ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΜΕΤΑΔΟΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ
ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΥΛΙΚΩΝ

Συγκριτική Μελέτη Μεθόδων Κατηγοριοποίησης σε Ιατρικά Δεδομένα

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Χρυσούλα Χ. Χαντζή

Επιβλέπων : Γεώργιος Κ. Ματσόπουλος
Επίκουρος Καθηγητής Ε.Μ.Π.

Αθήνα, Νοέμβριος 2011

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΜΕΤΑΔΟΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ
ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΥΛΙΚΩΝ

Συγκριτική Μελέτη Μεθόδων Κατηγοριοποίησης σε Ιατρικά Δεδομένα

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Χρυσούλα Χ. Χαντζή

Επιβλέπων : Γεώργιος Κ. Ματσόπουλος
Επίκουρος Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 28η Νοεμβρίου 2011.

.....	.....	.....
Γεώργιος Ματσόπουλος	Νικόλαος Ουζούνογλου	Δημήτριος Κουτσούρης
Επίκουρος Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Αθήνα, Νοέμβριος 2011

.....
Χρυσούλα Χ. Χαντζή

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Copyright © Χρυσούλα Χ. Χαντζή, 2011.
Με επιφύλαξη παντός δικαιώματος. All rights reserved.

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τον συγγραφέα.

Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν τις επίσημες θέσεις του Εθνικού Μετσόβιου Πολυτεχνείου.

Περίληψη

Ο σκοπός της παρούσας εργασίας είναι η μελέτη και η αξιολόγηση της εφαρμογής διαφόρων μεθόδων κατηγοριοποίησης σε ιατρικά δεδομένα προκειμένου να εξεταστεί κατά πόσο είναι δυνατόν δοθέντος ενός συνόλου δεδομένων να γίνει ασφαλής διάγνωση κάποιας ασθένειας με αυτόματο τρόπο. Για το σκοπό αυτό αντλήθηκαν από την βάση δεδομένων UCI δεδομένα από διάφορες διαγνωστικές ιατρικές εξετάσεις τα οποία φέρουν τον χαρακτηρισμό του ατόμου ως υγιές ή ασθενές, ο οποίος χρησιμοποιήθηκε για την αξιολόγηση των διαφόρων μεθόδων που χρησιμοποιήθηκαν.

Συγκεκριμένα, χρησιμοποιήθηκαν οι αλγόριθμοι επιβλεπόμενης αλλά και μη επιβλεπόμενης μάθησης. Στην κατηγορία της επιβλεπόμενης μάθησης χρησιμοποιήθηκαν τα Τεχνητά Νευρωνικά Δίκτυα (ANN), η Μηχανή Διανυσμάτων Υποστήριξης (SVM) και ο αλγόριθμος k Κοντινότερων Γειτόνων (kNN) ενώ στην κατηγορία της μη επιβλεπόμενης μάθησης χρησιμοποιήθηκαν οι Χάρτες Αυτο-Οργάνωσης (SOM) και ο Ασαφής c-Μέσος (FCM). Επιπλέον για την βελτίωση της απόδοσης των παραπάνω μεθόδων χρησιμοποιήθηκε και η μέθοδος επιλογής χαρακτηριστικών Σειριακής Εμπρόσθιας Μεταβλητής Επιλογής (SFFS) προκειμένου να αφαιρεθούν πλεονάζοντα χαρακτηριστικά των δεδομένων.

Η αξιολόγηση των αποτελεσμάτων έγινε με τη χρήση των στατιστικών μέτρων ακρίβεια (accuracy), ευαισθησία (sensitivity) και προσδιοριστικότητα (specificity) και της χαρακτηριστικής καμπύλης λειτουργίας (ROC).

Λέξεις Κλειδιά

Κατηγοριοποίηση, Επιλογή Χαρακτηριστικών, Τεχνητά Νευρωνικά Δίκτυα, Μηχανή Διανυσμάτων Υποστήριξης, k Κοντινότεροι Γείτονες Χάρτες Αυτο-Οργάνωσης, Ασαφής c-Μέσος, ,Σειριακή Εμπρόσθια Μεταβλητή Επιλογή.

Abstract

The scope of this thesis was the analysis and the evaluation of classification methods when applied on medical data in order to determine whether it is possible to diagnose a disease using machine learning. Therefore, a collection of data sets of diagnostic examinations was chosen from the UCI repository. The data sets contain the labels of the instances which are used to evaluate the performance the classifiers.

The assessed techniques were both supervised learning and unsupervised learning algorithms. As far as the supervised case is concerned, the employed methods were Artificial Neural Networks (ANN), Support Vector Machine (SVM) and k Nearest Neighbours (kNN) while for the unsupervised case there was made use of Self Organising Maps (SOM) and Fuzzy c-Means (FCM). Furthermore, in an attempt to optimise the classifier performance the Sequential Forward Floating Selection technique was applied so as to reduce the dimensionality of the data and remove redundant features.

The evaluation of the classification results was performed using the statistical measures accuracy, sensitivity and specificity while the Receiver Operating Characteristic (ROC) curve was also plotted.

Key Words

classification, pattern recognition, feature selection, Artificial Neural Networks, Support Vector Machine, k Nearest Neighbours, Self Organising Maps, Fuzzy c-Means, Sequential Forward Floating Selection

Ευχαριστίες

Στο σημείο αυτό θα ήθελα να ευχαριστήσω θερμά για τη συμβολή τους τον επιβλέποντα Επίκουρο Καθηγητή δόκτωρα Γέωργιο Ματσόπουλο καθώς και τον Καθηγητή Εφαρμογών δόκτωρα Παντελεήμονα Ασβεστά από το Τμήμα Ιατρικών Οργάνων του ΤΕΙ Αθηνών.

Επίσης θα ήθελα να ευχαριστήσω τους γονείς μου Χρήστο και Κατερίνα οι οποίοι με στήριξαν αμέριστα σε όλη τη διάρκεια των σπουδών μου.

Περιεχόμενα

1. ΕΙΣΑΓΩΓΗ

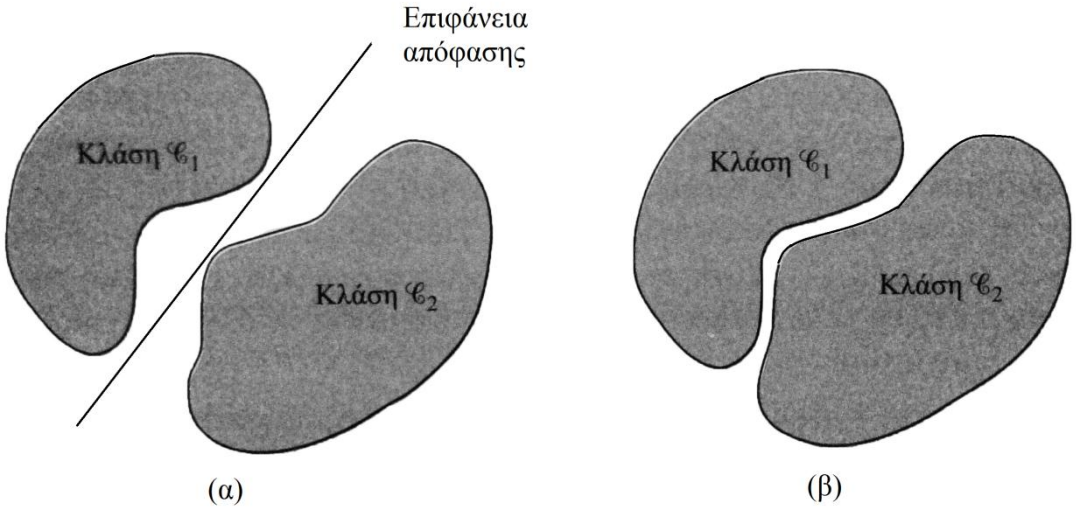
Η *αναγνώριση προτύπων (pattern recognition)* είναι ο επιστημονικός κλάδος ο οποίος στόχο έχει την περιγραφή και την κατάταξη αντικειμένων σε κατηγορίες ή *κλάσεις (classes)*. Ανάλογα με το είδος της εφαρμογής τα υπό κατάταξη αντικείμενα μπορεί να είναι οποιοδήποτε μετρήσιμο μέγεθος και αναφέρονται με τον γενικό όρο *πρότυπα (patterns)*. Η ανάπτυξη της αναγνώρισης προτύπων είναι ιδιαίτερα σημαντική τη σημερινή εποχή καθώς υπάρχει ανάγκη για περεταίρω αυτοματοποίηση διαδικασιών που έχει ήδη συντελεστεί σε μεγάλο βαθμό με την ευρύτατη χρήση των υπολογιστών. Η υποβοηθούμενη από υπολογιστή διάγνωση, η λήψη αποφάσεων με χρήση μηχανικής νοημοσύνης, η μηχανική όραση, η αναγνώριση φωνής και η αναγνώριση χαρακτήρων (αριθμών και γραμμάτων) είναι μερικές μόνο από τις ευρύτατες εφαρμογές της περιοχής της αναγνώρισης προτύπων.

Όπως προαναφέρθηκε, η αναγνώριση προτύπων συνίσταται στην *κατηγοριοποίηση (classification)* των αντικειμένων (προτύπων) στις κατηγορίες ή κλάσεις που καθορίζονται από τη φύση της εφαρμογής. Το πρώτο στάδιο της κατηγοριοποίησης είναι ο καθορισμός των μετρήσιμων μεγεθών που καθιστούν δυνατή την διάκριση των αντικειμένων στις επιθυμητές κατηγορίες οι οποίες είναι διακριτές μεταξύ τους. Τα μετρήσιμα αυτά μεγέθη αναφέρονται ως χαρακτηριστικά *(features).* Το σύνολο των χαρακτηριστικών που χρησιμοποιούνται για την κατηγοριοποίηση συγκροτεί το λεγόμενο *διάνυσμα χαρακτηριστικών (feature vector)*. Επομένως, αν μια εφαρμογή χρησιμοποιεί ℓ χαρακτηριστικά x_i, $i = 1,2, \dots, \ell$ το διάνυσμα χαρακτηριστικών που σχηματίζεται είναι το

$$\mathbf{x} = [x_1, x_2, \dots, x_\ell]^T$$

όπου το T συμβολίζει τον ανάστροφο πίνακα. Ένα διάνυσμα χαρακτηριστικών περιγράφει μοναδικά ένα συγκεκριμένο πρότυπο. Ακόμη, πρέπει να σημειωθεί ότι τα χαρακτηριστικά και το διάνυσμα χαρακτηριστικών μπορούν να χαρακτηριστούν ως τυχαίες μεταβλητές και τυχαίο διάνυσμα αντίστοιχα καθώς είναι προϊόντα μετρήσεων τυχαίων αντικειμένων.

Ο διαχωρισμός του χώρου του διανύσματος χαρακτηριστικών στις περιοχές που αντιστοιχούν σε κάθε κλάση γίνεται από τον λεγόμενο *κατηγοριοποιητή (classifier)* μέσω μίας διαδικασίας μηχανικής μάθησης. Το σύνορο που καθορίζει ο κατηγοριοποιητής μεταξύ των κλάσεων αναφέρεται ως *γραμμή απόφασης (decision line)*. Με αυτόν τον τρόπο, αν δοθεί ένα πρότυπο μαζί με το διάνυσμα χαρακτηριστικών του, ο κατηγοριοποιητής έχοντας καθορίσει τη γραμμή απόφασης κατατάσσει το πρότυπο αυτό στην κατάλληλη πλευρά της. Βέβαια, υπάρχει περίπτωση ο κατηγοριοποιητής να μην κατατάξει κάποιο αντικείμενο στη σωστή κλάση οπότε παρουσιάζεται *σφάλμα κατηγοριοποίησης (misclassification)*. Η γραμμή απόφασης μπορεί να είναι είτε μία ευθεία είτε μία καμπύλη, οπότε στην πρώτη περίπτωση πρόκειται για γραμμικό κατηγοριοποιητή ενώ σε αντίθετη περίπτωση ο κατηγοριοποιητής είναι μη γραμμικός. Αν το πρόβλημα είναι τέτοιο ώστε να υπάρχει τουλάχιστον μία ευθεία που να κατηγοριοποιεί σωστά όλα τα πρότυπα, τότε το σύνολο των προτύπων ονομάζεται *γραμμικά διαχωρίσιμο (linearly separable)* (σχήμα 1).



Σχήμα 1: **(α)** Ζεύγος γραμμικά διαχωρίσιμων προτύπων
(β) Ζεύγος μη γραμμικά διαχωρίσιμων προτύπων

Αν από το εξωτερικό περιβάλλον διατίθεται στο σύστημα ένα σύνολο προτύπων για τα οποία είναι γνωστή η κλάση στην οποία ανήκουν, η γνώση αυτή είναι χρήσιμη ώστε να καθοριστεί ακριβέστερα ο κατηγοριοποιητής. Στην περίπτωση αυτή τα πρότυπα αυτά αναφέρονται ως *πρότυπα εκπαίδευσης (training patterns)* και τα αντίστοιχα διανύσματα χαρακτηριστικών αναφέρονται ως *διανύσματα χαρακτηριστικών εκπαίδευσης (training feature vectors)* ενώ η μάθηση για την αναγνώριση προτύπων με αυτόν τον τρόπο ονομάζεται *επιβλεπόμενη (supervised)*. Βέβαια, δεν είναι πάντα διαθέσιμα τέτοια σύνολα οπότε η κατηγοριοποίηση βασίζεται στον εντοπισμό ομοιοτήτων μεταξύ των προτύπων προς κατηγοριοποίηση, οπότε η μάθηση ονομάζεται *μη επιβλεπόμενη (unsupervised)*. Ένας άλλος όρος που χρησιμοποιείται για τη διαδικασία αυτή είναι η *συσταδοποίηση (clustering)*. Η διάκριση ανάμεσα στις δύο αυτές μεθόδους θα αναλυθεί στο κεφάλαιο 2.

Στην πορεία προς την επιτυχή αναγνώριση προτύπων υπάρχουν κάποια σημεία ενδιαφέροντος που αξίζει να αναφερθούν. Αρχικά, ιδιαίτερα σημαντικός είναι ο καθορισμός και η *παραγωγή των χαρακτηριστικών (feature generation)* που θα χρησιμοποιηθούν για την κατηγοριοποίηση, διαδικασία που φαίνεται απλή σε κάποιες εφαρμογές, όμως σε ορισμένες άλλες μπορεί να αποτελέσει πολύ σύνθετη διαδικασία. Εν συνεχεία, αφού έχει παραχθεί ένας σημαντικός αριθμός χαρακτηριστικών πρέπει να καθοριστεί ποια ℓ το πλήθος χαρακτηριστικά από αυτά οδηγούν στο καλύτερο δυνατό αποτέλεσμα της κατηγοριοποίησης. Πρόκειται για τη διαδικασία της *επιλογής χαρακτηριστικών (feature selection)* και θα γίνει ειδική αναφορά σε αυτήν στο κεφάλαιο 3. Το επόμενο στάδιο που και το πλέον απαιτητικό είναι ο *σχεδιασμός του κατηγοριοποιητή (classifier design)*, όπου γίνεται η επιλογή ενός κριτηρίου βελτιστότητας και ενός τύπου μη-γραμμικότητας προκειμένου να πραγματοποιηθεί ληφθεί για κάθε κλάση το σωστό τμήμα του ℓ-διάστατου χώρου των χαρακτηριστικών. Τέλος, είναι απαραίτητο να αξιολογηθεί η απόδοση του συγκεκριμένου κατηγοριοποιητή, να διαπιστωθεί δηλαδή σε τι ποσοστό των περιπτώσεων ο κατηγοριοποιητής αποτυγχάνει, οπότε υπολογίζεται ο *ρυθμός σφαλμάτων κατηγοριοποίησης (classification error rate)*. Η αξιολόγηση των κατηγοριοποιητών θα παρουσιαστεί αναλυτικά στη συνέχεια αφού αποτελεί το άμεσο αντικείμενο της εργασίας.

2. ΔΙΑΔΙΚΑΣΙΕΣ ΜΑΘΗΣΗΣ

Στην εισαγωγή αναφερθήκαμε στην ανάγκη για μια διαδικασία μάθησης ώστε να μπορέσει ο κατηγοριοποιητής να καθορίσει την γραμμή απόφασης. Όπως συμβαίνει και με την διαδικασία μάθησης των ανθρώπων από το περιβάλλον τους, έτσι και η μηχανική μάθηση είναι δύο ειδών, μάθηση με εκπαιδευτή και μάθηση χωρίς εκπαιδευτή ή αλλιώς επιβλεπόμενη μάθηση και μη επιβλεπόμενη μάθηση. Η μέθοδος που ακολουθείται σε κάθε περίπτωση εξαρτάται από το αν υπάρχουν διαθέσιμα δεδομένα εκπαίδευσης. Αν υπάρχουν, τότε πρόκειται για την περίπτωση της επιβλεπόμενης μάθησης, ενώ σε αντίθετη περίπτωση η διαδικασία της μάθησης είναι μη επιβλεπόμενη.

2.1. Επιβλεπόμενη Μάθηση

Στην περίπτωση της επιβλεπόμενης μάθησης ή μάθησης με εκπαιδευτή, θεωρείται ότι ο εκπαιδευτής έχει γνώση του περιβάλλοντος και η γνώση αυτή αναπαρίσταται από ένα σύνολο ζευγών εισόδου-εξόδου, δηλαδή διανυσμάτων χαρακτηριστικών εκπαίδευσης. Αν το περιβάλλον τροφοδοτήσει τόσο τον εκπαιδευτή όσο και το σύστημα μάθησης με ένα διάνυσμα εκπαίδευσης τότε λόγω της γνώσης του ο εκπαιδευτής μπορεί να παράξει την επιθυμητή έξοδο δεδομένου του συγκεκριμένου διανύσματος.

Η επιθυμητή αυτή έξοδος είναι ουσιαστικά και η βέλτιστη έξοδος του κατηγοριοποιητή, επομένως υπό την συνδυασμένη επήρεια του διανύσματος εκπαίδευσης και ενός διανύσματος σφάλματος οι παράμετροι του συστήματος προσαρμόζονται κατάλληλα ώστε μετά από την επαναληπτική εφαρμογή της μεθόδου αυτής ο κατηγοριοποιητής να έχει την ίδια συμπεριφορά με τον εκπαιδευτή. Το διάνυσμα σφάλματος ορίζεται ως η διαφορά μεταξύ της επιθυμητής κατηγοριοποίησης και της πραγματικής κατηγοριοποίησης που εκτελεί ο κατηγοριοποιητής. Με τον τρόπο αυτό η γνώση του εκπαιδευτή μεταφέρεται στον κατηγοριοποιητή ώστε μετά το πέρας της εκπαίδευσης ο κατηγοριοποιητής να είναι ικανός να κατατάξει τα επόμενα πρότυπα που θα δεχθεί στη σωστή κλάση. Όπως αφήνει να εννοηθεί η παραπάνω περιγραφή, η επιβλεπόμενη μάθηση συνιστά ένα σύστημα κλειστού βρόχου με ανάδραση, στο οποίο όμως δεν περιλαμβάνεται το περιβάλλον το οποίο είναι άγνωστο προς τον κατηγοριοποιητή.

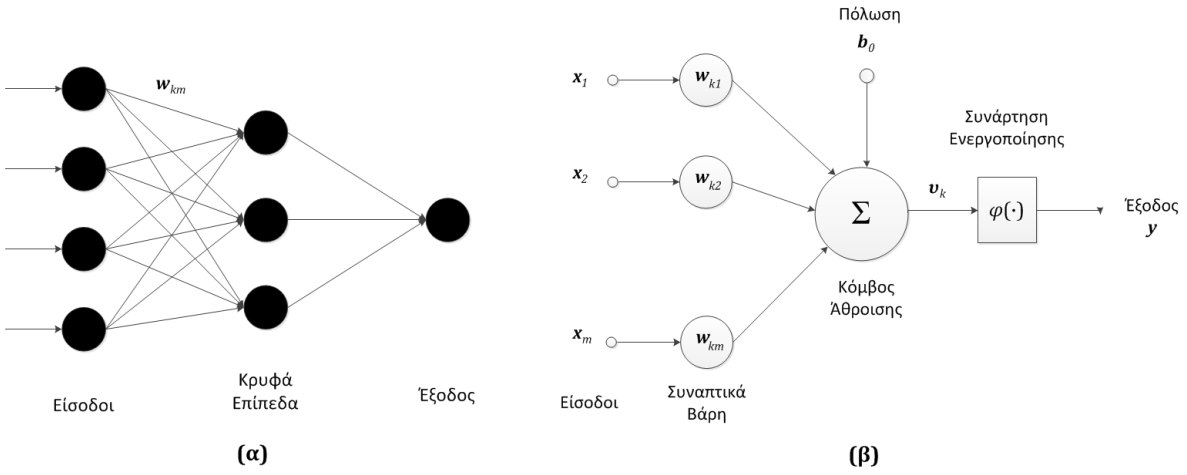
Έτσι, ένας κατηγοριοποιητής που χρησιμοποιεί επιβλεπόμενη μάθηση είναι ικανός με τη χρήση κατάλληλων αλγορίθμων να προσεγγίζει τη λειτουργία του εκπαιδευτή με την προσέγγιση να βελτιώνεται όσο αυξάνεται το σύνολο των διανυσμάτων εκπαίδευσης. Τέτοιοι αλγόριθμοι που χρησιμοποιούνται ευρέως είναι ο αλγόριθμος Perceptron, ο αλγόριθμος Ελάχιστου Μέσου Τετραγωνικού Σφάλματος (Least Mean Square, LMS), οι Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines, SVM), ο αλγόριθμος Οπισθοδιάδοσης (Back Propagation) και ο αλγόριθμος k-Κοντινότερων Γειτόνων (k-Nearest Neighbors, k-NN). Από τους αλγορίθμους αυτούς θα αναλυθούν και θα αξιολογηθούν στα πλαίσια αυτής της εργασίας ο αλγόριθμος Οπισθοδιάδοσης (Back Propagation), οι Μηχανές Διανυσμάτων Υποστήριξης (SVM) και ο αλγόριθμος k-Κοντινότερων Γειτόνων (k-NN).

