006*, LNCS 4224, pp. 99–107, *Springer-Verlag Berlin Heidelberg 2006*.

[29] Lusted L.B. (1971). Signal Detectability and Medical Decision Making. *Science*, **171**: 1217-1219.

[30] Noble WS. (2006). What is Support Vector Machine? *Nature Biotechnology*, **24(12)**: 1565-1567.

[31] Normand S-L.T. (1998). Tutorial in Biostatistics. Meta-analysis: Formulating, Evaluating, Combining, and Reporting. *Statistics in Medicine*, **18**: 321-359 (1999).

[32] Olkin I. (1995). Keynote Addresses. Meta – analysis: Reconciling the results of independent studies. *Statistics in Medicine*,**14**: 457-472.

[33] Peterson W.W. et al. (1954). The Theory of Signal Detection Theory. *Transactions of the IRE Professional Group of Information Theory*, 171-212.

[34] Ramaswamy S. et al. (2003). A molecular signature of metastasis in primary solid tumors. *Nature Genetics*, **33**: 49-54.

[35] Somorjai R.L. et al. (2003). Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. *Bioinformatics*, **19**: 1484-1491.

[36] Sorlie T. et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences of the United States of America*, **98**: 10869-10874.

[37] Spackman K.A. (1989). Signal Detection Theory: Valuable Tools for Evaluating Inductive Learning. *Proceedings of the Sixth International Workshop on Machine Learning*, San Mateo, CA, pp. 160-163, Morgan Kaufman.

[38] Swets J. (1988). Measuring the Accuracy of Diagnostic Systems. *Science*, **240**: 1285-1293.

[39] Tan P-N. et al. (2006). *Introduction to Data Mining*. Addison – Wesley Companion Book Site, 2006.

[40] Tong S. & Chang E. (2001): Support Vector Machine Active Learning for Image Retrieval. *Proceedings of the Ninth ACM International Conference on Multimedia*, September 30 – October 5, Ottawa, Canada.

[41] Tong S. & Koller D. (2000). Support Vector Machine Active Learning with Application to Text Classification. *Proceedings of the Seventeenth International Conference on Machine Learning*, p. 999-1006, June 29 – July 2, 2000.

[42] van't Veer L.J. et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**: 530-536.

[43] Vardhanabhuti S. et al. (2006). A comparison of statistical tests for detectin differential expression using Affymetrix oligonucleotide microarrays. *OMICS*, **10(4)**: 555-566.

[44] Warnat P. et al. (2005). Cross – platform analysis of cancer microarray data improves gene expression based classification of phenotypes. *BMC Bioinformatics*, **6**: 265.

[45] Whitehead A. & Whitehead J. (1991). A General Parametric Approach to the Meta – Analysis of Randomized Clinical Trials. *Statistics in Medicine*, **10**: 1665-1677.

[46] Γαλάνης Π. (2009). Συστηματική Ανασκόπηση και Μετα-Ανάλυση, *Αρχεία Ελληνικής Ιατρικής 2009* (*Archives of Hellenic Medicine 2009*), **26(6)**: 826-841.

[47] Πέτρογλου Ν. & Σπάρος Λ. (2004). Καμπύλη ROC στη Διαγνωστική Έρευνα. *Αρχεία Ελληνικής Ιατρικής*, **21(2)**: 179-194.

[48] Τσιάρα Χ. et al. (2011). Περιγραφή της Μεθοδολογίας της Μετα – Ανάλυσης με Παράθεση Παραδείγματος από τον Χώρο της Περιοδοντολογίας. *Στοματολογία*, **68(2)**: 63-71.

[49] Ανδρεάδης Ι.Ι. (2014). Ανάπτυξη Μεθοδολογιών για την Υποβοηθούμενη Διάγνωση Συμπλεγμάτων Μικροασβεστώσεων του Μαστού. Διδακτορική Διατριβή, ΕΜΠ.