2.1.1. Τεχνητά Νευρωνικά Δίκτυα (Artificial Neural Networks, ANN)

Εισαγωγή

Τα *Τεχνητά Νευρωνικά Δίκτυα (Artificial Neural Networks)* είναι μη γραμμικές δομές των οποίων η λειτουργία είναι εμπνευσμένη από τον ανθρώπινο εγκέφαλο. Είναι ισχυρά εργαλεία για τη διαδικασία της μοντελοποίησης, ειδικά σε περιπτώσεις που η σχέση μεταξύ των υποκείμενων δεδομένων είναι άγνωστη. Τα Τεχνητά Νευρωνικά Δίκτυα (ΤΝΔ) έχουν την ικανότητα να αναγνωρίζουν και να θυμούνται συσχετισμένα πρότυπα ανάμεσα σε ένα σύνολο δεδομένων εισόδου και συγκεκριμένες αντίστοιχες τιμές. Μετά τη διαδικασία της εκπαίδευσης, τα νευρωνικά δίκτυα μπορούν να χρησιμοποιηθούν για να προβλέψουν το αποτέλεσμα ενός νέου ανεξάρτητου συνόλου δεδομένων. Τα νευρωνικά δίκτυα μιμούνται την διαδικασία εκπαίδευσης του ανθρώπινου εγκεφάλου και μπορούν να επεξεργαστούν προβλήματα που περιλαμβάνουν μη γραμμικά και σύνθετα δεδομένα ακόμα κι αν τα δεδομένα είναι ανακριβή ή έχουν θόρυβο.

Τα νευρωνικά δίκτυα έχουν γίνει επίκεντρο πολλών ερευνών χάρη στο ευρύ φάσμα εφαρμογών όπου μπορούν να χρησιμοποιηθούν αλλά και λόγω της ικανότητας τους να αντιμετωπίσουν πολύπλοκα προβλήματα με ευκολία. Τα νευρωνικά δίκτυα είναι παράλληλα υπολογιστικά μοντέλα που αποτελούνται από πυκνά διασυνδεδεμένες προσαρμοστικές μονάδες επεξεργασίας. Ένα πολύ σημαντικό χαρακτηριστικό αυτών των δικτύων είναι η προσαρμοστικότητά τους, αφού ο προγραμματισμός αντικαθίσταται από τη γνώση μέσω του παραδείγματος, ιδιότητα πολύ χρήσιμη σε περιπτώσεις που το πρόβλημα δεν είναι πλήρως κατανοητό αλλά διατίθενται δεδομένα εκπαίδευσης.



Σχήμα 2: Μη γραμμικό μοντέλο νευρώνα. **(α)** Σχηματική αναπαράσταση **(β)** Μαθηματική αναπαράσταση

Ένα νευρωνικό δίκτυο είναι μια υπολογιστική δομή η οποία είναι εμπνευσμένη από την διαδικασία που έχει παρατηρηθεί στα φυσικά δίκτυα των βιολογικών νευρώνων στον

εγκέφαλο. Αποτελείται από απλές υπολογιστικές μονάδες που ονομάζονται *νευρώνες* οι οποίοι διασυνδέονται μεταξύ τους με πολύπλοκο τρόπο. Το σχήμα 2 παρουσιάζει το μοντέλο ενός νευρώνα που αποτελεί τη βάση για τη σχεδίαση μιας μεγάλης οικογένειας νευρωνικών δικτύων. Τα τρία βασικά στοιχεία του μοντέλου νευρώνα που διακρίνονται στο σχήμα 2β είναι τα εξής:

1. Ένα σύνολο *συνάψεων* οι οποίες αντιστοιχούν στην ιδέα των συνάψεων των νευρικών κυττάρων. Κάθε σύναψη χαρακτηρίζεται από ένα *βάρος*. Ένα σήμα εισόδου x_j στην είσοδο της σύναψης j που συνδέεται με το νευρώνα k πολλαπλασιάζεται επί το συναπτικό βάρος w_{kj}. Σε αντίθεση με τον ανθρώπινο εγκέφαλο, το συναπτικό βάρος ενός νευρώνα λαμβάνει και αρνητικές και θετικές τιμές.
2. Έναν *αθροιστή (adder)* για την άθροιση των σημάτων εισόδου, σταθμισμένων από τα αντίστοιχα συναπτικά βάρη του νευρώνα. Ουσιαστικά πρόκειται για έναν γραμμικό συνδυαστή (linear combiner).
3. Μια *συνάρτηση ενεργοποίησης (activation function)* για τον περιορισμό του πλάτους του σήματος εξόδου ενός νευρώνα. Η συνάρτηση ενεργοποίησης αναφέρεται επίσης και ως συνάρτηση περιορισμού (squashing function) επειδή περιορίζει το επιτρεπτό εύρος πλάτους του σήματος εξόδου σε κάποια πεπερασμένη τιμή. Συνήθως το κανονικοποιημένο εύρος τιμών πλάτους της εξόδου ενός νευρώνα γράφεται ως μοναδιαίο κλειστό διάστημα με τη μορφή [0,1] ή [-1,1].

Στο μοντέλο του νευρώνα του σχήματος 2 περιλαμβάνεται επίσης και μία εξωτερικά επιβαλλόμενη πόλωση b_k η οποία χρησιμεύει στην αύξηση ή τη μείωση της δικτυακής διέγερσης της συνάρτησης ενεργοποίησης ανάλογα με το αν είναι θετική ή αρνητική αντίστοιχα.

Η μαθηματική περιγραφή του μοντέλου ενός νευρώνα k γίνεται με τις δύο παρακάτω εξισώσεις

$$u_k = \sum_{j=1}^{m} w_{kj} x_j \tag{1}$$

$$y_k = \varphi(u_k + b_k) \tag{2}$$

όπου $x_1, x_2, \dots, x_m$ τα σήματα εισόδου, $w_{k1}, w_{k2}, \dots, w_{km}$ τα αντίστοιχα συναπτικά βάρη του νευρώνα k, u_k η έξοδος του γραμμικού συνδυαστή (δεν παρουσιάζεται στο σχήμα), b_k η πόλωση, $\varphi(\cdot)$ η συνάρτηση ενεργοποίησης και y_k το σήμα εξόδου του νευρώνα. Η χρήση της πόλωσης έχει ως αποτέλεσμα τη χρήση ενός αφινικού μετασχηματισμού στην έξοδο u_k του γραμμικού συνδυαστή σύμφωνα με τη σχέση

$$v_k = u_k + b_k \tag{3}$$

Η συνάρτηση ενεργοποίησης η οποία συμβολίζεται ως $\varphi(\cdot)$, ορίζει την έξοδο ενός νευρώνα βάσει του τοπικού πεδίου v. Δύο βασικοί τύποι συναρτήσεων ενεργοποίησης είναι η *συνάρτηση κατωφλίου (threshold function)* και η *σιγμοειδής συνάρτηση (sigmoid function)* που απεικονίζονται στο σχήμα 3.

Η συνάρτηση κατωφλίου (σχήμα 3α) είναι η γνωστή και ως βηματική συνάρτηση ή συνάρτηση Heaviside ορίζεται ως εξής

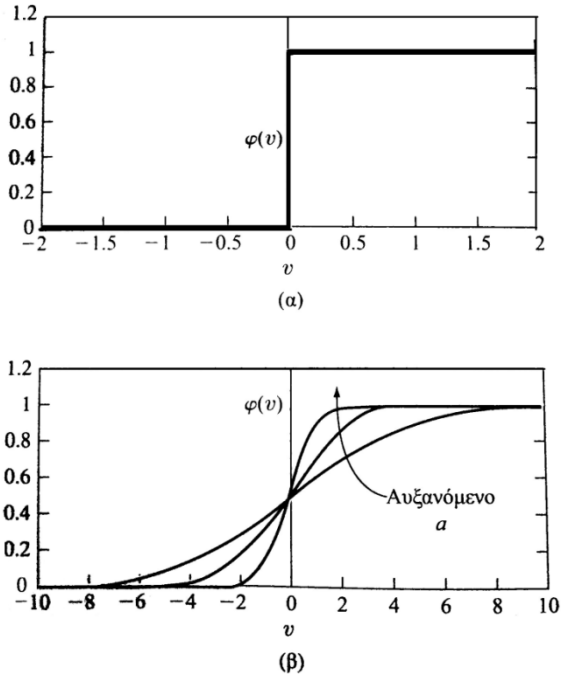
$$\varphi(\upsilon) = \begin{cases} 1\,, & \text{αν } \upsilon \geq 0 \\ 0\,, & \text{αν } \upsilon < 0 \end{cases} \tag{4}$$

Η έξοδος νευρώνα k που χρησιμοποιεί μια τέτοια συνάρτηση κατωφλίου προκύπτει

$$y_k = \begin{cases} 1\,, & \text{αν } \upsilon_k \geq 0 \\ 0\,, & \text{αν } \upsilon_k < 0 \end{cases} \tag{5}$$

όπου υ_k είναι το τοπικό πεδίο νευρώνα για το οποίο ισχύει

$$\upsilon_k = \sum_{j=1}^{m} w_{kj} x_j + b_k \tag{6}$$



Σχήμα 3: Συναρτήσεις Ενεργοποίησης
(α) Συνάρτηση Κατωφλίου
(β) Σιγμοειδής συνάρτηση για μεταβαλλόμενη παράμετρο κλίσης a

Το μοντέλο νευρώνα που περιγράφουν οι παραπάνω σχέσεις αναφέρεται ως *μοντέλο McCulloch-Pitts* (1946) Σύμφωνα με το μοντέλο αυτό, η έξοδος ενός νευρώνα λαμβάνει την τιμή 1 αν το τοπικό πεδίο του συγκεκριμένου νευρώνα είναι μη αρνητικό και 0 σε κάθε άλλη περίπτωση.

Η σιγμοειδής συνάρτηση της οποίας η γραφική παράσταση θυμίζει το λατινικό γράμμα «S» όπως φαίνεται στο σχήμα 3β είναι η πιο κοινή συνάρτηση που χρησιμοποιείται στα νευρωνικά δίκτυα. Ορίζεται ως γνησίως αύξουσα συνάρτηση και περιλαμβάνει τμήματα τόσο γραμμικής όσο και μη γραμμικής συμπεριφοράς. Επίσης είναι διαφορίσιμη, ιδιότητα ιδιαίτερα χρήσιμη για τα νευρωνικά δίκτυα. Επιπλέον, σε αντίθεση με τη συνάρτηση κατωφλίου, η σιγμοειδής συνάρτηση λαμβάνει τιμές στο συνεχές διάστημα [0,1]. Παράδειγμα σιγμοειδούς συνάρτησης είναι η *λογιστική συνάρτηση* η οποία ορίζεται ως

$$\varphi(\upsilon) = \frac{1}{1 + \exp(-\alpha\upsilon)} \tag{7}$$

όπου α η *παράμετρος κλίσης* της συνάρτησης. Μεταβάλλοντας την τιμή της παραμέτρου αυτής μεταβάλλεται και η κλίση της καμπύλης στην αρχή των αξόνων. Καθώς η

παράμετρος α τείνει στο άπειρο η καμπύλη προσεγγίζει τη συνάρτηση κατωφλίου. Άλλη μία ευρέως χρησιμοποιούμενη σιγμοειδής συνάρτηση είναι η συνάρτηση υπερβολικής εφαπτομένης:

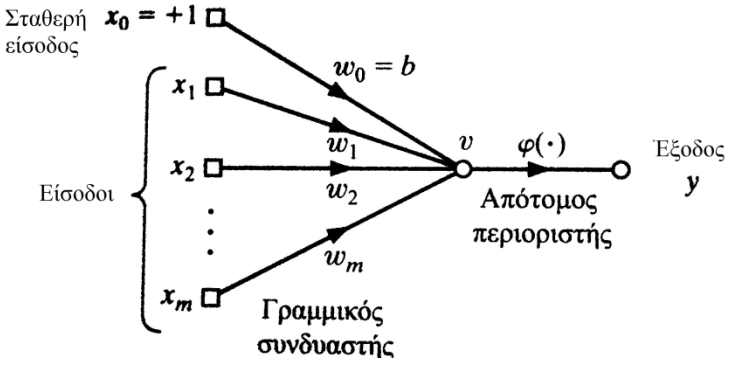
$$\varphi(v) = \tanh v \tag{8}$$

Επιπλέον, σε αντίθεση με τη συνάρτηση κατωφλίου, η σιγμοειδής συνάρτηση λαμβάνει τιμές στο συνεχές διάστημα [0,1].

Perceptron

Το *Perceptron* του Rosenblatt (1958) είναι το πρώτο νευρωνικό δίκτυο που είχε αλγοριθμική περιγραφή. Είναι η απλούστερη δυνατή μορφή ενός νευρωνικού δικτύου που χρησιμοποιείται για την κατηγοριοποίηση γραμμικά διαχωρίσιμων προτύπων. Ουσιαστικά αποτελείται από έναν και μόνο νευρώνα με προσαρμόσιμα βάρη και πόλωση. Υπό αυτό το πρίσμα, για τον αλγόριθμο που χρησιμοποιείται για την προσαρμογή αυτών των ελεύθερων παραμέτρων αποδεικνύεται ότι αν τα πρότυπα που χρησιμοποιούνται προέρχονται από δύο γραμμικά διαχωρίσιμες κλάσεις ο αλγόριθμος του perceptron συγκλίνει και τοποθετεί τη διαχωριστική επιφάνεια απόφασης με τη μορφή ενός υπερεπιπέδου μεταξύ των δύο κλάσεων.

Στο perceptron χρησιμοποιείται ένας μη γραμμικός νευρώνας σύμφωνα με το μοντέλο νευρώνα McCulloch-Pitts. Η απεικόνισή του με τη μορφή γραφήματος ροής φαίνεται στο σχήμα 4. Το μοντέλο αυτό είναι ισοδύναμο με αυτό του σχήματος 2β, με μόνη τροποποίηση ότι η πόλωση $b(n)$ αντιμετωπίζεται ως ένα συναπτικό βάρος το οποίο οδηγείται από σταθερή είσοδο +1.



Σχήμα 4:
Ισοδύναμο γράφημα ροής σήματος του perceptron. Χάρην σαφήνειας παραλείπεται ο χρόνος

Σύμφωνα με το μοντέλο αυτό ορίζουμε το διάνυσμα εισόδων διαστάσεων $(m+1) \times 1$ ως

$$\mathbf{x}(n) = [+1, x_1(n), x_2(n), \dots, x_m(n)]^T$$

όπου n το χρονικό βήμα εφαρμογής του αλγορίθμου. Αντίστοιχα ορίζεται το διάνυσμα βαρών διαστάσεων $(m+1) \times 1$ ως

$$\mathbf{w}(n) = [b, w_1(n), w_2(n), \dots, w_m(n)]^T$$

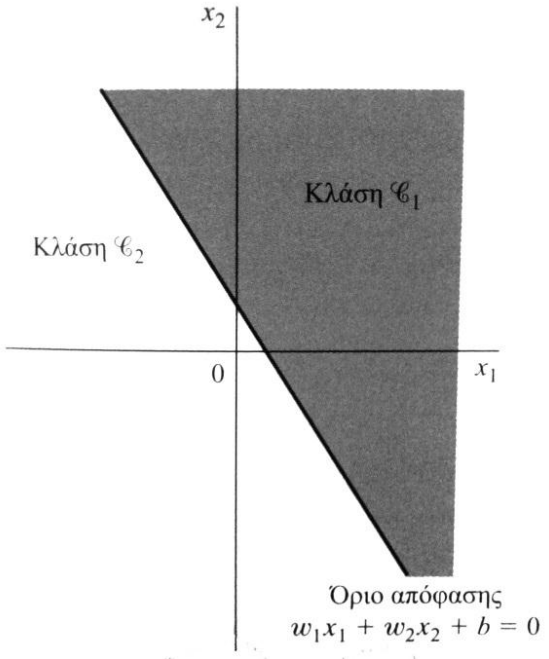
Επομένως η έξοδος του γραμμικού συνδυαστή, η οποία είναι και η είσοδος του απότομου περιοριστή ή τοπικό πεδίο, μπορεί να γραφεί σε συμπαγή μορφή

$$v(n) = \sum_{i=0}^{m} w_i(n) x_i(n) = \mathbf{w}^T(n)\mathbf{x}(n) \tag{9}$$

όπου $w_0(n) = b$. Στη συνέχεια, ο απότομος περιοριστής παράγει έξοδο ίση με +1 άν το τοπικό πεδίο $v(n)$ είναι θετικό, οπότε το πρότυπο κατατάσσεται στην κλάση $\mathcal{C}_1$ και -1 αν είναι αρνητικό, οπότε το πρότυπο κατατάσσεται στην κλάση $\mathcal{C}_2$.

Για σταθερό n η εξίσωση $\mathbf{w}^T\mathbf{x} = 0$ αν απεικονιστεί σε ένα m-διάστατο χώρο με συντεταγμένες $x_1, x_2, \ldots, x_m$ ορίζει το υπερεπίπεδο που δρα ως διαχωριστική επιφάνεια μεταξύ των δύο κλάσεων εισόδου $\mathcal{C}_1$ και $\mathcal{C}_2$. Ένα παράδειγμα σε χώρο δύο διαστάσεων φαίνεται στο σχήμα 5. Δοθέντος ενός συνόλου εκπαίδευσης $\mathcal{T}$, ο σκοπός της κατηγοριοποίησης είναι να βρεθεί το κατάλληλο διάνυσμα βαρών $\mathbf{w}$ ώστε να ισχύουν:

$$\begin{aligned} &\mathbf{w}^T\mathbf{x} > 0 && \text{για τα διανύσματα εισόδων } \mathbf{x} \text{ που ανήκουν στην κλάση } \mathcal{C}_1 \\ &\mathbf{w}^T\mathbf{x} \leq 0 && \text{για τα διανύσματα εισόδων } \mathbf{x} \text{ που ανήκουν στην κλάση } \mathcal{C}_2 \end{aligned} \tag{10}$$



Σχήμα 5:
Το υπερεπίπεδο ως όριο απόφασης για ένα πρόβλημα ταξινόμησης προτύπων σε δύο κλάσεις σε χώρο διαστάσεων (για σταθερό χρόνο)

Διατυπώνεται επομένως ο αλγόριθμος προσαρμογής του διανύσματος βαρών του perceptron ως εξής:

1. Εάν το n-οστό μέλος του συνόλου εκπαίδευσης $\mathbf{x}(n)$ ταξινομείται σωστά από το διάνυσμα βαρών $\mathbf{w}(n)$ που υπολογίζεται στη n-οστή επανάληψη του αλγορίθμου, δεν γίνεται καμία διόρθωση στο διάνυσμα βαρών του perceptron σύμφωνα με τον κανόνα

$$\begin{aligned} &\mathbf{w}(n+1) = \mathbf{w}(n) && \text{Αν } \mathbf{w}^T\mathbf{x} > 0 \text{ και } \mathbf{x}(n) \text{ ανήκει στην κλάση } \mathcal{C}_1 \\ &\mathbf{w}(n+1) = \mathbf{w}(n) && \text{Αν } \mathbf{w}^T\mathbf{x} \leq 0 \text{ και } \mathbf{x}(n) \text{ ανήκει στην κλάση } \mathcal{C}_2 \end{aligned} \tag{11}$$

2. Ειδάλλως, το διάνυσμα βαρών ενημερώνεται σύμφωνα με τον κανόνα

$$\begin{aligned} &\mathbf{w}(n+1) = \mathbf{w}(n) - \eta(n)\mathbf{x}(n) && \text{Αν } \mathbf{w}^T\mathbf{x} > 0 \text{ και } \mathbf{x}(n) \in \mathcal{C}_1 \\ &\mathbf{w}(n+1) = \mathbf{w}(n) - \eta(n)\mathbf{x}(n) && \text{Αν } \mathbf{w}^T\mathbf{x} \leq 0 \text{ και } \mathbf{x}(n) \in \mathcal{C}_2 \end{aligned} \tag{12}$$

όπου η παράμετρος $\eta(n)$ ονομάζεται *ρυθμός μάθησης* και ελέγχει την προσαρμογή που εφαρμόζεται στο διάνυσμα βαρών στην επανάληψη n.

Αν $\eta(n) = \eta > 0$, όπου η σταθερά ανεξάρτητη του αριθμού επανάληψης n, τότε πρόκειται για την περίπτωση του λεγόμενου *κανόνα προσαρμογής μέσω σταθερής αύξησης*.

Αποδεικνύεται με αυτά τα δεδομένα ότι ο αλγόριθμος με σταθερή αύξηση συγκλίνει σε μία λύση. Το θεώρημα σύγκλισης με σταθερή αύξηση διατυπώνεται ως εξής (Rosenblatt, 1962):

> *Έστω ότι τα υποσύνολα των διανυσμάτων εκπαίδευσης $\mathcal{T}_1$ και $\mathcal{T}_2$ είναι γραμμικά διαχωρίσιμα. Έστω ότι οι είσοδοι που παρουσιάζονται στο perceptron προέρχονται από αυτά τα δύο υποσύνολα. Το perceptron συγκλίνει μετά από κάποιο αριθμό n_0 επαναλήψεων υπό την έννοια ότι το*
>
> $$\boldsymbol{w}(n_0) = \boldsymbol{w}(n_0 + 1) = \boldsymbol{w}(n_0 + 2) = \cdots$$
>
> *είναι ένα διάνυσμα λύσεων για $n_0 \leq n_{max}$, όπου n_{max} ο αριθμός επαναλήψεων μετά από τις οποίες τερματίζει ο αλγόριθμος.*

Στον πίνακα 1 παρουσιάζεται συνοπτικά ο *αλγόριθμος σύγκλισης του perceptron* [Lippmann, 1987].

Πίνακας 1: Σύνοψη του Αλγορίθμου Σύγκλισης του Perceptron

Μεταβλητές και Παράμετροι:

$\mathbf{x}(n) = [+1, x_1(n), x_2(n), \ldots, x_m(n)]^T$ (διάνυσμα εισόδων)
$\mathbf{w}(n) = [b, w_1(n), w_2(n), \ldots, w_m(n)]^T$ (διάνυσμα βαρών)
$b =$ πόλωση
$y(n) =$ πραγματική απόκριση (κβαντισμένη)
$d(n) =$ επιθυμητή απόκριση
$\eta =$ παράμετρος ρυθμού μάθησης, $0 < \eta < 1$

1. ***Αρχικοποίηση***. Θέσε $\mathbf{w}(0) = \mathbf{0}$. Στη συνέχεια, εκτέλεσε τους ακόλουθους υπολογισμούς για χρονικό βήμα $n = 1,2, \ldots$
2. ***Ενεργοποίηση***. Στο χρονικό βήμα $\boldsymbol{n}$ ενεργοποίησε το perceptron εφαρμόζοντας το διάνυσμα εισόδων $\mathbf{x}(n)$ και την επιθυμητή απόκριση $d(n)$.
3. ***Υπολογισμός της Πραγματικής Απόκρισης***. Υπολόγισε την πραγματική απόκριση του perceptron ως
   $$y(n) = \mathbf{sgn}[\mathbf{w}^T(n)\mathbf{x}(n)]$$
   όπου $\mathrm{sgn}(\cdot)$ η συνάρτηση προσήμου
4. ***Προσαρμογή του Διανύσματος Βαρών***. Ενημέρωσε το διάνυσμα βαρών του perceptron ώστε να καταλήξεις στο
   $$\mathbf{w}(n+1) = \mathbf{w}(n) + \eta[d(n) - y(n)]\mathbf{x}(n)$$
   όπου $$d(n) = \begin{cases} +1, & \text{αν } \mathbf{x}(n) \in \mathcal{C}_1 \\ -1, & \text{αν } -\mathbf{x}(n) \in \mathcal{C}_2 \end{cases}$$
5. *Συνέχιση.* Αύξησε το χρονικό βήμα n κατά 1 και επίστρεψε στο βήμα 2.

Στην μέχρι τώρα ανάλυση δεν έχει γίνει αναφορά σε σε κάποια συνάρτηση κόστους. Σε αυτό το σημείο θα γίνει αυτή η προσθήκη ενώ θα διατυπωθεί μία μέθοδος *μαζικής (batch)* κατηγοριοποίησης. Η έννοια της μαζικότητας έγκειται στο γεγονός ότι σε κάθε χρονικό βήμα χρησιμοποιείται το σύνολο των εσφαλμένα ταξινομημένων δειγμάτων για τον υπολογισμό της προσαρμογής.

Η *συνάρτηση κόστους του perceptron* ορίζεται ως

$$J(\mathbf{w}) = \sum_{\mathbf{x}\in\mathcal{X}} -\mathbf{w}^T\mathbf{x} \tag{13}$$

όπου $\mathcal{X}$ είναι το σύνολο των δειγμάτων **x** που ταξινομούνται εσφαλμένα από ένα perceptron που το οποίο χρησιμοποιεί το **w** ως διάνυσμα βαρών. Εάν τα δείγματα ταξινομούνται σωστά τότε το σύνολο $\mathcal{X}$ είναι κενό οπότε και η συνάρτηση κόστους $J(\mathbf{w})$ είναι 0. Το σημαντικότερο χαρακτηριστικό της συνάρτησης κόστους $J(\mathbf{w})$ είναι ότι η συνάρτηση αυτή είναι διαφορίσιμη ως προς το διάνυσμα βαρών **w.** Επομένως διαφορίζοντας τη συνάρτηση $J(\mathbf{w})$ ως προς το διάνυσμα βαρών **w** προκύπτει το *διάνυσμα κλίσεων*

$$\nabla J(\mathbf{w}) = \sum_{\mathbf{x}\in\mathcal{X}} -\mathbf{x} \tag{14}$$

όπου ο τελεστής κλίσης

$$\nabla = \left[\frac{\partial}{\partial w_1}, \frac{\partial}{\partial w_2}, \dots, \frac{\partial}{\partial w_m}\right]^T \tag{15}$$

Σύμφωνα με τη *μέθοδο της πλέον απότομης κατάβασης (steepest descent),* η προσαρμογή που γίνεται στο διάνυσμα βαρών **w** σε κάθε χρονικό βήμα του αλγορίθμου εφαρμόζεται σε κατεύθυνση αντίθετη ως προς το διάνυσμα κλίσεων $\nabla J(\mathbf{w})$. Κατά συνέπεια ο αλγόριθμος παίρνει τη μορφή

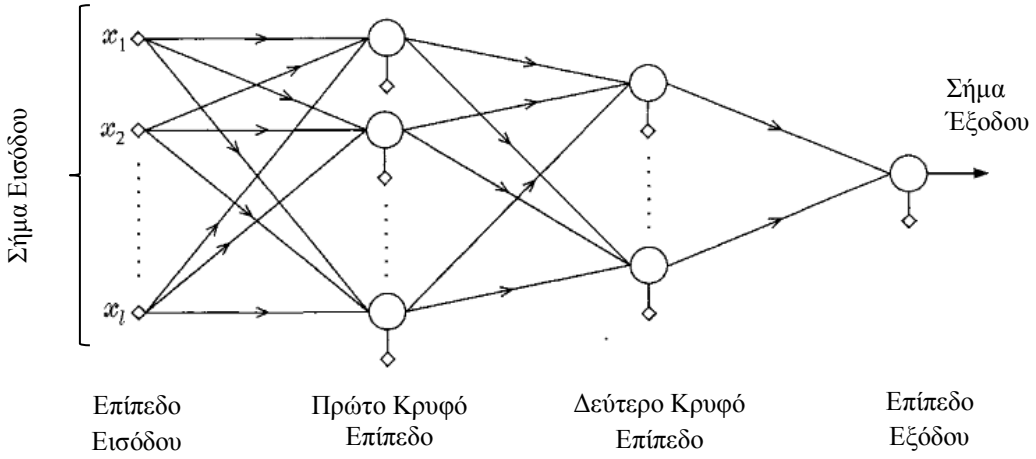
$$\mathbf{w}(n+1) = \mathbf{w}(n) - \eta(n)\nabla J(\mathbf{w}) = \mathbf{w}(n) + \eta(n)\sum_{\mathbf{x}\in\mathcal{X}} \mathbf{x} \tag{16}$$

Η μορφή αυτή περιλαμβάνει την περίπτωση του αλγορίθμου σύγκλισης για προσαρμογή βάσει ενός δείγματος που περιγράφηκε προηγουμένως ως ειδική περίπτωση. Σύμφωνα με την εξίσωση (16) που εκφράζει τον *μαζικό αλγόριθμο του perceptron*, το διάνυσμα βαρών στο βήμα n+1 του αλγόριθμου προσαρμόζεται μέσω του αθροίσματος των δειγμάτων που ταξινομούνται εσφαλμένα από το διάνυσμα βαρών $\mathbf{w}(n)$. Η επίδραση του αθροίσματος αυτού στην προσαρμογή καθορίζεται από την παράμετρο ρυθμού μάθησης $\eta(n)$.

Perceptron πολλών επιπέδων

Το perceptron που μελετήθηκε στην προηγούμενη ενότητα είναι το απλούστερο νευρωνικό δίκτυο, το οποίο όπως φαίνεται στην εικόνα 2α έχει μόνο ένα κρυφό επίπεδο νευρώνων. Οι δυνατότητες του δικτύου αυτού περιορίζονται στην λύση προβλήματος κατηγοριοποίησης όπου τα πρότυπα είναι γραμμικά διαχωρίσιμα. Σε πραγματικές συνθήκες όμως υπάρχει ανάγκη για κατηγοριοποίηση μη γραμμικά διαχωρίσιμων προτύπων. Για το σκοπό αυτό είναι απαραίτητη η εισαγωγή περεταίρω επιπέδων νευρώνων εσωτερικά του δικτύου, τα οποία παραμένουν *κρυφά* για τους κόμβους των επιπέδων εισόδου και εξόδου. Η δομή ενός τέτοιου δικτύου με δύο κρυφά επίπεδα και μία έξοδο φαίνεται στην εικόνα 6.

Οι νευρώνες εξόδου συγκροτούν το επίπεδο εξόδου του δικτύου. Οι υπόλοιποι νευρώνες συγκροτούν τα κρυφά επίπεδα του δικτύου. Οι κρυφές μονάδες δεν αποτελούν τμήμα ούτε της εισόδου είτε της εξόδου του δικτύου. Το πρώτο κρυφό επίπεδο τροφοδοτείται από το επίπεδο εισόδου του δικτύου το οποίο αποτελείται από αισθητηριακές μονάδες. Οι έξοδοι που παράγονται από το πρώτο κρυφό επίπεδο εφαρμόζονται στο επόμενο κρυφό επίπεδο και ούτω καθ' εξής για το υπόλοιπο του δικτύου.



Σχήμα 6: Αρχιτεκτονική ενός perceptron πολλών επιπέδων με δύο κρυφά επίπεδα νευρώνων και μία έξοδο

Κάθε κρυφός νευρώνας ή νευρώνας εξόδου ενός perceptron πολλών επιπέδων σχεδιάζεται ώστε να εκτελεί δύο υπολογισμούς. Πρώτον, εκτελεί τον υπολογισμό του σήματος που εμφανίζεται στην έξοδο κάθε νευρώνα του προηγούμενου επιπέδου (*λειτουργικό σήμα*) ο οποίος εκφράζεται ως μία συνεχής μη γραμμική συνάρτηση του σήματος εισόδου και των συναπτικών βαρών που σχετίζονται μ' αυτό τον νευρώνα. Δεύτερον, εκτελείται ο υπολογισμό μιας εκτίμησης του διανύσματος κλίσης, δηλαδή των κλίσεων της επιφάνειας σφάλματος σε σχέση με τα βάρη που είναι συνδεδεμένα στις εισόδους ενός νευρώνα. Ο υπολογισμός αυτός χρησιμοποιείται στη φάση εξέλιξης του δικτύου προς τα πίσω, τη φάση δηλαδή που τα συναπτικά βάρη προσαρμόζονται σύμφωνα με τα σφάλματα που έχουν υπολογιστεί.

Οι κρυφοί νευρώνες στο perceptron πολλών επιπέδων δρουν ως ανιχνευτές χαρακτηριστικών (feature detectors). Καθώς προχωρά η διαδικασία μάθησης οι κρυφοί νευρώνες προσδιορίζουν τα χαρακτηριστικά εκείνα των προτύπων εκπαίδευσης που είναι αυτά που είναι καθοριστικότερα στην διαδικασία της κατηγοριοποίησης. Αυτό επιτυγχάνεται μετασχηματίζοντας γραμμικά τα δεδομένα εισόδου σε ένα νέο χώρο, τον *χώρο χαρακτηριστικών*. Με αυτόν τον μετασχηματισμό ξεχωρίζουν τα χαρακτηριστικά που ενδιαφέρουν την διαδικασία της κατηγοριοποίησης από τυχόν περιττές πληροφορίες που τίθενται στην είσοδο. Αυτή ακριβώς η ιδιότητα είναι που καθιστά το perceptron πολλών επιπέδων ικανό να λύσει το μη γραμμικά διαχωρίσιμο πρόβλημα κατηγοριοποίησης.

Ανάλογα με το πώς εκτελείται η επιβλεπόμενη μάθηση στο perceptron πολλών επιπέδων έχουν αναπτυχθεί δύο στρατηγικές, η *μαζική μάθηση (batch)* και η *on-line μάθηση*. Το χαρακτηριστικό της μαζικής μάθησης είναι ότι οι προσαρμογές στα συναπτικά βάρη εκτελούνται μετά την επεξεργασία του συνόλου των προτύπων του συνόλου εκπαίδευσης. Από την άλλη πλευρά, στην on-line μάθηση τα συναπτικά βάρη προσαρμόζονται παράδειγμα προς παράδειγμα, δηλαδή σε κάθε εμφάνιση κάποιου προτύπου του συνόλου εκπαίδευσης.

Αλγορίθμος Οπισθοδιάδοσης (Back Propagation, ΒΚ)

Η μέθοδος της on-line μάθησης για την επιβλεπόμενη εκπαίδευση έγινε πολύ δημοφιλής χάρη και στην ανάπτυξη του αλγορίθμου *Οπισθοδιάδοσης (Back Propagation, ΒΚ)* τον οποίο θα αναλύσουμε εις βάθος στη συνέχεια.

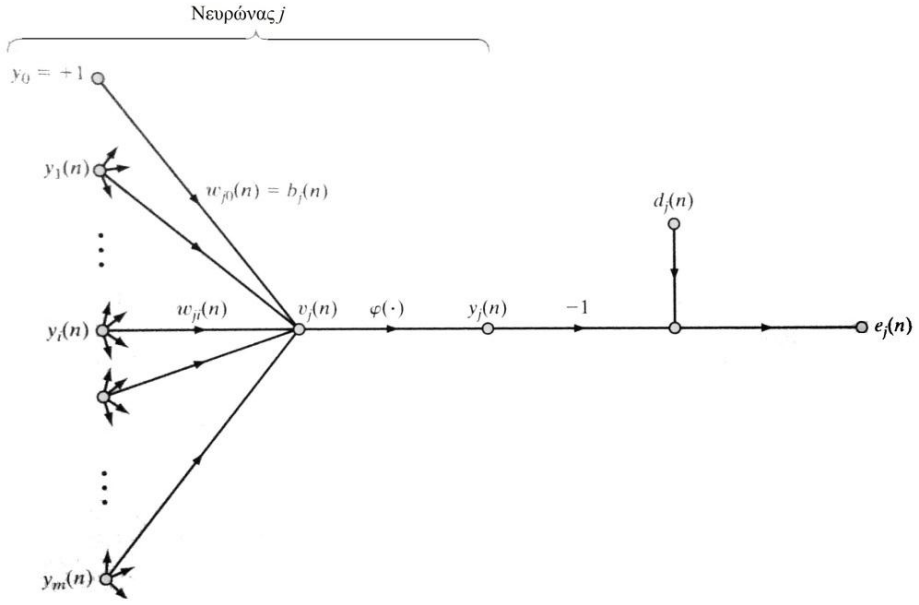
Στο σχήμα 7 φαίνονται οι συμβολισμοί που θα χρησιμοποιηθούν για την περιγραφή του αλγορίθμου. Έστω ότι διατίθεται ένα σύνολο εκπαίδευσης $\mathcal{T} = \{\mathbf{x}(n), \mathbf{d}(n)\}_{n=1}^{N}$. Αν με $y_j(n)$ συμβολίζεται το λειτουργικό σήμα που παράγεται στην έξοδο του νευρώνα j στο επίπεδο εξόδου από το ερέθισμα $\mathbf{x}(n)$ που εφαρμόζεται στο επίπεδο εισόδου, τότε αυτό αρχικά τουλάχιστον θα είναι διαφορετικό από την επιθυμητή έξοδο $\mathbf{d}(n)$ οπότε ορίζεται το σήμα σφάλματος που παράγεται στην έξοδο του νευρώνα j ως

$$e_j(n) = d_j(n) - y_j(n) \tag{17}$$

όπου $d_j(n)$ είναι το j-οστό στοιχείο του διανύσματος επιθυμητών εξόδων $\mathbf{d}(n)$. Ο σκοπός του αλγορίθμου ΒΚ είναι έχοντας καθορίσει μία κατάλληλη για το πρόβλημα συνάρτηση κόστους J που να εξαρτάται από τα $\mathbf{d}(n)$ και $\boldsymbol{y}(n)$, να προσαρμόζονται τα συναπτικά βάρη του δικτύου ώστε η συνάρτηση κόστους J να ελαχιστοποιείται. Στη συγκεκριμένη περίπτωση η συνάρτηση κόστους που είναι κατάλληλη είναι η συνολική στιγμιαία ενέργεια σφάλματος ολόκληρου του δικτύου, δηλαδή η συνάρτηση:

$$\mathcal{E}(n) = \frac{1}{2}\sum_{j \in C} e_j^2(n) \tag{18}$$

όπου C το σύνολο των νευρώνων του επιπέδου εξόδου.



Σχήμα 7: Γράφημα ροής σήματος που περιγράφει τις λεπτομέρειες του νευρώνα εξόδου *j*

Το τοπικό πεδίο $v_j(n)$ που παράγεται στην είσοδο της συνάρτησης ενεργοποίησης φ_j και σχετίζεται με το νευρώνα j είναι:

$$v_j(n) = \sum_{i=0}^{m} w_{ji}(n) y_i(n) \tag{19}$$

όπου m είναι το πλήθος των εισόδων (εξαιρουμένης της πόλωσης) που εφαρμόζονται στον νευρώνα j. Το συναπτικό βάρος w_{j0} αντιστοιχεί στην σταθερή είσοδο και είναι ίση με την πόλωση b_j που εφαρμόζεται στο νευρώνα j. Επομένως το λειτουργικό σήμα που λαμβάνεται στην έξοδο $y_i(n)$ του νευρώνα j κατά την επανάληψη n είναι:

$$y_i(n) = \varphi_j\left(\upsilon_j(n)\right) \tag{20}$$

Η διόρθωση που $\Delta w_{ji}(n)$ εφαρμόζει στο συναπτικό βάρος $w_{ji}(n)$ ο αλγόριθμος ΒΚ σε κάθε επανάληψη είναι ανάλογη της μερικής παραγώγου $\partial\mathcal{E}(n)/\partial w_{ji}(n)$. Η μερική παράγωγος αυτή αντιπροσωπεύει ένα *συντελεστή ευαισθησίας* ο οποίος καθορίζει την κατεύθυνση της αναζήτησης στο χώρο βαρών για το συναπτικό βάρος $w_{ji}(n)$. Επομένως, η διόρθωση $\Delta w_{ji}(n)$ που εφαρμόζεται στο $w_{ji}(n)$ ορίζεται από τον κανόνα Δέλτα ως:

$$\Delta w_{ji}(n) = -\eta\frac{\partial\mathcal{E}(n)}{\partial w_{ji}(n)} \tag{21}$$

όπου η η παράμετρος ρυθμού μάθησης. Το αρνητικό πρόσημο της σχέσης (21) υποδηλώνει τη χρήση της μεθόδου βαθμωτής κατάβασης (gradient descent) στο χώρο βαρών, δηλαδή αναζητείται η κατεύθυνση για τη μεταβολή των βαρών η οποία μειώνει την τιμή του $\mathcal{E}(n)$. Με χρήση του κανόνα της αλυσίδας και των σχέσεων (19) και (20) και μετά από σειρά πράξεων προκύπτει τελικά

$$\Delta w_{ji}(n) = \eta\delta_j(n)y_i(n) \tag{22}$$

όπου η *τοπική κλίση* $\delta_j(n)$ ορίζεται ως

$$\delta_j(n) = \frac{\partial\mathcal{E}(n)}{\partial\upsilon_j(n)} = \frac{\partial\mathcal{E}(n)}{\partial e_j(n)}\frac{\partial e_j(n)}{\partial y_j(n)}\frac{\partial y_j(n)}{\partial\upsilon_j(n)} = e_j(n)\varphi_j{}'\left(\upsilon_j(n)\right) \tag{23}$$

Παρατηρώντας τις σχέσεις (22) και (23) είναι εμφανές ότι το σφάλμα $e_j(n)$ για το νευρώνα εξόδου j είναι είναι σημαντικός παράγοντας στον υπολογισμό της προσαρμογής $\Delta w_{ji}(n)$. Επομένως, πρέπει να γίνει μία διάκριση μεταξύ της περίπτωσης που ο νευρώνας j είναι ένας κόμβος εξόδου και της περίπτωσης που ο νευρώνας j είναι κρυφός κόμβος.