[50] Βαλάντη Ε. (2011). Ανάλυση των Καμπυλών ROC και Εφαρμογή τους σε Πραγματικά Βιοϊατρικά Δεδομένα. Διπλωματική Εργασία, ΕΜΠ.

[51] Γιαννούλη Π.Δ. (2014). Εφαρμογές των Μηχανών Διανυσματικής Υποστήριξης σε Προβλήματα Ταξινόμησης και Παλινδρόμησης. Διπλωματική Εργασία, ΕΜΠ.

[52] Γύπας Φ. (2011). Ανάπτυξη Νευρωνικού Μοντέλου για Γονιδιακή Ανάλυση. Διπλωματική Εργασία, Πολυτεχνείο Κρήτης.

[53] Δρόσου Π.Κ. (2013). Στατιστικές Μέθοδοι για την Ανάλυση Δεδομένων Υψηλής Διάστασης. Διπλωματική Εργασία, ΕΜΠ.

[54] Εγγλέζου Π.Γ. (2014). Επιλογή Μεταβλητών για την Ταξινόμηση με Μηχανές Διανυσματικής Υποστήριξης. Διπλωματική Εργασία, ΕΜΠ.

[55] Θεοδόση – Κοκκίνου Λ. (2013). Τεχνητά Νευρωνικά Δίκτυα και Εφαρμογές στα Συστήματα Αυτόματου Ελέγχου. Διπλωματική Εργασία, Πανεπιστήμιο Πατρών.

[56] Κυρίτσης Κ. (2014). Νευρωνικά Δίκτυα και Μηχανές Διανυσματικής Υποστήριξης. Διπλωματική Εργασία, Πανεπιστήμιο Πατρών.

[57] Λόκας Μ. (2012). Ταξινόμηση Ανεπιθύμητης Αλληλογραφίας εφαρμόζοντας Στατιστικές Τεχνικές Ταξινόμησης με την Γλώσσα Προγραμματισμού R. Διπλωματική Εργασία, ΑΠΘ.

[58] Λουκίνα Β. (2012). Διαχωριστική Ανάλυση, Ταξινόμηση και Ομαδοποίηση Δεδομένων με Εφαρμογές στο SPSS. Μεταπτυχιακή Διπλωματική Εργασία, Πανεπιστήμιο Πατρών.

[59] Ξενή Μ. (2012). Λογιστική Παλινδρόμηση και Διαχωριστική Ανάλυση. Μεταπτυχιακή Διπλωματική Εργασία, Πανεπιστήμιο Πατρών.

[60] Πετρόχειλος Ο. (2009). Επιλογή Χαρακτηριστικών για Προβλήματα Ταξινόμησης. Μεταπτυχιακή Διπλωματική Εργασία.

[61] Στούφη Ι. Ε. (2015). Κριτήρια Πληροφορίας για την Επιλογή Μεταβλητών στις Μηχανές Διανυσματικής Υποστήριξης και Εφαρμογές. Μεταπτυχιακή Διπλωματική Εργασία, ΕΜΠ.

[62] Στρατή Ι.Π. (2007). Meta – Analysis. Μεταπτυχιακή Διπλωματική Εργασία, ΑΠΘ.

[63] Τσακανίκας Π. (2005). Αναγνώριση Γονιδιακών Εκφράσεων Νεοπλασιών με σε Microarrays. Διπλωματική Εργασία, Πανεπιστήμιο Πατρών.

[64] Τσουχνικά Μ. (2007). Νευρωνικά Δίκτυα και Εφαρμογές. Μεταπτυχιακή Διπλωματική Εργασία, ΑΠΘ.

[65] Χατζηζαχαρίας Κ. (2014). Επιλογή Χαρακτηριστικών για Ταξινόμηση με την Βοήθεια Μέτρων Πληροφορίας. Διπλωματική Εργασία, ΕΜΠ.