Στην περίπτωση που *ο νευρώνας j βρίσκεται στο επίπεδο εξόδου* του δικτύου τροφοδοτείται με τη δική του επιθυμητή απόκριση. Το σήμα σφάλματος $e_j(n)$ που σχετίζεται με αυτό το νευρώνα υπολογίζεται συγκεκριμένα από τη σχέση (17) (βλ. και σχήμα 7). Δεδομένου του σφάλματος $e_j(n)$ η τοπική κλίση $\delta_j(n)$ υπολογίζεται απλά από τη σχέση (23).

Στην περίπτωση που *ο νευρώνας j βρίσκεται σε ένα κρυφό επίπεδο* του δικτύου δεν υπάρχει καθορισμένη επιθυμητή απόκριση γι' αυτόν. Το σήμα σφάλματος που αντιστοιχεί σε ένα τέτοιο νευρώνα, επομένως, θα πρέπει να καθοριστεί αναδρομικά, χρησιμοποιώντας τα σήματα σφάλματος όλων των νευρώνων με τους οποίους συνδέεται άμεσα ο εν λόγω νευρώνας. Η περίπτωση αυτή απαιτεί ειδική μελέτη, για την οποία χρήσιμη είναι η εικόνα 8, όπου βλέπουμε τον κρυφό νευρώνα j να συνδέεται με ένα νευρώνα εξόδου k.

Αρχικά, τροποποιούμε τη σχέση (23) όπου υπολογίζεται η τοπική κλίση $\delta_j(n)$ για τον κρυφό νευρώνα j ως εξής

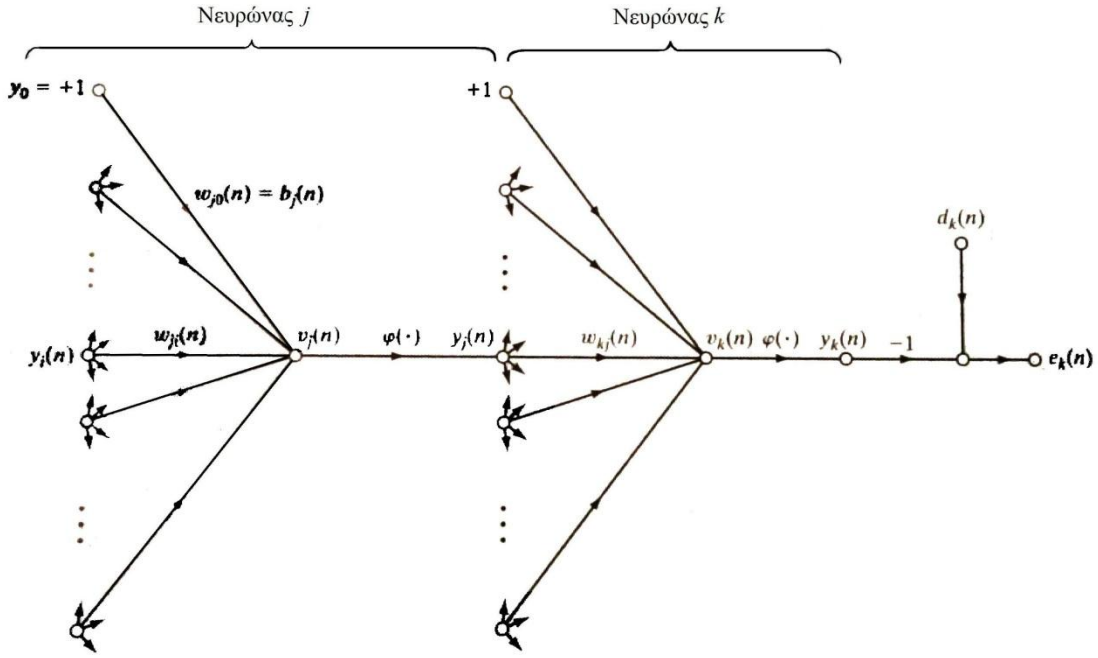
$$\delta_j(n) = \frac{\partial \mathcal{E}(n)}{\partial y_j(n)} \frac{\partial y_j(n)}{\partial \upsilon_j(n)} = -\frac{\partial \mathcal{E}(n)}{\partial y_j(n)} \varphi_j{}'\left(\upsilon_j(n)\right) \tag{24}$$

ενώ εφαρμόζοντας τη σχέση (18) για το νευρώνα εξόδου k ισχύει

$$\mathcal{E}(n) = \frac{1}{2} \sum_{k \in C} e_k^2(n) \tag{25}$$

και τελικά διαφορίζοντας τη σχέση (25) ως προς το λειτουργικό σήμα $y_j(n)$ προκύπτει

$$\frac{\partial \mathcal{E}(n)}{\partial y_j(n)} = \sum_k e_k(n) \frac{\partial e_k(n)}{\partial y_j(n)} = \sum_k e_k(n) \frac{\partial e_k(n)}{\partial \upsilon_k(n)} \frac{\partial \upsilon_k(n)}{\partial y_j(n)} \tag{26}$$



Σχήμα 8: Γράφημα ροής σήματος που περιγράφει τις λεπτομέρειες του νευρώνα εξόδου ***k*** ο οποίος συνδέεται στον κρυφό νευρώνα ***j***

Επίσης, σύμφωνα με το σχήμα 8 για το σφάλμα εξόδου και το τοπικό πεδίο του νευρώνα k ισχύει

$$e_k(n) = d_k(n) - y_k(n) \tag{27}$$

$$\upsilon_k(n) = \sum_{i=0}^{m} w_{kj}(n) y_j(n) \tag{28}$$

όμοια με τις σχέσεις (17) και (19). Τελικά, η ζητούμενη μερική παράγωγος προκύπτει

$$\frac{\partial \mathcal{E}(n)}{\partial y_j(n)} = -\sum_k e_k(n) \varphi_k{}'\big(\upsilon_k(n)\big)\, w_{kj}(n) = -\sum_k \delta_k(n)\, w_{kj}(n) \tag{29}$$

όπου η τοπική κλίση $\delta_k(n)$ ορίζεται όπως στη σχέση (23) αλλά αυτή τη φορά για το νευρώνα k. Έτσι, από τις σχέσεις (24) και (29) λαμβάνεται ο *τύπος οπισθοδιάδοσης* για την τοπική κλίση $\delta_j(n)$:

$$\delta_j(n) = \varphi_j{}'\left(v_j(n)\right)\sum_k \delta_k(n)\, w_{kj}(n) \tag{30}$$

Στη σχέση (30) εξωτερικός παράγοντας $\varphi_j{}'(v_j(n))$ εξαρτάται αποκλειστικά από τη συνάρτηση ενεργοποίησης που σχετίζεται με το νευρώνα j. Ο παράγοντας του αθροίσματος για όλα τα k που περιλαμβάνεται επίσης σ' αυτόν τον υπολογισμό εξαρτάται από δύο σύνολα όρων. Το πρώτο σύνολο όρων, $\delta_k(n)$, απαιτεί γνώση των σημάτων σφάλματος $e_k(n)$ για όλους τους νευρώνες που βρίσκονται στο αμέσως δεξιότερο επίπεδο του νευρώνα j και συνδέονται άμεσα με το νευρώνα j. Το δεύτερο σύνολο όρων $w_{kj}(n)$ αποτελείται από τα συναπτικά βάρη που σχετίζονται με αυτές τις συνδέσεις.

Συμπερασματικά, η διόρθωση $\Delta w_{ji}(n)$ που εφαρμόζεται από τον αλγόριθμο Οπισθοδιάδοσης στο συναπτικό βάρος $w_{ji}(n)$ που συνδέει ένα νευρώνα i με ένα νευρώνα j δίνεται από τη σχέση (22). Στη σχέση (22), η τοπική κλίση $\delta_j(n)$ διαφέρει ανάλογα με το αν ο νευρώνας j είναι νευρώνας του επιπέδου εξόδου ή όχι. Η τοπική κλίση υπολογίζεται από τη σχέση (23) εάν ο νευρώνας j είναι νευρώνας εξόδου, ειδάλλως χρησιμοποιείται η σχέση (30).

Στο σημείο αυτό πρέπει να γίνει μία σειρά από επιλογές για τον καθορισμό των παραμέτρων που υπεισέρχονται στου μαθηματικό μοντέλο του αλγορίθμου ΒΚ, συγκεκριμένα για τη συνάρτηση ενεργοποίησης, το ρυθμό μάθησης και τα κριτήρια τερματισμού του αλγορίθμου. Κατ' αρχάς, όσον αφορά στη συνάρτηση ενεργοποίησης $\varphi(\cdot)$, όπως υποδεικνύεται από τις μαθηματικές σχέσεις που προηγήθηκαν, είναι απαραίτητο η συνάρτηση αυτή να είναι διαφορίσιμη. Η συνηθέστερη περίπτωση είναι η χρήση σιγμοειδούς μη γραμμικότητας, όπως παρουσιάστηκε στο σχήμα 3β, με τις δύο μορφές που χρησιμοποιούνται ευρύτερα, τη λογιστική συνάρτηση και τη συνάρτηση υπερβολικής εφαπτομένης.

Αν επιλεγεί η λογιστική συνάρτηση η οποία ορίζεται ως

$$\varphi_j\left(v_j(n)\right) = \frac{1}{1+\exp\left(-av_j(n)\right)}, \qquad a > 0$$

η σχέση (23) για νευρώνα στο επίπεδο εξόδου και η σχέση (28) για κρυφό νευρώνα γίνονται αντίστοιχα

$$\delta_j(n) = \alpha\left[d_j(n) - o_j(n)\right]o_j(n)\left[1 - o_j(n)\right] \tag{31}$$

$$\delta_j(n) = ay_j(n)\left[1 - y_j(n)\right]\sum_k \delta_k(n)\, w_{kj}(n) \tag{32}$$

όπου προκειμένου για νευρώνα εξόδου, $o_j(n) = y_j(n)$.

Ομοίως, αν επιλεγεί η συνάρτηση υπερβολικής εφαπτομένης με τύπο

$$\varphi_j\left(v_j(n)\right) = \alpha\tanh\left(bv_j(n)\right), \qquad a, b > 0$$

η σχέση (23) για νευρώνα στο επίπεδο εξόδου και η σχέση (30) για κρυφό νευρώνα γίνονται αντίστοιχα

$$\delta_j(n) = \frac{b}{a}\left[d_j(n) - o_j(n)\right]\left[a - o_j(n)\right]\left[a + o_j(n)\right] \tag{33}$$

$$\delta_j(n) = \frac{b}{a}\left[a - y_j(n)\right]\left[a + y_j(n)\right]\sum_k \delta_k(n)\, w_{kj}(n) \tag{34}$$

όπου προκειμένου για νευρώνα εξόδου, $o_j(n) = y_j(n)$.

Αναφορικά με την παράμετρο του ρυθμού μάθησης η που έχει εισαχθεί στη διόρθωση $\Delta w_{ji}(n)$ των συναπτικών βαρών, όπως αναφέρθηκε και προηγουμένως, αυτή καθορίζει την ισχύ του διανύσματος κλίσης $\partial\mathcal{E}(n)/\partial w_{ji}(n)$ επί της διόρθωσης $\Delta w_{ji}(n)$. Με άλλα λόγια, ορίζεται από τον αλγόριθμο ΒΚ μια τροχιά που διαγράφεται στο χώρο των βαρών σύμφωνα με τη μέθοδο της πλέον απότομης κατάβασης. Όσο μικρότερη είναι η παράμετρος ρυθμού μάθησης τόσο πιο αποδυναμωμένες είναι οι αλλαγές που θα επιβληθούν στα συναπτικά βάρη του δικτύου από τη μία επανάληψη στην επόμενη και επίσης, τόσο πιο ομαλή θα είναι η τροχιά στο χώρο των βαρών. Όπως είναι αναμενόμενο, σ' αυτή την περίπτωση η διαδικασία μάθησης είναι πιο αργή, απαιτούνται δηλαδή περισσότερες επαναλήψεις. Στην περίπτωση που επιλεγεί πολύ μεγάλη τιμή της παραμέτρου του ρυθμού μάθησης, η διαδικασία μάθησης επιταχύνεται μεν, υπάρχει όμως ο κίνδυνος οι μεγάλες αλλαγές στα συναπτικά βάρη να πάρουν τη μορφή ταλάντωσης οπότε το σύστημα είναι ασταθές.

Προκειμένου να βρεθεί μία ισορροπία μεταξύ μεγάλου ρυθμού μάθησης και ευστάθειας, εισάγεται στον κανόνα Δέλτα για τη διόρθωση των συναπτικών βαρών $\Delta w_{ji}(n)$ ένας νέος όρος ορμής, δηλαδή:

$$\Delta w_{ji}(n) = \alpha\Delta w_{ji}(n-1) + \eta\delta_j(n)y_i(n) \tag{35}$$

όπου ο (συνήθως) θετικός αριθμός α ονομάζεται *σταθερά ορμής*. Η σταθερά αυτή έχει το ρόλο ελέγχου της ανάδρασης που δέχεται η διόρθωση $\Delta w_{ji}(n)$. Λύνοντας την εξίσωση (35) σαν εξίσωση διαφορών πρώτης τάξης ως προς $\Delta w_{ji}(n)$ προκύπτει:

$$\Delta w_{ji}(n) = \eta\sum_{t=0}^{n} a^{n-t}\delta_j(t)y_i(t) = -\eta\sum_{t=0}^{n} a^{n-t}\frac{\partial\mathcal{E}(t)}{\partial w_{ji}(t)} \tag{36}$$

Η σχέση (36) αποτελεί μία εκθετικά σταθμισμένη χρονοσειρά μήκους $n+1$. Για να συγκλίνει η χρονοσειρά, η σταθερά ορμής πρέπει να κινείται στο διάστημα $0 \le |\alpha| < 1$. Για $\alpha = 0$ μεταπίπτουμε στην περίπτωση όπου δεν χρησιμοποιείται ορμή. Γενικότερα, όταν υπάρχει ταλάντωση του προσήμου της προσαρμογής η ύπαρξη της ορμής δρα σταθεροποιητικά ενώ όταν το πρόσημο είναι σταθερό τότε η χρήση της ορμής επιταχύνει την κατάβαση.

Στο σημείο αυτό πρέπει να σημειωθεί ότι μέχρι στιγμής γινόταν αναφορά σε παράμετρο ρυθμού μάθησης η, σταθερής τιμής. Στην πραγματικότητα όμως η παράμετρος ρυθμού μάθησης μπορεί να είναι διαφορετική για κάθε νευρώνα, έστω η_{ij}. Τέλος, υπάρχει δυνατότητα να μην είναι όλα τα συναπτικά βάρη μεταβλητά, αλλά να υπάρχουν μερικοί νευρώνες με σταθερό βάρος, οπότε για τους νευρώνες αυτούς θα ισχύει $\eta_{ij} = 0$.

Η εφαρμογή του αλγορίθμου ΒΚ γίνεται σε δύο φάσεις υπολογισμών. Το πρώτο πέρασμα είναι το λεγόμενο πέρασμα με κατεύθυνση προς τα εμπρός ενώ το δεύτερο πέρασμα είναι το λεγόμενο πέρασμα με κατεύθυνση προς τα πίσω. Κατά το *πέρασμα με κατεύθυνση προς τα εμπρός* δεν γίνονται αλλαγές στα συναπτικά βάρη του δικτύου, παρά μόνο υπολογίζονται τα λειτουργικά σήματα στην έξοδο των νευρώνων των διαδοχικών κρυφών επιπέδων μέχρι το επίπεδο εξόδου, οπότε και υπολογίζεται το σήμα σφάλματος. Το

πέρασμα με κατεύθυνση προς τα πίσω ξεκινά από το επίπεδο εξόδου διαδίδοντας το σφάλμα διαδοχικά στα προηγούμενα κρυφά επίπεδα όπου υπολογίζεται αναδρομικά η τοπική κλίση κάθε νευρώνα. Με αυτόν τον τρόπο μεταβάλλονται τα συναπτικά βάρη σύμφωνα με τη διόρθωση του κανόνα Δέλτα.

Ο αλγόριθμος ΒΚ στη γενική περίπτωση δεν μπορεί να αποδειχτεί ότι συγκλίνει σε κάποια λύση, άρα παρουσιάζεται επίσης δυσκολία και στην υιοθέτηση κριτηρίων τερματισμού που να είναι καλά ορισμένα. Υπάρχουν όμως πρακτικοί κανόνες οι οποίοι μπορούν να χρησιμοποιηθούν για τον τερματισμό της διαδικασίας προσαρμογής των βαρών. Αρχικά, αναφέρεται το εξής κριτήριο [Kramer et al., 1989]

> *Ο αλγόριθμος ΒΚ θεωρείται ότι έχει συγκλίνει όταν η Ευκλείδεια νόρμα του διανύσματος κλίσης φτάσει σε ένα επαρκώς μικρό κατώφλι κλίσης.*

Το κριτήριο αυτό έχει το μειονέκτημα ότι προκύπτουν μεγάλοι χρόνοι μάθησης, όπως επίσης απαιτεί τον υπολογισμό του διανύσματος κλίσης. Άλλη μία πρόταση που μπορεί να χρησιμοποιηθεί ως κριτήριο σύγκλισης είναι:

> *Ο αλγόριθμος ΒΚ θεωρείται ότι έχει συγκλίνει όταν ο απόλυτος ρυθμός μεταβολής του μέσου τετραγωνικού σφάλματος ανά εποχή*[1] *είναι επαρκώς μικρός.*

Μια τυπική τιμή τερματισμού για το μέσο τετραγωνικό σφάλμα αν ακολουθηθεί αυτό το κριτήριο είναι στο διάστημα 0.1% ως 1% ανά εποχή, αλλά μπορεί να επιλεγεί ακόμα μικρότερη τιμή κατωφλίου στο 0.01% ανά εποχή. Η επιλογή της τιμής τερματισμού είναι πολύ κρίσιμη καθώς το κριτήριο αυτό πολλές φορές επιφέρει πρόωρο τερματισμό της διαδικασίας μάθησης.

Τέλος, για τον τερματισμό του αλγορίθμου χρησιμοποιείται ακόμη ένα κριτήριο, με χρήση της *διασταυρωμένης επικύρωσης (cross validation)*, οπότε μετά από κάθε επανάληψη της διαδικασίας μάθησης το δίκτυο ελέγχεται ως προς την ικανότητα γενίκευσής του. Η αξιολόγηση αυτή επιτυγχάνεται χωρίζοντας το σύνολο εκπαίδευσης σε δύο ξένα μεταξύ τους υποσύνολα, εκ των οποίων το πρώτο χρησιμοποιείται για τη γνωστή διαδικασία προσαρμογής των βαρών ενώ το δεύτερο χρησιμοποιείται στο να διαπιστωθεί κατά πόσο το δίκτυο είναι σε θέση να ταξινομήσει με ικανοποιητικό σφάλμα τα μελλοντικά πρότυπα.
Έτσι, μπορεί να παρουσιαστεί πλέον η πλήρης εικόνα του αλγορίθμου ΒΚ όπως παρουσιάζεται στον πίνακα 2.