ΠΑΡΑΡΤΗΜΑ 1: Εφαρμογή της μεθόδου SVM – RFE: εντολές στην R

1. Εισαγωγή δεδομένων “Statlog (Heart) Data Set”

```
> data<-read.table("data.txt")
> heartdis<-as.matrix(data)
> age<-heartdis[,1]
> x2<-heartdis[,2]
> x3<-heartdis[,3]
> trestbps<-heartdis[,4]
> chol<-heartdis[,5]
> x6<-heartdis[,6]
> restecg<-heartdis[,7]
> thalach<-heartdis[,8]
> x9<-heartdis[,9]
> oldpeak<-heartdis[,10]
> x11<-heartdis[,11]
> ca<-heartdis[,12]
> x13<-heartdis[,13]
> num<-heartdis[,14]
> Y<-as.factor(num)
> sex<-as.factor(x2)
> cp<-as.factor(x3)
> fbs<-as.factor(x6)
> exang<-as.factor(x9)
> slope<-as.factor(x11)
> thal<-as.factor(x13)
>heart<-cbind(age,sex,cp,trestbps,chol,fbs,restecg,thalach,exang,oldpeak,
slope,ca,thal,Y)
```

2. Δημιουργία training set και test set

```
> i<-1:270
> j<-sample(i,size=trunc(length(i)/4))
> testset<- heart[j,]
> trainingset<- heart[-j,]
```

3. Λήψη και φόρτωση του πακέτου ‘e1071’

```
> install.packages("e1071")
> library(e1071)
```

4. Κατασκευή κώδικα SVM - RFE

```
> SVMRFE<-function(x,y){
+ i<-1
+ s<-1:ncol(x)
+ R<-vector(mode="integer",length=ncol(x))
+ while(length(s)>0){
+ SVMtrain<-svm(x[,s],y,scale=FALSE,type='C-classification',kernel='sigmoid')
+ w<-crossprod(SVMtrain$coefs,SVMtrain$SV)
+ c<-w^2
+ argminc<-sort(c,index.return=TRUE)$ix
+ R[i]<-s[argminc[1]]
+ i<-i+1
+ s<-s[-argminc[1]]
+ }
+ return(R)
+ }
```

5. Εφαρμογή κώδικα στα δεδομένα του σετ “Statlog (Heart) Data Set”

```
> features<-trainingset[,-14]
> responce<-trainingset[,14]
> SVMRFE(features,responce)
```

ΠΑΡΑΡΤΗΜΑ 2: Spearman correlation

Η συσχέτιση του Spearman είναι κατάλληλη και για συνεχείς και για διακριτές μεταβλητές. Έστω, λοιπόν, ότι θέλουμε να εξετάσουμε την συσχέτιση μεταξύ δύο μεταβλητών $X = (X_1, X_2, \dots, X_n)$ και $Y = (Y_1, Y_2, \dots, Y_n)$. Τότε, αρχικά, πρέπει να βρούμε τις κατατάξεις τους $x = (x_1, x_2, \dots, x_n)$ και $y = (y_1, y_2, \dots, y_n)$ αντίστοιχα (βλέπε παράδειγμα) και στη συνέχεια να χρησιμοποιήσουμε τον ακόλουθο τύπο που δίνει την Spearman correlation.

$$p = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Ωστόσο, αν $\forall i$ τα X_i είναι διαφορετικά μεταξύ τους, δηλαδή δεν υπάρχουν επαναλαμβανόμενες τιμές, όμοια και για τα Y_i, τότε μπορούμε να χρησιμοποιήσουμε τον απλούστερο τύπο (Corder & Foreman, 2009):

$$p = 1 - \frac{6\sum_{i=1}^{n} {d_i}^2}{n(n^2 - 1)}$$

όπου

$$d_i = x_i - y_i$$

Η Spearman correlation κυμαίνεται από -1 έως 1 και με τις τιμές 1 και -1 να συμβολίζουν μια τέλεια συσχέτιση και την τιμή μηδέν να συμβολίζει την ανυπαρξία συσχέτισης μεταξύ των μεταβλητών.

Για να διευκρινίσουμε το θέμα των κατατάξεων παρουσιάζουμε το εξής παράδειγμα: έστω ότι θέλουμε να εξετάσουμε την συσχέτιση μεταξύ του IQ ενός ατόμου με τις εβδομαδιαίες ώρες που το ίδιο παρακολουθεί τηλεόραση. Στον παρακάτω πίνακα δίνονται τα δεδομένα.

Πίνακας 7: Πίνακας δεδομένων. Η μεταβλητή X αναπαριστά τα επίπεδα IQ και η μεταβλητή Y τις ώρες παρακολούθησης τηλεόρασης / εβδομάδα.

X		Y	
X_1	106	Y_1	7
X_2	86	Y_2	0
X_3	100	Y_3	27
X_4	101	Y_4	50
X_5	99	Y_5	28
X_6	103	Y_6	29
X_7	97	Y_7	20
X_8	113	Y_8	12
X_9	112	Y_9	6
X_{10}	110	Y_{10}	17

Βήμα 1: Ταξινόμηση των X_i σε αύξουσα σειρά και δημιουργία των κατατάξεων τους x_i.

X		x	
X_2	86	x_1	1
X_7	97	x_2	2
X_5	99	x_3	3
X_3	100	x_4	4
X_4	101	x_5	5
X_6	103	x_6	6
X_1	106	x_7	7
X_{10}	110	x_8	8
X_9	112	x_9	9
X_8	113	x_{10}	10

Βήμα 2: Αντιστοίχηση των τιμών Y_i με την ταξινομημένη λίστα των X_i βάση του αρχικού πίνακα δεδομένων (Πίνακας 7) και δημιουργία των κατατάξεων τους y_i (από το μικρότερο Y_i στο μεγαλύτερο).

X		X		Y		y	
X_2	86	x_1	1	Y_2	0	y_1	1
X_7	97	x_3	2	Y_7	20	y_2	6
X_5	99	x_3	3	Y_5	28	y_3	8
X_3	100	x_4	4	Y_3	27	y_4	7
X_4	101	x_5	5	Y_4	50	y_5	10
X_6	103	x_6	6	Y_6	29	y_6	9
X_1	106	x_7	7	Y_1	7	y_7	3
X_{10}	110	x_8	8	Y_{10}	17	y_8	5

X_9	112	x_9	9	Y_9	6	y_9	2
X_8	113	x_{10}	10	Y_8	12	y_{10}	4

Βήμα 3: Παρατηρούμε ότι δεν υπάρχουν επαναλαμβανόμενες τιμές, άρα μπορούμε να χρησιμοποιήσουμε τον απλούστερο τύπο με τα d_i (Corder & Foreman, 2009). Αν υπήρχε επαναλαμβανόμενη τιμή, τότε η κατάταξη της υπολογίζεται ως ο μέσος όρος των θέσεων που αυτή καταλαμβάνει, για παράδειγμα:

X		Θέσεις	x	
X_1	0,8	1	x_1	1
X_2	1,2	2	x_2	$2+3/2$
X_3	1,2	3	x_3	$2+3/2$
X_4	2,3	4	x_4	4
X_5	18	5	x_5	5

Υπολογίζουμε λοιπόν τα d_i.

d_i	$d_i{}^2$
0	0
-4	16
-5	25
-3	9
-5	25
-3	9
4	16
3	9
7	49
6	36

Οπότε έχουμε ότι $\sum_{i=1}^{10} d_i{}^2 = 194$ και άρα $p = 1 - \frac{6 \cdot 194}{10 \cdot (10^2-1)} \approx -0{,}176$.

Συμπεραίνουμε λοιπόν ότι εξαιτίας της αρκετά χαμηλής τιμής της Spearman correlation δεν μπορούμε να πούμε ότι υπάρχει κάποια συσχέτιση μεταξύ του IQ ενός ατόμου και των εβδομαδιαίων ωρών που παρακολουθεί αυτό τηλεόραση.