Πίνακας 2: Σύνοψη του Αλγορίθμου Οπισθοδιάδοσης (Backpropagation)

1. ***Αρχικοποίηση***
   Αρχικοποίησε τα συναπτικά βάρη χρησιμοποιώντας μια ομοιόμορφη κατανομή.

2. ***Παρουσιάσεις των Παραδειγμάτων Εκπαίδευσης***
   Παρουσίασε στο δίκτυο μια εποχή παραδειγμάτων εκπαίδευσης.

[1] Ο όρος εποχή αναφέρεται στην *εποχή εκπαίδευσης*, δηλαδή το διάστημα παρουσίασης στον αλγόριθμο όλων των παραδειγμάτων ενός δείγματος εκπαίδευσης $\mathcal{T}$. Το ίδιο σύνολο εκπαίδευσης μπορεί να χρησιμοποιηθεί περισσότερες από μία φορές ώστε να επιτευχθεί καλύτερη σύγκλιση.

3. ***Υπολογισμός προς τα Εμπρός.***
   Για κάθε διάνυσμα εκπαίδευσης υπολόγισε το τοπικό πεδίο για το νευρώνα j στο επίπεδο l $(l = 1,2, \dots, L)$

$$v_j{}^{(l)}(n) = \sum_i w_{ji}{}^{(l)}(n) y_i{}^{(l-1)}(n)$$

   Υπολόγισε το σήμα σφάλματος

$$e_j{}^{(L)}(n) = d_j(n) - y_j{}^{(L)}(n) = d_j(n) - o_j(n)$$

4. ***Υπολογισμός προς τα Πίσω.***
   Υπολόγισε τις τοπικές κλίσεις του δικτύου

$$\delta_j^{(l)} = \begin{cases} e_j{}^{(L)}(n)\varphi_j{}'\left(v_j{}^{(l)}(n)\right), & \text{νευρώνας } j \text{ στο επίπεδο εξόδου L} \\ \varphi_j{}'\left(v_j{}^{(l)}(n)\right)\sum_k \delta_k{}^{(l+1)}(n)\, w_{kj}{}^{(l+1)}(n), & \text{νευρώνας } j \text{ στο κρυφό επίπεδο } l \end{cases}$$

   Προσάρμοσε τα συναπτικά βάρη του δικτύου στο επίπεδο l

$$w_{ji}{}^{(l)}(n+1) = w_{ji}{}^{(l)}(n) + a w_{ji}{}^{(l)}(n-1) + \eta \delta_j^{(l)}(n)\, y_i{}^{(l-1)}(n)$$

5. ***Συνέχιση.*** Αύξησε το χρονικό βήμα n κατά 1 και επίστρεψε στο βήμα 3 μέχρι να ικανοποιηθεί το επιλεγμένο κριτήριο τερματισμού.

Στο σημείο αυτό να σημειωθεί ότι με κατάλληλες τροποποιήσεις ο αλγόριθμος BK μπορεί να εφαρμοστεί και στην περίπτωση όπου ζητείται να ακολουθηθεί η στρατηγική της μαζικής μάθησης. Η εφαρμογή της μαζικής μάθησης προσφέρει καλύτερη εκτίμηση των τοπικών κλίσεων και επομένως καλύτερη σύγκλιση. Η online μάθηση, από την άλλη πλευρά, προσδίδει μεγαλύτερη τυχαιότητα κατά την εκπαίδευση, γεγονός που μικραίνει την πιθανότητα να παγιδευτεί ο αλγόριθμος σε κάποιο τοπικό ελάχιστο της συνάρτησης κόστους. Η επιλογή μεταξύ των δύο προσεγγίσεων γίνεται ανάλογα με τη φύση του προβλήματος κατηγοριοποίησης.

Όταν η διαδικασία της εκπαίδευσης έχει ολοκληρωθεί οι τιμές στις οποίες τα συναπτικά βάρη έχουν συγκλίνει σταθεροποιούνται και το δίκτυο πλέον μπορεί να μπει στη διαδικασία της κατηγοριοποίησης. Αυτή είναι μια διαδικασία πολύ απλούστερη από την εκπαίδευση. Ένα άγνωστο διάνυσμα παρουσιάζεται στην είσοδο του δικτύου και ταξινομείται στην κλάση που υποδεικνύει η έξοδος του δικτύου. Οι υπολογισμοί που εκτελούνται από τους νευρώνες είναι προσθέσεις και πολλαπλασιασμοί που ακολουθούνται από μη γραμμικότητες. Για το λόγο αυτό έχουν προταθεί διάφορες υλοποιήσεις σε επίπεδο υλικού, όπως οπτικά κυκλώματα και κυκλώματα VLSI. Επίσης λόγω του εγγενούς παραλληλισμού που παρουσιάζουν τα νευρωνικά δίκτυα, για τους υπολογισμούς μπορεί να χρησιμοποιηθούν συστήματα παράλληλης επεξεργασίας. Με αυτή τη λογική έχουν αναπτυχθεί και ειδικά συστήματα νευροϋπολογιστών (neurocomputers).

2.1.2. Μηχανή Διανυσμάτων Υποστήριξης (Support Vector Machine, SVM)

Στην ενότητα αυτή θα μελετηθεί σε βάθος η κατηγοριοποίηση με χρήση των μηχανών διανυσμάτων υποστήριξης. Η μέθοδος αυτή αναφέρεται σε προβλήματα κατηγοριοποίησης

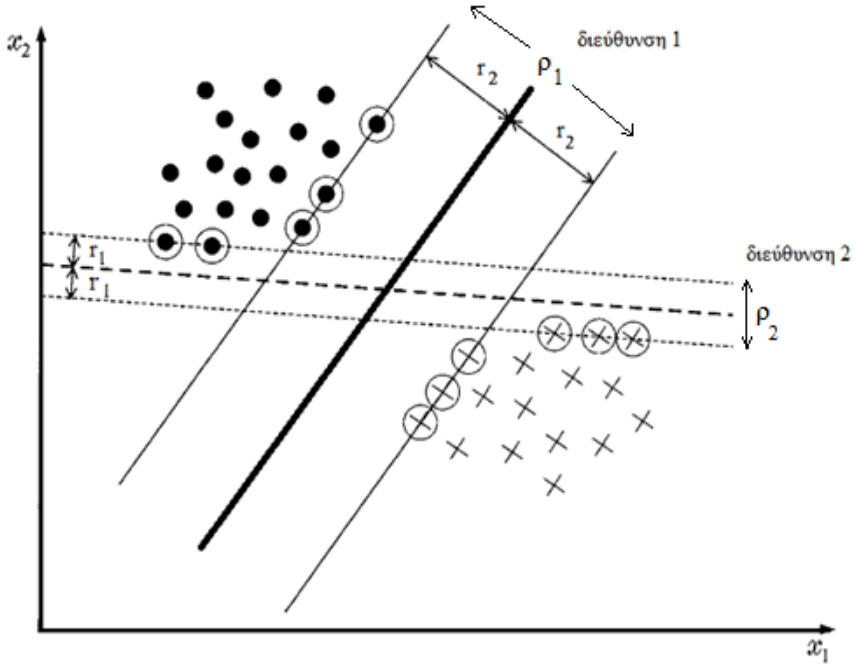
σε δύο ομάδες και η κεντρική ιδέα της είναι η αντιστοίχιση των διανυσμάτων με κάποιο κριτήριο που επιλέγεται εξ αρχής σε έναν πολυδιάστατο χώρο χαρακτηριστικών. Στον χώρο αυτό κατασκευάζεται μία γραμμική επιφάνεια απόφασης με ειδικές ιδιότητες ώστε ο κατηγοριοποιητής να έχει την ικανότητα να λειτουργεί στη γενική περίπτωση διανυσμάτων εισόδου. Το πρόβλημα που τίθεται είναι πώς θα καταστεί δυνατό να προσδιοριστεί ένα *υπερεπίπεδο (hyperplane)* στην επιφάνεια αυτή ώστε η γενίκευση να είναι ικανοποιητική. Η λύση του προβλήματος αυτού διαφοροποιείται ανάλογα με το αν τα πρότυπα είναι γραμμικώς διαχωρίσιμα ή όχι.

Γραμμικά Διαχωρίσιμα Πρότυπα

Ξεκινώντας από την περίπτωση των γραμμικά διαχωρίσιμων προτύπων, θεωρούμε ένα δείγμα εκπαίδευσης $\{(\mathbf{x_i}, d_i)\}_{i=1}^{N}$ όπου $\mathbf{x}_i$ είναι το i-οστό διάνυσμα εισόδου και d_i το αντίστοιχο επιθυμητό αποτέλεσμα της κατηγοριοποίησης και έστω ότι για τη μία κλάση ισχύει $d_i = +1$ και για την άλλη $d_i = -1$. Η εξίσωση της επιφάνειας απόφασης με τη μορφή ενός υπερεπιπέδου που πραγματοποιεί τον διαχωρισμό είναι

$$\mathbf{w}^T\mathbf{x} + b = 0 \tag{37}$$

όπου $\mathbf{x}$ είναι ένα διάνυσμα εισόδου, $\mathbf{w}$ ένα προσαρμόσιμο διάνυσμα βαρών και b είναι μία πόλωση. Το διάνυσμα $\mathbf{w}$ αντιπροσωπεύει τη διεύθυνση του υπερεπιπέδου ενώ το b αντιπροσωπεύει την ακριβή θέση του υπερεπιπέδου στο χώρο.

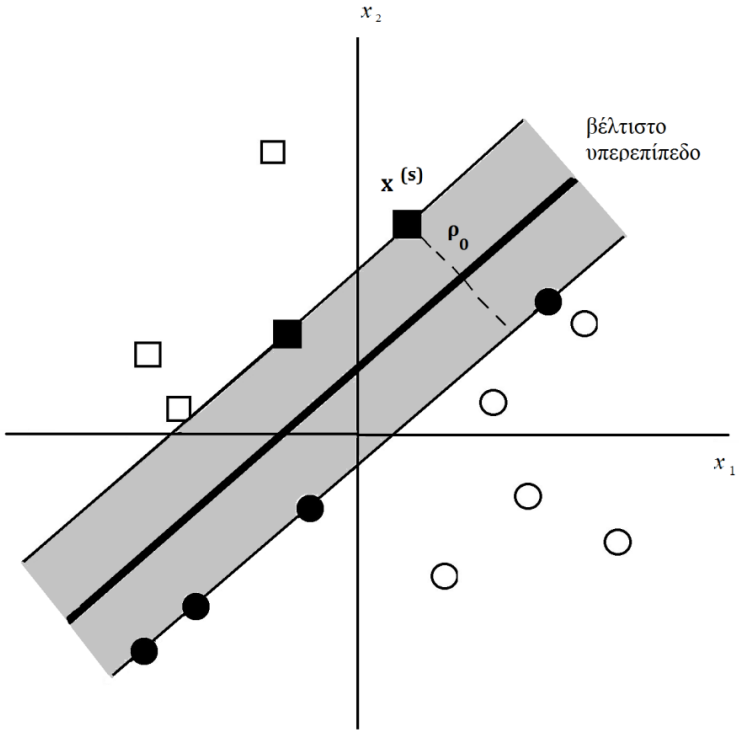


Σχήμα 9: Δύο πιθανά υπερεπίπεδα για ένα σύνολο εκπαίδευσης σε διδιάστατο χώρο χαρακτηριστικών. Το υπερεπίπεδο διεύθυνσης 1 είναι προτιμότερο του άλλου καθώς το περιθώριο διαχωρισμού είναι μεγαλύτερο από αυτό της διεύθυνσης 2. Τα κυκλωμένα σημεία αποτελούν τα διανύσματα υποστήριξης που αντιστοιχούν στο κάθε υπερεπίπεδο**.**

Για ένα δεδομένο σύνολο εκπαίδευσης μπορούν να προκύψουν περισσότερα από ένα υπερεπίπεδα που να διαχωρίζουν τις δύο κλάσεις, όπως φαίνεται στο σχήμα 9. Επομένως, για τα πρότυπα που είναι κατανεμημένα εκατέρωθεν του υπερεπιπέδου θα ισχύει

$$\mathbf{w}^T\mathbf{x} + b \geq 0, \text{για } d_i = +1$$
$$\mathbf{w}^T\mathbf{x} + b < 0, \text{για } d_i = -1$$

Για ένα δεδομένο διάνυσμα βαρών $\mathbf{w}$ και πόλωση b ο διαχωρισμός μεταξύ του υπερεπιπέδου που ορίζει η εξίσωση (37) και του πλησιέστερου σημείου δεδομένων ονομάζεται *περιθώριο διαχωρισμού (margin of separation)* και συμβολίζεται με το γράμμα ρ. Ο στόχος της μηχανής διανυσμάτων υποστήριξης είναι να προσδιορίσει το υπερεπίπεδο εκείνο για το οποίο το περιθώριο διαχωρισμού μεγιστοποιείται. Η επιφάνεια απόφασης που πληροί αυτή τη συνθήκη ονομάζεται *βέλτιστο υπερεπίπεδο (optimal hyperplane)*, ενώ το αντίστοιχο περιθώριο διαχωρισμού ονομάζεται *βέλτιστο περιθώριο (optimal margin)*. Παρατηρώντας το σχήμα 9, μεταξύ των υπερεπιπέδων διεύθυνσης 1 και 2, το υπερεπίπεδο διεύθυνσης 1 είναι το βέλτιστο υπερεπίπεδο καθ' ότι το περιθώριο διαχωρισμού είναι μεγαλύτερο. Ο λόγος για τον οποίο είναι επιθυμητή μία τέτοια συνθήκη είναι ότι η κατηγοριοποίηση των μελλοντικών προτύπων εισόδου είναι πολύ πιο ασφαλής, δεδομένου ότι παρέχεται καλύτερη δυνατότητα να ταξινομηθούν πρότυπα που διαφέρουν από αυτά του συνόλου εκπαίδευσης. Αυτή είναι και η έννοια της γενικευμένης απόδοσης του κατηγοριοποιητή στην όποια έγινε νύξη προηγουμένως. Προκειμένου για ένα διδιάστατο χώρο χαρακτηριστικών η απεικόνιση του βέλτιστου υπερεπιπέδου φαίνεται στο σχήμα 10.



Σχήμα 10: Γραφική απεικόνιση της έννοιας του βέλτιστου υπερεπιπέδου για γραμμικά διαχωρίσιμα πρότυπα σε χώρο 2 διαστάσεων.
Τα σημεία δεδομένων με μαύρο χρώμα είναι τα διανύσματα υποστήριξης,

Η εξίσωση του βέλτιστου υπερεπιπέδου προκύπτει από την εξίσωση (37) αν διατίθενται οι βέλτιστες τιμές του διανύσματος βαρών $\mathbf{w}_0$ και της πόλωσης b ως:

$$\mathbf{w_0}^T\mathbf{x} + b_0 = 0 \qquad (38)$$

απ' όπου προκύπτει και η συνάρτηση διάκρισης

$$g(\mathbf{x}) = \mathbf{w_0}^T\mathbf{x} + b_0 \tag{39}$$

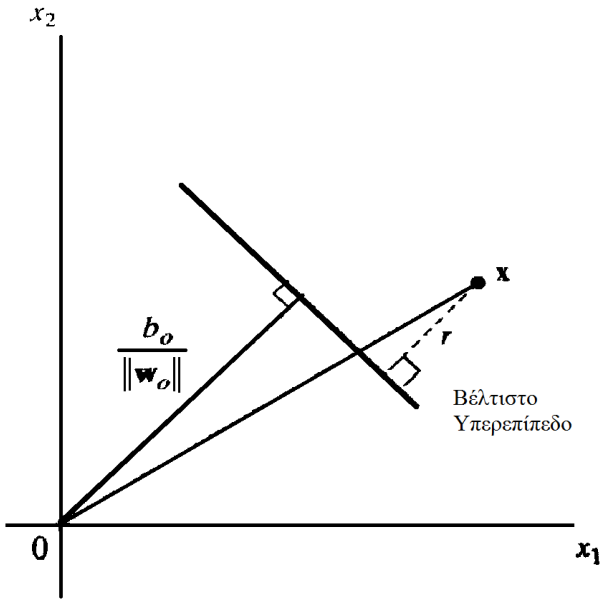
η οποία δίνει ένα αλγεβρικό μέτρο της απόστασης του $\mathbf{x}$ από το βέλτιστο υπερεπίπεδο. Αναλυτικότερα, το $\mathbf{x}$ μπορεί να εκφραστεί ως

$$\mathbf{x} = \mathbf{x}_p + r\frac{\mathbf{w}_0}{\|\mathbf{w}_0\|} \tag{40}$$

όπου $\mathbf{x}_p$ είναι η κανονική προβολή του $\mathbf{x}$ στο βέλτιστο επίπεδο και r είναι η επιθυμητή αλγεβρική απόσταση, η οποία είναι θετική αν το $\mathbf{x}$ βρίσκεται στη θετική πλευρά του βέλτιστου υπερεπιπέδου και αρνητική αν το $\mathbf{x}$ βρίσκεται στην αρνητική πλευρά. Όμως εξ' ορισμού ισχύει $g(\mathbf{x}_p) = 0$, άρα

$$g(\mathbf{x}) = \mathbf{w_0}^T\mathbf{x} + b_0 = r\|\mathbf{w}_0\| \Leftrightarrow$$
$$r = \frac{g(\mathbf{x})}{\|\mathbf{w}_0\|} \tag{41}$$

Η απόσταση του βέλτιστου υπερεπιπέδου από το σημείο αρχής ($\mathbf{x} = \mathbf{0}$) είναι $b_0/\|\mathbf{w}_0\|$. Αν $b_0 > 0$ το σημείο αρχής βρίσκεται στη θετική πλευρά του υπερεπιπέδου, αν $b_0 < 0$ το σημείο αρχής βρίσκεται στην αρνητική πλευρά του υπερεπιπέδου, ενώ αν $b_0 = 0$ το υπερεπίπεδο διέρχεται από το σημείο αρχής. Οι παρατηρήσεις αυτές απεικονίζονται στο σχήμα 3.



Σχήμα 11: Γεωμετρική ερμηνεία των αλγεβρικών αποστάσεων ενός σημείου από το βέλτιστο υπερεπίπεδο στην περίπτωση χώρου 2 διαστάσεων

Δοθέντος λοιπόν του συνόλου προτύπων εκπαίδευσης $\mathcal{T} = \{(\mathbf{x_i}, d_i)\}$ οι παράμετροι $\mathbf{w}_0$και b_0 για το βέλτιστο υπερεπίπεδο πρέπει να ικανοποιούν τη σχέση

$$d_i(\mathbf{w}_0{}^T\mathbf{x}_i + b_0) \geq 1, \text{για } i = 1,2,\dots N \tag{42}$$

Τα σημεία δεδομένων $(\mathbf{x_i}, d_i)$ για τα οποία στην παραπάνω σχέση ο ισχύει με ισότητα ονομάζονται *διανύσματα υποστήριξης (support vectors)* και σε αυτά οφείλει την ονομασία

της η μέθοδος των Μηχανών Διανυσμάτων Υποστήριξης. Τα διανύσματα αυτά έχουν ιδιαίτερη θέση στη μέθοδο αυτή καθώς είναι τα πλησιέστερα στο βέλτιστο υπερεπίπεδο και επομένως είναι και αυτά που παρουσιάζουν τη μεγαλύτερη δυσκολία στην κατηγοριοποίηση και επομένως έχουν άμεση επίδραση στη βέλτιστη θέση της επιφάνειας απόφασης.

Έστω ένα διάνυσμα υποστήριξης $\mathbf{x}^{(s)}$. Κατά συνέπεια του ορισμού των διανυσμάτων υποστήριξης ισχύει

$$g\left(\mathbf{x}^{(s)}\right) = \mathbf{w_0}^T\mathbf{x}^{(s)} + b_0 = \pm 1 \text{ για } \boldsymbol{d}^{(s)} = \pm 1 \tag{43}$$

Επομένως το βέλτιστο περιθώριο κέρδους προσδιορίζεται ως ο χώρος μεταξύ των παράλληλων επιφανειών $\mathbf{w_0}^T\mathbf{x}^{(s)} + b_0 = \pm 1$. Από την εξίσωση (41) προκύπτει η αλγεβρική απόσταση του διανύσματος υποστήριξης $\mathbf{x}^{(s)}$ από το βέλτιστο υπερεπίπεδο

$$r = \frac{g\left(\mathbf{x}^{(s)}\right)}{\|\mathbf{w}_0\|} = \begin{cases} \dfrac{1}{\|\mathbf{w}_0\|}, & \text{για } \boldsymbol{d}^{(s)} = +1 \\ -\dfrac{1}{\|\mathbf{w}_0\|}, & \text{για } \boldsymbol{d}^{(s)} = -1 \end{cases} \tag{44}$$

Το θετικό πρόσημο υποδεικνύει ότι το $\mathbf{x}^{(s)}$ βρίσκεται στη θετική πλευρά του βέλτιστου υπερεπιπέδου ενώ το αρνητικό πρόσημο ότι το $\mathbf{x}^{(s)}$ βρίσκεται στη αρνητική πλευρά του βέλτιστου υπερεπιπέδου. Έτσι, από την εξίσωση (44) προκύπτει η *βέλτιστη τιμή του περιθωρίου διαχωρισμού* ρ μεταξύ των δύο κλάσεων που στις οποίες ανήκουν τα πρότυπα εκπαίδευσης του συνόλου $\mathcal{T}$:

$$\rho = 2r = \frac{2}{\|\mathbf{w}_0\|} \tag{45}$$

Από την εξίσωση (45) προκύπτει ότι προκειμένου να μεγιστοποιηθεί το περιθώριο διαχωρισμού μεταξύ των δύο κλάσεων η Ευκλείδεια νόρμα του διανύσματος βαρών πρέπει να ελαχιστοποιείται. Η συνθήκη αυτή καθιστά το βέλτιστο υπερεπίπεδο που ορίζει η εξίσωση (39) *μοναδικό*. Η ελαχιστοποίηση της νόρμας αυτής είναι μία διαδικασία μη γραμμικής δευτεροβάθμιας βελτιστοποίησης που υπόκειται σε ένα σύνολο περιορισμών οι οποίοι είναι γραμμικές ανισότητες.

Η μέθοδος της μηχανής διανυσμάτων υποστήριξης ικανοποιεί μια καλώς ορισμένη συνθήκη βελτιστότητας αφού η διαδικασία που ακολουθείται είναι μία ειδική περίπτωση της *βελτιστοποίησης κυρτών συναρτήσεων*. Η διαδικασία της βελτιστοποίησης αυτής αποτελείται από τέσσερα κύρια βήματα. Αρχικά, γίνεται η διατύπωση του προβλήματος στον πρωτεύοντα χώρο βαρών. Στη συνέχεια, κατασκευάζεται η συνάρτηση Lagrange του προβλήματος και έπειτα διατυπώνονται οι συνθήκες για την βελτιστότητα της μηχανής. Τέλος, γίνεται η επίλυση του προβλήματος στον δυϊκό χώρο των πολλαπλασιαστών Lagrange.

Η φορμαλιστική διατύπωση του προβλήματος βελτιστοποίησης με περιορισμούς η οποία περιγράφει το πρωτεύον πρόβλημα είναι η εξής:

Δοθέντος ενός δείγματος εκπαίδευσης $\{(\mathbf{x_i}, d_i)\}_{i=1}^N$*, να βρεθούν οι βέλτιστες τιμές του διανύσματος βαρών* $\mathbf{w}$ και της πόλωσης b *ώστε να ικανοποιούν τους περιορισμούς*

$$d_i(\mathbf{w}_0{}^T\mathbf{x}_i + b_0) \geq 1, \text{ για } i = 1,2,\dots N$$

και το διάνυσμα βαρών $\mathbf{w}$ *να ελαχιστοποιεί τη συνάρτηση κόστους*

$$\Phi(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

Δεδομένου ότι η $\Phi(\mathbf{w})$ είναι μία κυρτή συνάρτηση του $\mathbf{w}$ και οι περιορισμοί είναι γραμμικοί ως προς το $\mathbf{w}$ είναι δυνατή η λύση του προβλήματος με τη χρήση της *μεθόδου πολλαπλασιαστών Lagrange*. Επομένως, η συνάρτηση Lagrange που κατασκευάζεται είναι η εξής

$$\mathcal{L}(\mathbf{w}, b, \boldsymbol{\lambda}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{N} \lambda_i[d_i(\mathbf{w_0}^T\mathbf{x}_i + b) - 1] \tag{46}$$

όπου οι μεταβλητές λ_i ονομάζονται πολλαπλασιαστές Lagrange και $\boldsymbol{\lambda}$ είναι το αντίστοιχο διάνυσμα των πολλαπλασιαστών Lagrange. Οι συνθήκες βελτιστότητας Karush-Kuhn-Tucker (KKT) που πρέπει να ικανοποιεί ο ελαχιστοποιητής είναι οι παρακάτω:

$$\frac{\partial \mathcal{L}(\mathbf{w}, b, \boldsymbol{\lambda})}{\partial \mathbf{w}} = \mathbf{0} \tag{47}$$

$$\frac{\partial \mathcal{L}(\mathbf{w}, b, \boldsymbol{\lambda})}{\partial b} = \mathbf{0} \tag{48}$$

$$\lambda_i \geq 0, \ i = 1,2, \dots N \tag{49}$$

$$\lambda_i[d_i(\mathbf{w_0}^T\mathbf{x}_i + b_0) - 1] = 0, i = 1,2, \dots N \tag{50}$$

Με αντικατάσταση των συνθηκών (47) και (47) στην σχέση (46) προκύπτει

$$\mathbf{w} = \sum_{i=1}^{N} \lambda_i d_i \, \mathbf{x}_i \tag{51}$$

$$\sum_{i=1}^{N} \lambda_i d_i = 0 \tag{52}$$

Σ' αυτό το σημείο είναι απαραίτητο να υπολογιστούν οι παράμετροι που έχουν υπεισέλθει στον ορισμό του προβλήματος που αναφέρθηκε προηγουμένως. Λόγω της κυρτότητας της συνάρτησης κόστους και της γραμμικότητας των περιορισμών είναι δυνατόν να χρησιμοποιηθεί η λεγόμενη *Λαγκραντζιανή δυϊκότητα (Lagrangian duality)*. Το πρωτεύον πρόβλημα δηλαδή μετασχηματίζεται στο δυϊκό του στον χώρο των πολλαπλασιαστών Lagrange, το οποίο έχει την ίδια βέλτιστη τιμή με το πρωτεύον όμως τη λύση παρέχουν οι πολλαπλασιαστές Lagrange. Σύμφωνα με το θεώρημα περί δυϊσμού, για να είναι το $\mathbf{w}_0$ μια βέλτιστη λύση στο πρωτεύον πρόβλημα και το $\boldsymbol{\lambda}_0$ μια βέλτιστη λύση για το δυϊκό πρόβλημα, η ικανή και αναγκαία συνθήκη είναι το $\mathbf{w}_0$ να είναι εφικτό για το πρωτεύον πρόβλημα και

$$\Phi(\mathbf{w}_0) = \mathcal{L}(\mathbf{w}_0, b_0, \boldsymbol{\lambda_0}) = \min_{\mathbf{w}} \mathcal{L}(\mathbf{w}, b, \boldsymbol{\lambda}) \tag{53}$$

Προκειμένου να διατυπωθεί το δυϊκό πρόβλημα στην παρούσα περίπτωση, η εξίσωση (46) αναλύεται περεταίρω:

$$\mathcal{L}(\mathbf{w}, b, \boldsymbol{\lambda}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{N} \lambda_i \, d_i \mathbf{w}_0{}^T \mathbf{x}_i - b\sum_{i=1}^{N} \lambda_i d_i + \sum_{i=1}^{N} \lambda_i \tag{53}$$

από τη συνθήκη βελτιστότητας (52) ο τρίτος όρος του δεξιού μέλους μηδενίζεται. Ακόμη από τη σχέση (51) προκύπτει

$$\mathbf{w}^T\mathbf{w} = \sum_{i=1}^{N} \lambda_i d_i \, \mathbf{w}^T\mathbf{x}_i = \sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_i\lambda_j d_i d_j \, \mathbf{x}_i{}^T\mathbf{x}_j \tag{54}$$

Θέτοντας την αντικειμενική συνάρτηση $\mathcal{L}(\mathbf{w}, b, \boldsymbol{\lambda}) = Q(\boldsymbol{\lambda})$ ώστε να γίνει πιο εμφανής ο μετασχηματισμός του προβλήματος στο δυϊκό του, η σχέση (53) αναδιατυπώνεται ως εξής

$$Q(\boldsymbol{\lambda}) = \sum_{i=1}^{N} \lambda_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_i\lambda_j d_i d_j \, \mathbf{x}_i{}^T\mathbf{x}_j \tag{55}$$

Άρα, το πρόβλημα βελτιστοποίησης αναδιατυπώνεται πλέον ως εξής:

Δοθέντος ενός δείγματος εκπαίδευσης $\mathcal{T} = \{(\mathbf{x_i}, d_i)\}_{i=1}^{N}$, *να βρεθούν οι πολλαπλασιαστές Lagrange* $\{\boldsymbol{\lambda}_i\}_{i=1}^{N}$ *που μεγιστοποιούν την αντικειμενική συνάρτηση*

$$Q(\boldsymbol{\lambda}) = \sum_{i=1}^{N} \lambda_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_i\lambda_j d_i d_j \, \mathbf{x}_i{}^T\mathbf{x}_j$$

υπό τους περιορισμούς

(1) $$\sum_{i=1}^{N} \lambda_i d_i = 0$$

(2) $$\lambda_i \geq 0, \; i = 1,2, \ldots N$$

Στο σημείο αυτό, η χρησιμότητα του μετασχηματισμού του προβλήματος είναι πιο σαφής. Το δυϊκό πρόβλημα, σε αντίθεση με το πρωτεύον, βασίζεται μόνο στα δεδομένα εκπαίδευσης. Επιπλέον, η συνάρτηση $Q(\boldsymbol{\lambda})$ που πρέπει να μεγιστοποιηθεί εξαρτάται μόνο από τα πρότυπα εισόδου με τη μορφή ενός αθροίσματος εσωτερικών γινομένων. Τα διανύσματα υποστήριξης στην πλειονότητα των περιπτώσεων αποτελούν ένα υποσύνολο μόνο του συνόλου εκπαίδευσης, επομένως το διάνυσμα των λύσεων είναι αραιό. Στον περιορισμό (2) του δυϊκού προβλήματος η ανισότητα ισχύει για τα διανύσματα υποστήριξης ενώ η ισότητα ισχύει για όλα τα άλλα διανύσματα εκπαίδευσης, οπότε όλοι οι πολλαπλασιαστές Lagrange είναι μηδέν. Επομένως, υπολογίζοντας τους βέλτιστους πολλαπλασιαστές Lagrange, $\lambda_{o,i}$, προκύπτει από τη σχέση (51) το βέλτιστο διάνυσμα βαρών

$$\mathbf{w}_0 = \sum_{i=1}^{N_s} \lambda_{0,i} d_i \, \mathbf{x}_i \tag{56}$$

όπου N_s το πλήθος των διανυσμάτων υποστήριξης. Η βέλτιστη πόλωση b_0 υπολογίζεται με αντικατάσταση του $\mathbf{w}_0$ που υπολογίστηκε προηγουμένως στη σχέση (43).

Σε μία μηχανή διανυσμάτων υποστήριξης, όπως περιγράφηκε παραπάνω, επιβάλλεται μία δομή στο σύνολο των υπερεπιπέδων διαχωρισμού που προκύπτουν μέσω περιορισμών

(ελαχιστοποίηση) στην Ευκλείδεια νόρμα του διανύσματος βαρών. Η δομή αυτή περιγράφεται από το ακόλουθο θεώρημα (Vapnik, 1995,1998)

> *Έστω D η διάμετρος της μικρότερης σφαίρας που περιέχει όλα τα διανύσματα εισόδου* $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$. *Το σύνολο των βέλτιστων υπερεπιπέδων που περιγράφεται από την εξίσωση*
>
> $$\mathbf{w_0}^T\mathbf{x} + b_0 = 0$$
>
> *έχει διάσταση VC (Vapnik-Chervonekis)*[2]*, h, με άνω φράγμα*
>
> $$h \leq \min\left\{\left\lceil\frac{D^2}{\rho^2}\right\rceil, m_0\right\} + 1$$
>
> *όπου το* $\lceil\cdot\rceil$ *συμβολίζει τον μικρότερο ακέραιο που είναι μεγαλύτερος ή ίσος με τον αριθμό που περικλείει,* ρ *είναι το περιθώριο διαχωρισμού, ίσο με* $2/\|\mathbf{w}_0\|$*, και* m_0 *είναι η διαστατικότητα του χώρου εισόδου.*

Επομένως, σύμφωνα με το παραπάνω θεώρημα, είναι δυνατό μέσω της σωστής επιλογής του περιθωρίου διαχωρισμού ρ να ελεγχθεί η πολυπλοκότητα του βέλτιστου υπερεπιπέδου, δηλαδή η διάσταση VC, ανεξάρτητα από τη διαστατικότητα m_0 του χώρου εισόδου.

Μη γραμμικά Διαχωρίσιμα Πρότυπα

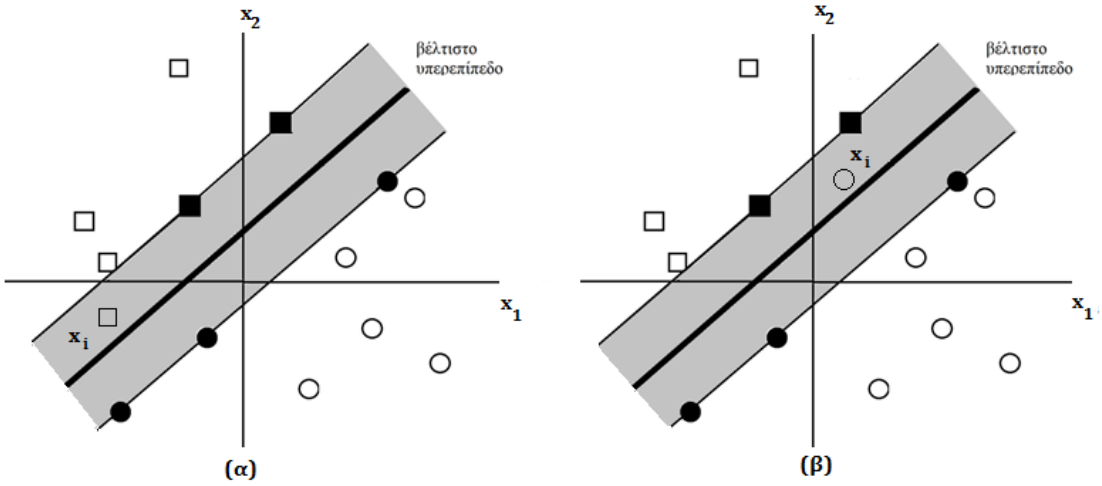
Στην περίπτωση των μη γραμμικά διαχωρίσιμων προτύπων, το σύνολο εκπαίδευσης είναι τέτοιο που δεν επιτρέπει την κατασκευή ενός υπερεπιπέδου διαχωρισμού χωρίς την να υπεισέλθουν σφάλματα στην κατηγοριοποίηση. Το έργο της κατηγοριοποίησης περιορίζεται στην εύρεση ενός βέλτιστου υπερεπιπέδου το οποίο να ελαχιστοποιεί την πιθανότητα σφάλματος, υπολογισμένου επί του συνόλου του συνόλου εκπαίδευσης.

Το περιθώριο διαχωρισμού μεταξύ των κλάσεων χαρακτηρίζεται *ελαστικό (soft)* αν υπάρχει ένα σημείο δεδομένων $(\mathbf{x}_i, d_i)$ το οποίο παραβιάζει τη συνθήκη της σχέσης (42):

$$d_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1, \text{ για } i = 1,2, \ldots N$$

Η παραβίαση της συνθήκης αυτής μπορεί να προκύψει με δύο τρόπους. Είτε το διάνυσμα $(\mathbf{x}_i, d_i)$ εμπίπτει στην περιοχή διαχωρισμού και στη σωστή πλευρά της επιφάνειας απόαφασης (σχήμα 12α) οπότε η κατηγοριοποίηση είναι σωστή, είτε το διάνυσμα $(\mathbf{x}_i, d_i)$ εμπίπτει στη λάθος πλευρά της επιφάνειας απόφασης (σχήμα 12β) οπότε η κατηγοριοποίηση αποτυγχάνει. Λόγω του σφάλματος που υπεισέρχεται στην κατηγοριοποίηση στην περίπτωση που εξετάζουμε είναι επόμενο η αφαίρεση από το σύνολο εκπαίδευσης κάποιου από τα παραπάνω σημεία τα οποία προκαλούν το σφάλμα αυτό να προκαλέσει και την αλλαγή της επιφάνειας απόφασης.

[2] Η *διάσταση VC* είναι μία παράμετρος η οποία αποτελεί μέτρο της χωρητικότητας ή εκφραστικής ισχύος (expressive power) μιας οικογένειας δυαδικών συναρτήσεων ταξινόμησης που υλοποιούνται από μία μηχανή μάθησης. Όσο πιο πολλά φαινόμενα μπορούν να περιγραφούν από την οικογένεια συναρτήσεων τόσο μεγαλύτερη είναι η τιμή της h. Στη συγκεκριμένη περίπτωση η διάσταση VC αποτελεί μέτρο της πολυπλοκότητας του χώρου συναρτήσεων.



Σχήμα 12: Υπερεπίπεδο ελαστικού περιθωρίου διαχωρισμού σε διδιάστατο χώρο.
(α) Το σημείο δεδομένων $\mathbf{x}_i$ βρίσκεται στη σωστή πλευρά του υπερεπιπέδου αλλά εντός της περιοχής διαχωρισμού.
(β) Το σημείο δεδομένων $\mathbf{x}_i$ βρίσκεται στη λάθος πλευρά του υπερεπιπέδου

Για την εισαγωγή μίας ενιαίας περιγραφής τόσο των δύο παραπάνω προβληματικών περιπτώσεων όσο και της περίπτωσης ενός διανύσματος που ικανοποιεί τον περιορισμό της σχέσης (42) εισάγεται στον ορισμό του διαχωριστικού υπερεπιπέδου ένα καινούριο σύνολο μη αρνητικών μεταβλητών $\{\xi_i\}_{i=1}^{N}$ οι οποίες ονομάζονται *μεταβλητές χαλάρωσης (slack variables)* και εισάγονται ως εξής:

$$d_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \text{για } i = 1,2,\dots N \tag{57}$$

Οι μεταβλητές αυτές αποτελούν ουσιαστικά το μέτρο της απόκλισης ενός σημείου δεδομένων από την ιδανική συνθήκη διαχωρισιμότητας των προτύπων. Αν $0 < \xi_i \leq 1$, πρόκειται για ένα διάνυσμα εντός της περιοχής διαχωρισμού αλλά στη σωστή πλευρά της επιφάνειας απόφασης (σχήμα 4α). Αν $\xi_i > 1$, πρόκειται για ένα διάνυσμα στη λάθος πλευρά της επιφάνειας απόφασης (σχήμα 4β). Αν $\xi_i = 0$ η εξίσωση (57) μεταπίπτει στην περίπτωση της προηγούμενης ενότητας των γραμμικά διαχωρίσιμων προτύπων, πρόκειται δηλαδή για τα σημεία που είναι σωστά ταξινομημένα και βρίσκονται εκτός της περιοχής διαχωρισμού. Τα διανύσματα υποστήριξης ορίζονται επομένως όπως και στην προηγούμενη περίπτωση και είναι τα σημεία για τα οποία ισχύει η ισότητα της σχέσης (57) με $\xi_i = 0$. Ο στόχος του κατηγοριοποιητή είναι επομένως να προσδιοριστεί ένα υπερεπίπεδο το οποίο να μεγιστοποιεί το περιθώριο διαχωρισμού περιλαμβάνοντας ταυτόχρονα στην περιοχή διαχωρισμού το μικρότερο δυνατό αριθμό διανυσμάτων για τα οποία να ισχύει $\xi_i > 0$.
Προκειμένου να διατυπωθεί φορμαλιστικά το πρόβλημα και να επιτευχθεί αυτός ο στόχος ορίζουμε την συναρτησιακή εξίσωση κόστους

$$\Phi(\xi) = \sum_{i=1}^{N} I(\xi_i - 1)$$

Η παραπάνω συνάρτηση πρέπει να ελαχιστοποιηθεί ως προς το διάνυσμα βαρών $\mathbf{w}$ υπό τον περιορισμό της σχέσης (57). Η συνάρτηση $I(\xi)$ είναι μία *συνάρτηση δείκτης (indicator function)* η οποία ορίζεται ως εξής

$$I(\xi) = \begin{cases} 0, & \xi \leq 0 \\ 1, & \xi > 0 \end{cases}$$

Η ελαχιστοποίηση του συναρτησιακού $\Phi(\xi)$ είναι ένα πρόβλημα ελαχιστοποίησης μιας μη κυρτής συνάρτησης, το οποίο ανήκει στην κλάση των NP-πλήρων προβλημάτων, είναι επομένως αδύνατη η λύση του σε ικανοποιητικό χρόνο. Προκειμένου να αρθεί αυτή η δυσκολία το συναρτησιακό προσεγγίζεται ως εξής

$$\Phi(\xi) = \sum_{i=1}^{N} \xi_i$$

ενώ αν αναδιατυπωθεί το συναρτησιακό ώστε να διευκολύνεται η ελαχιστοποίησή του ως προς το διάνυσμα βαρών **w**

$$\Phi(\mathbf{w}, \xi) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N} \xi_i \tag{58}$$

Η ελαχιστοποίηση του πρώτου όρου σχετίζεται με τα διανύσματα υποστήριξης ενώ το άθροισμα του δεύτερου όρου αποτελεί το άνω φράγμα του αριθμού εσφαλμένων ταξινομήσεων των προτύπων εκπαίδευσης. Η παράμετρος *C* καθορίζει ποιος από τους δύο όρους υπερισχύει στην συνάρτηση, επιλέγεται από τον χρήστη και προσδιορίζεται είτε πειραματικά είτε με κάποια άλλη μέθοδο (π.χ. διασταυρωμένης επικύρωσης, cross validation). Όταν το σύνολο εκπαίδευσης θεωρείται αξιόπιστο, τότε η τιμή της σταθεράς αυτής προκύπτει μεγάλη. Αντίθετα, αν θεωρηθεί ότι το σύνολο εκπαίδευσης μπορεί να περιέχει θόρυβο τότε η σταθερά παίρνει πιο μικρές τιμές.
Θα χρησιμοποιηθεί όπως και στην περίπτωση των γραμμικά διαχωρίσιμων προτύπων ο μετασχηματισμός του πρωτεύοντος προβλήματος ελαχιστοποίησης του συναρτησιακού (58) υπό τον περιορισμό της σχέσης (57) στο δυϊκό του στο χώρο των πολλαπλασιαστών Lagrange. Το πρωτεύον πρόβλημα επομένως ορίζεται φορμαλιστικά ως εξής:

Δοθέντος ενός δείγματος εκπαίδευσης $\mathcal{T} = \{(\mathbf{x_i}, d_i)\}_{i=1}^{N}$ *να βρεθούν βέλτιστες τιμές του διανύσματος βαρών* **w** *και της πόλωσης b ώστε να ικανοποιούν τον περιορισμό*

$$d_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \text{ για } i = 1,2,\dots N$$

$$\xi_i \geq 0, \text{ για } i = 1,2,\dots N$$

και επίσης τέτοιες ώστε το διάνυσμα **w** *και οι μεταβλητές* ξ_i *να ελαχιστοποιούν τη συναρτησιακή εξίσωση κόστους*

$$\Phi(\mathbf{w}, \xi) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N} \xi_i$$

όπου C παράμετρος που καθορίζεται από τον χρήστη.

Προχωρώντας στον μετασχηματισμό του προβλήματος στο δυϊκό του ορίζουμε την κυρτή συνάρτηση Lagrange του προβλήματος

$$\mathcal{L}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N} \xi_i - C\sum_{i=1}^{N} \mu_i\xi_i - \sum_{i=1}^{N} \lambda_i \left[d_i(\mathbf{w_0}^T\mathbf{x}_i + b_0) - 1 + \xi_i\right] \tag{59}$$

όπου οι μεταβλητές λ_i είναι οι πολλαπλασιαστές Lagrange και $\boldsymbol{\lambda}$ είναι το αντίστοιχο διάνυσμα των πολλαπλασιαστών Lagrange ενώ μ_i είναι οι πολλαπλασιαστές Lagrange που

αντιστοιχούν στις μεταβλητές ξ_i και $\boldsymbol{\mu}$ το αντίστοιχο διάνυσμα των πολλαπλασιαστών Lagrange. Οι συνθήκες βελτιστότητας ΚΚΤ που πρέπει να ικανοποιεί ο ελαχιστοποιητής είναι οι παρακάτω:

$$\frac{\partial \mathcal{L}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\mu})}{\partial \mathbf{w}} = \mathbf{0} \tag{60}$$

$$\frac{\partial \mathcal{L}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\mu})}{\partial b} = 0 \tag{61}$$

$$\frac{\partial \mathcal{L}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\mu})}{\partial \xi} = 0 \tag{62}$$

$$\mu_i \xi_i = 0, \ i = 1,2, \dots N \tag{63}$$

$$\mu_i \geq 0 \text{ και } \lambda_i \geq 0, \ i = 1,2, \dots N \tag{64}$$

$$\lambda_i \left[d_i (\mathbf{w}_0{}^T \mathbf{x}_i + b_0) - 1 + \xi_i \right] = 0, \ i = 1,2, \dots N \tag{65}$$

Με εφαρμογή των συνθηκών (60), (61) και (62) στην σχέση (59) προκύπτουν αντίστοιχα οι σχέσεις:

$$\mathbf{w} = \sum_{i=1}^{N} \lambda_i d_i \, \mathbf{x}_i \tag{66}$$

$$\sum_{i=1}^{N} \lambda_i d_i = 0 \tag{67}$$

$$C - \mu_i - \lambda_i = 0, i = 1,2, \dots N \tag{68}$$

Εξετάζοντας ιδιαίτερα την περίπτωση των προβληματικών διανυσμάτων ($\xi_i > 0$), λύνοντας τις συνθήκες ΚΚΤ για $\xi_i \neq 0$ για τους αντίστοιχους πολλαπλασιαστές προκύπτει $\mu_i = 0$ και $\lambda_i = C$. Αυτό το αποτέλεσμα σημαίνει ότι τα προβληματικά αυτά σημεία έχουν τη μέγιστη επιρροή στην τελική λύση για το διάνυσμα βαρών **w.**

Με αντικατάσταση των σχέσεων (66), (67) και (68) στη Λαγκραντζιανή της σχέσης (59) και θέτοντας την αντικειμενική συνάρτηση $\mathcal{L}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = Q(\boldsymbol{\lambda})$ προκύπτει η αντικειμενική συνάρτηση που θα μεγιστοποιηθεί στα πλαίσια του δυϊκού προβλήματος:

$$Q(\boldsymbol{\lambda}) = \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j d_i d_j \, \mathbf{x}_i{}^T \mathbf{x}_j \tag{69}$$

Άρα, το πρόβλημα βελτιστοποίησης αναδιατυπώνεται ως εξής:

Δοθέντος ενός δείγματος εκπαίδευσης $\mathcal{T} = \{(\mathbf{x}_i, d_i)\}_{i=1}^{N}$, *να βρεθούν οι πολλαπλασιαστές Lagrange* $\{\boldsymbol{\lambda}_i\}_{i=1}^{N}$ *που μεγιστοποιούν την αντικειμενική συνάρτηση*

$$Q(\boldsymbol{\lambda}) = \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j d_i d_j \, \mathbf{x}_i{}^T \mathbf{x}_j$$

υπό τους περιορισμούς

$$\sum_{i=1}^{N} \lambda_i d_i = 0 \quad (1)$$

$$0 \leq \lambda_i \leq C, \ i = 1,2, \dots N \quad (2)$$

όπου C είναι μια καθοριζόμενη από τον χρήστη θετική παράμετρος.

Από τη σχέση (67) προκύπτει το διάνυσμα βαρών **w** όπως και στην περίπτωση των διαχωρίσιμων προτύπων και στη συνέχεια υπολογίζεται η πόλωση b.

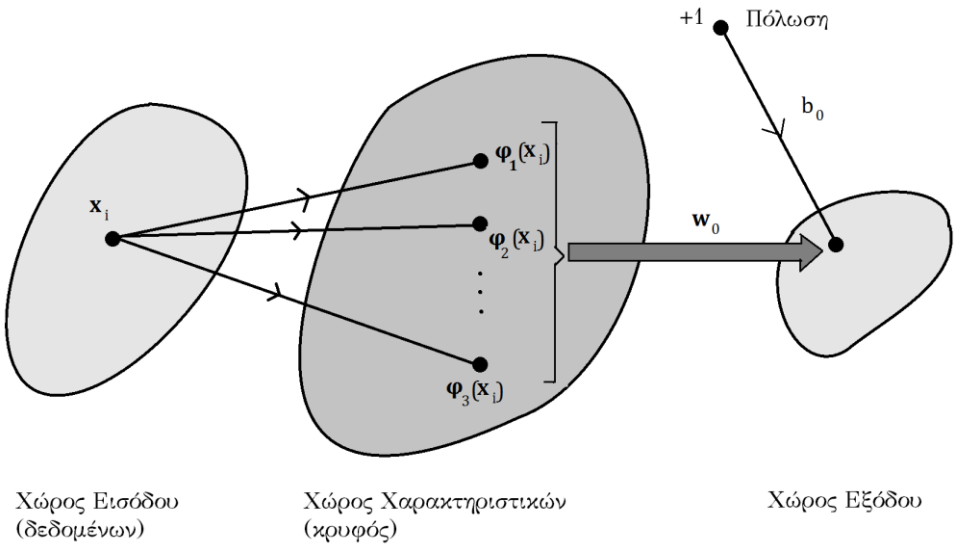
Όπως έχει αναφερθεί προηγουμένως, τα διανύσματα υποστήριξης προσδιορίζονται με τον ίδιο τρόπο είτε πρόκειται για διαχωρίσιμα είτε μη διαχωρίσιμα πρότυπα. Συγκρίνοντας τώρα την εξίσωση (69) με την αντίστοιχη του διαχωρίσιμου προβλήματος εξίσωση (55) προκύπτει ότι η συνάρτηση που πρέπει να μεγιστοποιηθεί είναι επίσης η ίδια και στις δύο περιπτώσεις. Η μεταβλητές ξ_i καθώς και οι αντίστοιχοι πολλαπλασιαστές Lagrange μ_i απαλείφονται από την αντικειμενική συνάρτηση. Η επίδρασή τους μεταφέρεται στη σταθερά C και η διαφορά μεταξύ των δύο περιπτώσεων έγκειται στον περιορισμό που αφορά στους πολλαπλασιαστές Lagrange λ_i. Στην περίπτωση των μη διαχωρίσιμων προτύπων οι πολλαπλασιαστές λ_i περιορίζονται από τη σταθερά C, περιορισμός ο οποίος στην απλή περίπτωση δεν υπάρχει, μπορεί δηλαδή να θεωρηθεί ότι $C \to \infty$.

Το επόμενο σημείο στο οποία πρέπει να γίνει ιδιαίτερη αναφορά είναι το υπολογιστικό φορτίο που δημιουργούν οι μηχανές διανυσμάτων υποστήριξης. Οι διαδικασίες της εκπαίδευσης και του ελέγχου είναι αυξημένης τόσο χρονικής όσο και υπολογιστικής πολυπλοκότητας. Όσον αφορά το στάδιο της εκπαίδευσης, όταν το σύνολο εκπαίδευσης είναι σχετικά μικρό οι συνήθεις αλγόριθμοι βελτιστοποίησης γενικής χρήσης αρκούν για να έρθει εις πέρας η εργασία αυτή. Για μεγάλα σύνολα εκπαίδευσης όμως (της τάξης μερικών χιλιάδων διανυσμάτων) πρέπει να υπάρξει ιδιαίτερη αντιμετώπιση. Το σύνολο εκπαίδευσης αποσυντίθεται με τέτοιο τρόπο ώστε τα υποσύνολα που δημιουργούνται να μπορούν να χωρέσουν στην διαθέσιμη μνήμη (Bose 1997, Osun 1997, Chan 2000). Επί των υποσυνόλων που προκύπτουν εφαρμόζεται κάποιος αλγόριθμος βελτιστοποίησης γενικής χρήσης ενώ όσο προχωρά η διαδικασία από το ένα υποσύνολο στο επόμενο τα διανύσματα υποστήριξης ενημερώνονται ώστε να ισχύουν οι περιορισμοί του προβλήματος στο σύνολο που αποτελούν τα υποσύνολα που έχουν μέχρι στιγμής υποστεί επεξεργασία. Παρόμοια προβλήματα προκύπτουν και στο στάδιο του ελέγχου για μεγάλα σύνολα προτύπων ελέγχου τα οποία αντιμετωπίζονται με κατάλληλες τεχνικές που έχουν προταθεί (Burg 1997).

Μηχανή Πυρήνα

Μετά από την μελέτη όλων των παραπάνω δεδομένων είναι πλέον δυνατή η φορμαλιστική περιγραφή της διαδικασίας κατασκευής μίας μηχανής διανυσμάτων υποστήριξης για την κατηγοριοποίηση προτύπων. Η διαδικασία αυτή αποτελείται από δύο επιμέρους μαθηματικές διαδικασίες όπως φαίνεται στο σχήμα 13. Αρχικά γίνεται αντιστοίχιση ενός διανύσματος εισόδου σε ένα χώρο χαρακτηριστικών μεγαλύτερης διαστατικότητας. Η αντιστοίχιση αυτή είναι μη γραμμική. Επίσης, ο ενδιάμεσος χώρος χαρακτηριστικών είναι εσωτερικός του συστήματος, δηλαδή άγνωστος πρός την είσοδο και την έξοδο. Η διαστατικότητά του χώρου αυτού καθορίζεται από τον αριθμό των

διανυσμάτων υποστήριξης ώστε να διασφαλίζεται η βελτιστότητα της κατηγοριοποίησης. Το επόμενο στάδιο περιλαμβάνει την κατασκευή του βέλτιστου υπερεπιπέδου για τον διαχωρισμό των διανυσμάτων του χώρου χαρακτηριστικών που προσδιορίστηκαν στο προηγούμενο βήμα.



Σχήμα 13: Απεικόνιση των λειτουργιών που συγκροτούν τη διαδικασία κατασκευής μιας μηχανής διανυσμάτων υποστήριξης για την αναγνώριση προτύπων. Αρχικά γίνεται η μη γραμμική αντιστοίχιση του χώρου εισόδου στο χώρο χαρακτηριστικών και στη συνέχεια η γραμμική αντιστοίχιση από το χώρο χαρακτηριστικών στο χώρο εισόδου.

Το πρώτο βήμα ακολουθεί το *θεώρημα του Cover* για τη γραμμική διαχωρισιμότητα των προτύπων. Αν το διαθέσιμο σύνολο των προτύπων είναι μη διαχωρίσιμο. Σύμφωνα με το θεώρημα Cover ένας τέτοιος πολυδιάστατος χώρος μπορεί να μετασχηματιστεί σε ένα νέο χώρο χαρακτηριστικών όπου τα πρότυπα είναι γραμμικά διαχωρίσιμα με μεγάλη πιθανότητα υπό δύο συνθήκες. Η πρώτη συνθήκη είναι ότι ο μετασχηματισμός είναι μη γραμμικός ενώ η δεύτερη υπαγορεύει ότι η διαστατικότητα του χώρου χαρακτηριστικών είναι αρκετά μεγάλη. Βέβαια, το θεώρημα Cover δεν εξετάζει την βελτιστότητα του διαχωριστικού υπερεπιπέδου. Αυτή η έλλειψη καλύπτεται στο δεύτερο βήμα, όπου το βέλτιστο υπερεπίπεδο υπολογίζεται κατά τα γνωστά, με τη βασική διαφορά όμως ότι το διαχωριστικό υπερεπίπεδο ορίζεται ως μία γραμμική συνάρτηση στο χώρο χαρακτηριστικών αντί του αρχικού χώρου εισόδου. Η κατασκευή του γίνεται με τη χρήση του λεγόμενου πυρήνα εσωτερικού γινομένου που εξηγείται στη συνέχεια.

Ακολουθώντας αυτή τη λογική ορίζουμε ως $\mathbf{x}$ ένα διάνυσμα του χώρου εισόδου διάστασης m_0, ως $\{\varphi_j(\mathbf{x})\}_{j=1}^{\infty}$ ένα σύνολο μη γραμμικών συναρτήσεων οι οποίες μετασχηματίζουν τον χώρο διάστασης m_0 στο χώρο χαρακτηριστικών άπειρων διαστάσεων και ως $\{w_j\}_{j=1}^{\infty}$ το απείρως μεγάλο σύνολο βαρών που μετασχηματίζει τον χώρο χαρακτηριστικών στο χώρο εξόδου. Χρησιμοποιώντας αυτό το μετασχηματισμό ο ορισμός ενός υπερεπιπέδου το οποίο δρα ως επιφάνεια απόφασης διαμορφώνεται ως εξής

$$\sum_{j=1}^{\infty} w_j \varphi_j(\mathbf{x}) = 0 \tag{70}$$

Η απόφαση για την κλάση στην οποία θα καταταχθεί το διάνυσμα εισόδου $\mathbf{x}$ λαμβάνεται στον χώρο εξόδου. Η εξίσωση (70) σε περιγραφή πίνακα γράφεται

$$\mathbf{w}^T \boldsymbol{\varphi}(\mathbf{x}) = 0 \tag{71}$$

όπου το $\boldsymbol{\varphi}(\mathbf{x})$ είναι το διάνυσμα χαρακτηριστικών και $\mathbf{w}$ το διάνυσμα βαρών. Προσαρμόζοντας τη σχέση (56) στην παρούσα περίπτωση προκύπτει

$$\mathbf{w} = \sum_{i=1}^{N_s} \lambda_i d_i \, \boldsymbol{\varphi}(\mathbf{x}_i) \tag{72}$$

όπου N_s είναι το πλήθος των διανυσμάτων υποστήριξης και το διάνυσμα χαρακτηριστικών εκφράζεται ως

$$\boldsymbol{\varphi}(\mathbf{x}_i) = [\varphi_1(\mathbf{x}_i), \varphi_2(\mathbf{x}_i), \dots]^{\boldsymbol{T}} \tag{73}$$

Αντικαθιστώντας την εξίσωση (71) στην εξίσωση (72) προκύπτει μία νέα έκφραση για την επιφάνεια απόφασης στο χώρο εξόδου

$$\sum_{i=1}^{N_s} \lambda_i d_i \, \boldsymbol{\varphi}^T(\mathbf{x}_i)\boldsymbol{\varphi}(\mathbf{x}) = 0 \tag{74}$$

Ο βαθμωτός όρος $\boldsymbol{\varphi}^T(\mathbf{x}_i)\boldsymbol{\varphi}(\mathbf{x})$ της εξίσωσης αυτής αποτελεί ένα εσωτερικό γινόμενο. Ο όρος αυτός που συμβολίζεται με $k(\mathbf{x}, \mathbf{x}_i)$ ονομάζεται *πυρήνας εσωτερικού γινομένου (inner product kernel)* ή απλά *πυρήνας (kernel)* και ορίζεται ως (Shawe-Taylor, Christianini 2004):

> *Ο πυρήνας $k(\mathbf{x}, \mathbf{x}_i)$ είναι μία συνάρτηση η οποία υπολογίζει το εσωτερικό γινόμενο των εικόνων που παράγονται στο χώρο χαρακτηριστικών βάσει του διανύσματος $\boldsymbol{\varphi}$ δύο σημείων δεδομένων στο χώρο εισόδου*

Και ο αντίστοιχος μαθηματικός μαθηματικός ορισμός:

$$k(\mathbf{x}, \mathbf{x}_i) = \boldsymbol{\varphi}^T(\mathbf{x}_i)\boldsymbol{\varphi}(\mathbf{x}) = \sum_{j=1}^{\infty} \varphi_j(\mathbf{x}_i)\varphi_j(\mathbf{x}) \tag{75}$$

Η έκφραση της βέλτιστης επιφάνειας γίνεται επομένως

$$\sum_{i=1}^{N_s} \lambda_i d_i \, k(\mathbf{x}, \mathbf{x}_i) = 0 \tag{76}$$

Το σημαντικό όφελος αυτής της προσέγγισης είναι ότι με τον προσδιορισμό μιας τέτοιας συνάρτησης πυρήνα είναι δυνατόν να έρθει εις πέρας η εργασία της κατηγοριοποίησης προτύπων χωρίς να εμπλακεί καθόλου το διάνυσμα βαρών. Ο χώρος χαρακτηριστικών θεωρήθηκε απείρων διαστάσεων, η σχέση (74) όμως υποδεικνύει ότι ο αριθμός των διαστάσεων είναι ίσος με το πλήθος των των προτύπων εκπαίδευσης που χρησιμοποιεί-ται.

Έλεγχος της ικανότητας γενίκευσης των μηχανών διανυσμάτων υποστήριξης

Η ικανότητα γενίκευσης μιας μηχανής μάθησης εξαρτάται από δύο παράγοντες, το ρυθμό σφάλματος στο σύνολο εκπαίδευσης και τη χωρητικότητα της μηχανής όπως αυτή εκφράζεται μέσω της VC διάστασης. Η πιθανότητα λάθος κατηγοριοποίησης κάποιου προτύπου αγνώστου στη μηχανή (*σφάλμα ελέγχου*) είναι φραγμένη με την εξής μορφή:

Με πιθανότητα $1-\eta$ *ισχύει η ανισότητα*

$$R(\mathbf{w},b) \leq R_{train}(\mathbf{w},b) + \sqrt{\frac{h\log\left(\frac{2N}{h}\right) - \log\frac{\eta}{4}}{N}} \tag{77}$$

Στην σχέση (42), $R(\mathbf{w},b)$ η πιθανότητα σφάλματος ελέγχου, $R_{train}(\mathbf{w},b)$ η πιθανότητα σφάλματος εκπαίδευσης και h η διάσταση VC. Δηλαδή η σχέση αυτή υπαγορεύει ότι:

$$P(\text{σφάλμα ελέγχου}) \leq \text{Συχνότητα(σφάλμα εκπαίδευσης)} + \begin{matrix}\text{Διάστημα}\\ \text{Εμπιστοσύνης}\end{matrix}$$

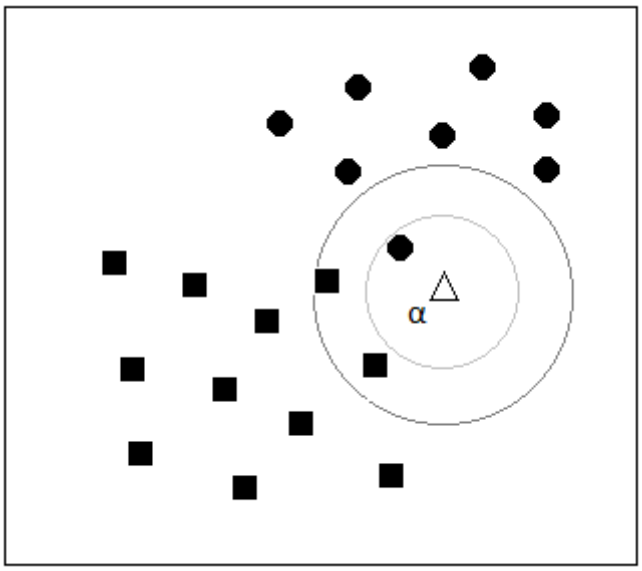
Στο φράγμα της παραπάνω σχέσης το διάστημα εμπιστοσύνης εξαρτάται από τη VC διάσταση της μηχανής, το πλήθος των προτύπων του συνόλου εκπαίδευσης και την τιμή του η. Οι δύο όροι του φράγματος είναι ανταγωνιστικοί μεταξύ τους επομένως πρέπει να γίνει ένας συμβιβασμός μεταξύ τους. Όσο μικρότερη είναι η VC διάσταση τόσο μικρότερο είναι το διάστημα εμπιστοσύνης και επίσης τόσο μεγαλύτερη είναι η συχνότητα εμφάνισης σφάλματος εκπαίδευσης. Για την διευθέτηση αυτού του προβλήματος υιοθετείται η αρχή της *ελαχιστοποίησης του δομικού κινδύνου (structural risk minimization)* η οποία έχει τις ρίζες της στη θεωρία της VC διάστασης και σύμφωνα με την οποία για την εύρεση λύσης αρκεί να ελαχιστοποιήσουμε το άθροισμα των δύο ανταγωνιστικών όρων.

2.1.3. Αλγόριθμος k-Κοντινότερων Γειτόνων (k-Nearest Neighbor)

Η λογική της κατηγοριοποίησης κοντινότερων γειτόνων η οποία είναι μη γραμμική είναι πολύ απλή καθώς το υπό κατηγοριοποίηση πρότυπο κατατάσσεται στην κλάση που είναι ανήκουν τα κοντινότερα σε αυτό πρότυπα εκπαίδευσης. Αν πρόκειται για την απλή κατηγοριοποίηση κοντινότερου γείτονα τότε το υπό κατηγοριοποίηση πρότυπο κατατάσσεται στην κλάση που ανήκει το κοντινότερο πρότυπο. Στην περίπτωση που το σύνολο εκπαίδευσης είναι μεγάλο, το απλό αυτό κριτήριο παρουσιάζει μάλιστα πολύ καλές επιδόσεις. Για περισσότερη ασφάλεια στην κατηγοριοποίηση όμως είναι δυνατόν να ληφθούν υπ' όψιν περισσότερα του ενός κοντινά πρότυπα του δείγματος εκπαίδευσης, έστω k το πλήθος. Σ' αυτό συνίσταται και η μέθοδος των *k-Κοντινότερων Γειτόνων (kΚΓ) (k-Nearest Neighbor, kNN).* Στην περίπτωση που τα k κοντινότερα πρότυπα δεν είναι όλα της ίδιας κλάσης τότε χρησιμοποιείται η ψήφος πλειοψηφίας είτε η ψήφος σταθμισμένων αποστάσεων. Το k επιλέγεται ώστε να μην είναι πολλαπλάσιο του πλήθους Μ των κλάσεων που υπάρχουν στο πρόβλημα. Όλα τα πρότυπα εκπαίδευσης πρέπει να είναι συνεχώς διαθέσιμα στην μνήμη κατά την εκτέλεση της κατηγοριοποίησης και για το λόγο αυτό πολλές φορές η μέθοδος αυτή αναφέρεται ως κατηγοριοποίηση *Βάσει της Μνήμης (Memory-Based)*. Άλλες ονομασίες της μεθόδου που αναφέρονται είναι Κατηγοριοποίηση *Βάσει της Περίπτωσης (Case-Based)* ή *Βάσει του Παραδείγματος (Example-Based)* καθώς η κατηγοριοποίηση βασίζεται άμεσα στα πρότυπα εκπαίδευσης.

Παρατηρώντας το σχήμα 14, το πρότυπο α με εφαρμογή του κριτηρίου κοντινότερου γείτονα θα καταταχθεί στην κλάση των κύκλων. Στο σχήμα 14, αν εφαρμοστεί κατηγοριοποίηση τριών κοντινότερων γειτόνων τότε το πρότυπο α θα καταταχθεί στην

κλάση των τετραγώνων, καθώς τα δύο από τα τρία κοντινότερα πρότυπα είναι τετράγωνα, χρησιμοποιείται η ψήφος πλειοψηφίας.



Σχήμα 14: Παράδειγμα ταξινόμησης κοντινότερων γειτόνων. Ο μικρός ανοιχτός γκρι κύκλος αναφέρεται στην ταξινόμηση κοντινότερου γείτονα ενώ ο μεγάλος σκούρος γκρι κύκλος αναφέρεται στην ταξινόμηση 3 κοντινότερων γειτόνων.

Το πρόβλημα της κατηγοριοποίησης kNN ορίζεται ως εξής: Έστω ότι διαθέτουμε ένα σύνολο εκπαίδευσης $\mathcal{T} = \{\mathbf{x}_i\}_{i=1}^{N}$, όπου N το πλήθος των προτύπων του συνόλου $\mathcal{T}$. Τα χαρακτηριστικά των προτύπων αυτών συγκροτούν σύνολο $\mathcal{F}$ και οι τιμές τους κανονικοποιούνται ώστε να βρίσκονται στο διάστημα [0,1]. Κάθε πρότυπο του $\mathcal{T}$ αντιστοιχίζεται σε μία κλάση y_j, $j = 1,2,\dots,M$, όπου M το πλήθος των κλάσεων. Το ζητούμενο είναι να ταξινομηθεί ένα άγνωστο πρότυπο $\boldsymbol{\alpha}$. Η κατηγοριοποίηση αποτελείται από δύο επιμέρους διαδικασίες, τον προσδιορισμό των k κοντινότερων γειτόνων και τον καθορισμό της κλάσης με χρήση των k κοντινότερων γειτόνων. Έτσι, αρχικά υπολογίζεται η απόσταση του $\boldsymbol{\alpha}$ από κάθε $\mathbf{x}_i$:

$$d(\boldsymbol{\alpha}, \mathbf{x}_i) = \sum_{f \in \mathcal{F}} w_f \, \delta(\alpha_f, x_{if}) \tag{78}$$

όπου w_f είναι ο συντελεστής βάρους του χαρακτηριστικού f. Υπάρχει μεγάλο ποικιλία δυνατοτήτων για το μετρικό δ. Με βάση αυτό το μετρικό υπολογίζονται οι k κοντινότεροι γείτονες του προτύπου $\boldsymbol{\alpha}$. Σύμφωνα με τον μαθηματικό ορισμό ένα μετρικό πρέπει να ικανοποιεί τα παρακάτω τέσσερα κριτήρια:

(1) $d(x,y) \geq 0$ (μη αρνητικότητα)

(2) $d(x,y) = 0 \Leftrightarrow x = y$ (ταυτότητα)

(3) $d(x,y) = d(y,x)$ (συμμετρικότητα)

(4) $d(x,z) \geq d(x,y) + d(y,z)$ (τριγωνική ανισότητα)

Είναι δυνατόν να κατασκευαστεί κατηγοριοποιητής kNN που να μην είναι μετρικό, κάποιες βελτιστοποιήσεις της μεθόδου για την βελτίωση της απόδοσης όμως απαιτούν τη χρήση μετρικού και κυρίως να ισχύει η τριγωνική ανισότητα.

Αν τα χαρακτηριστικά των προτύπων παίρνουν συνεχείς τιμές, χρησιμοποείται το μετρικό *απόστασης Minkowski*, ο γενικός τύπος του οποίου είναι

$$MD_p(\boldsymbol{\alpha}, \mathbf{x}_i) = \left(\sum_{f \in \mathcal{F}} \left| \alpha_f - \mathrm{x}_{if} \right|^p \right)^{\frac{1}{p}} \tag{79}$$

Αν επιλεχθεί p=1, τότε προκύπτει η λεγόμενη *απόσταση* L_1 ή *απόσταση Manhattan,* ενώ αν επιλεχθεί p=2 τότε προκύπτει η *απόσταση* L_2 η οποία είναι η γνωστή Ευκλείδεια απόσταση. Οι δύο αυτές περιπτώσεις είναι οι πιο συνήθεις, υπάρχει όμως δυνατότητα να χρησιμοποιηθούν και μεγαλύτερες τιμές του p. Όσο μεγαλύτερη είναι η τιμή του p τόσο μεγαλύτερη βαρύτητα δίνεται στα χαρακτηριστικά τα οποία διαφέρουν πολύ μεταξύ των δύο προτύπων. Μια ακόμη περίπτωση είναι να χρησιμοποιηθεί η απόσταση Minkowski για $p \to \infty$, η οποία συμβολίζεται ως L_∞ και ονομάζεται *απόσταση Chebyshev*:

$$MD_\infty(\boldsymbol{\alpha}, \mathbf{x}_i) = \max_{f \in \mathcal{F}} \left| \alpha_f - \mathrm{x}_{if} \right| \tag{80}$$

Η απόσταση αυτή είναι ουσιαστικά η διαφορά μεταξύ των τιμών του χαρακτηριστικού στο οποίο τα δύο πρότυπα διαφέρουν περισσότερο μεταξύ τους.

Αν τώρα τα χαρακτηριστικά των προτύπων λαμβάνουν δυαδικές τιμές, μπορεί να χρησιμοποιηθεί η *απόσταση Hamming*

$$HD(\boldsymbol{\alpha}, \mathbf{x}_i) = \left| \{ j \mid \alpha_j \neq \mathrm{x}_{ij} \} \right| \tag{81}$$

η οποία ουσιαστικά υπολογίζει τον αριθμό των σημείων στα οποία τα διανύσματα των δύο πρότυπων διαφέρουν. Η απόσταση Hamming συμπίπτει με την απόσταση L_1 στην περίπτωση δυαδικών χαρακτηριστικών. Αν τα χαρακτηρστικά λαμβάνουν διακριτές τιμές χρησιμοποιείται το μέτρο

$$D(\boldsymbol{\alpha}, \mathbf{x}_i) = \sum_{f \in \mathcal{F}} d_f \left(\alpha_f, \mathrm{x}_{if} \right) \tag{82}$$

όπου d_f είναι κάποια συνάρτηση απόστασης για το κάθε χαρακτηριστικό. Έτσι, χρησιμοποιώντας κάποιο από τα παραπάνω μετρικά δημιουργείται μία υπερσφαίρα με κέντρο το διάνυσμα του προτύπου **α** ακτίνα την απόσταση του k-οστού μακρύτερου γείτονα από το πρότυπο **α.**

Ο τρόπος με τον οποίο θα προσδιοριστεί η κλάση του **α** αξιοποιώντας τη γνώση για την κλάση των k κοντινότερων γειτόνων που περικλείονται στην υπερσφαίρα παρουσιάζει επίσης ποικιλία. Η πιο απλή περίπτωση είναι να επιλεχθεί η κλάση κατά πλειοψηφία των k κοντινότερων γειτόνων. Μία πιο κομψή προσέγγιση είναι να σταθμιστούν οι ψήφοι των γειτόνων ως προς την απόσταση τους από το υπό κατηγοριοποίηση πρότυπο, δηλαδή τα πιο κοντινά στο **α** πρότυπα να επηρεάζουν περισσότερο το αποτέλεσμα. Οι ψήφοι που θα δίνονται σε κάθε κλάση y_j είναι επομένως

$$\mathrm{Vote}(y_j) = \sum_{c=1}^{k} \frac{1}{[d(\boldsymbol{\alpha}, \mathbf{x}_i)]^n} 1(y_j, y_c) \tag{83}$$

όπου

$$1(y_j, y_c) = \begin{cases} 1, & y_j = y_c \\ 0, & \text{αλλιώς} \end{cases}$$

και το n επιλέγεται από τον χρήστη ανάλογα με το πόσο επιθυμεί να μειωθεί η επίδραση των μακρινότερων από τους k κοντινότερους γείτονες (συνήθως ισούται με 1). Μία άλλη προσέγγιση [Shepard, 1987], χρησιμοποιεί εκθετική μείωση της επίδρασης των μακρύτερων γειτόνων αντί του αντιστρόφου της απόστασης:

$$\text{Vote}(y_j) = \sum_{c=1}^{k} e^{-\frac{d(\boldsymbol{\alpha},\mathbf{x}_i)}{h}} 1(y_j, y_c) \tag{84}$$

Έτσι, με κάποιο από τους τύπους (83) και (84) υπολογίζονται οι ψήφοι κάθε κλάσης και το πρότυπο $\boldsymbol{\alpha}$ κατατάσσεται στην κλάση με τις περισσότερες ψήφους.

Όσον αφορά τώρα στην στατιστική συμπεριφορά του κατηγοριοποιητή kNN, αυτός χαρακτηρίζεται *μη βέλτιστος (suboptimal)*. Παρ' όλα αυτά έχει γίνει πολύ δημοφιλής, καθώς ενώ δεν ανήκει στους κατηγοριοποιητές Bayes, ασυμπτωτικά, δηλαδή για $k \to \infty$, προσεγγίζει τη βέλτιστη συμπεριφορά ενός κατηγοριοποιητή Bayes. Ενδεικτικά για την περίπτωση δύο κλάσεων, η πιθανότητα σφάλματος του κατηγοριοποιητή kNN, P_{kNN}, προκύπτει

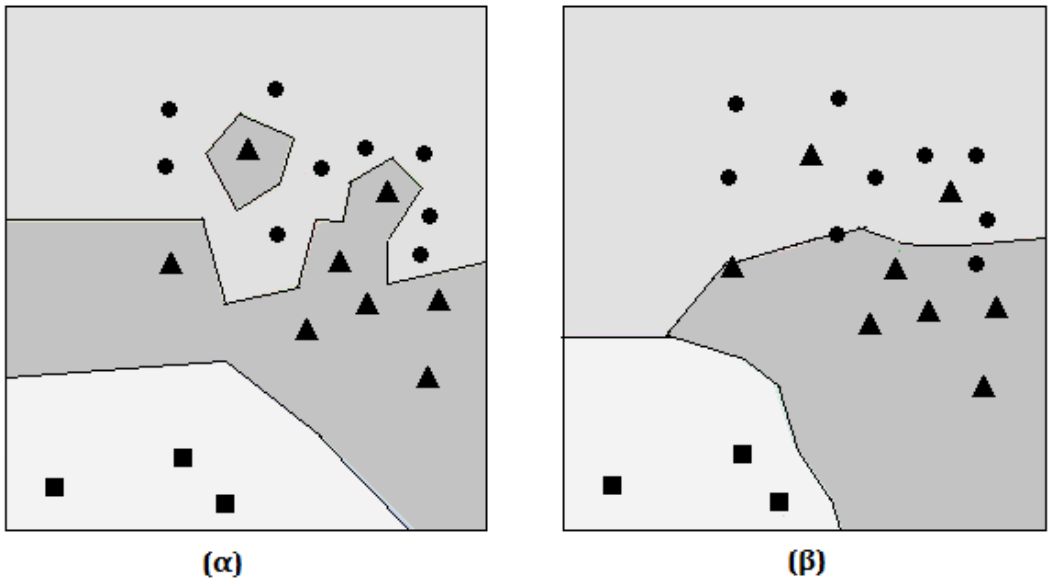
$$P_B \leq P_{kNN} \leq P_B + \frac{1}{\sqrt{ke}} \tag{85}$$

όπου P_B η πιθανότητα σφάλματος ενός κατηγοριοποιητή Bayes, η οποία ως γνωστόν είναι ελάχιστη. Από τη σχέση αυτή είναι εμφανές ότι για μεγάλες τιμές του k η συμπεριφορά του κατηγοριοποιητή προσεγγίζει τη βέλτιστη.

Για το συμπέρασμα αυτό μπορεί να δοθεί μία διαισθητική ερμηνεία. Αν υποτεθεί ότι διαθέτουμε μεγάλο σύνολο εκπαίδευσης, δηλαδή το N είναι αρκετά μεγάλο, η ακτίνα της υπερσφαίρας που έχει κέντρο το υπό κατηγοριοποίηση πρότυπο $\boldsymbol{\alpha}$ και περικλείει τους k κοντινότερους γείτονές του τείνει στο 0, καθώς σ' αυτή την περίπτωση τα πρότυπα θα είναι κατανεμημένα με μεγάλη πυκνότητα στο χώρο κι έτσι οι k κοντινότεροι γείτονες του $\boldsymbol{\alpha}$ θα βρίσκονται πολύ κοντά του. Επομένως, η οι δεσμευμένες πιθανότητες κάθε κλάσης ω_i θα είναι περίπου ίσες με την πιθανότητα $P(\omega_i|\boldsymbol{\alpha})$. Ακόμη, για μεγάλες τιμές του k, το οποίο όμως είναι απείρως μικρό ποσοστό του N, η πλειοψηφία των σημείων εντός της υπερσφαίρας θα ανήκουν στην κλάση με τη μεγαλύτερη δεσμευμένη πιθανότητα. Άρα, ο κατηγοριοποιητής kNN ασυμπτωτικά τείνει στον κατηγοριοποιητή Bayes. Βέβαια, στην πεπερασμένη περίπτωση είναι πολύ πιθανό η κατηγοριοποίηση κοντινότερου γείτονα $(k = 1)$ να παρουσιάζει μικρότερη πιθανότητα σφάλματος απ' ότι η κατηγοριοποίηση για μεγαλύτερο k. Ενδεικτικά παρουσιάζεται ένα συγκριτικό παράδειγμα στο σχήμα 15.

Υπολογιστική Πολυπλοκότητα

Ο σημαντικότερος περιορισμός της κατηγοριοποίησης kNN όμως είναι η πολυπλοκότητα της διαδικασίας προσδιορισμού των k κοντινότερων γειτόνων μεταξύ του συνόλου εκπαίδευσης. Η εξαντλητική αναζήτηση των γειτόνων έχει πολυπλοκότητα της τάξης του $O(kN)$, είναι δηλαδή ψευδοπολυωνυμικού χρόνου, η οποία είναι αρκετά κακή. Η λύση του προβλήματος γίνεται ακόμα πιο δύσκολη όταν ο χώρος των χαρακτηριστικών έχει πολλές διαστάσεις. Ακόμη, με χρήση ενός μετρικού απόστασης Minkowski η αναζήτηση εκτελείται σε χρόνο $O(|\mathcal{F}|N)$, οπότε και πάλι προκύπτει μία μεγάλη χρονική πολυπλοκότητα.



Σχήμα 15: Περιοχές απόφασης που δημιουργεί ο αλγόριθμος κοντινότερων γειτόνων kNN σε ένα δεδομένο σύνολο εκπαίδευσης με τρεις κλάσεις σε δύο διαφορετικές περιπτώσεις
(α) κατηγοριοποιητής ενός κοντινότερου γείτονα (k=1) (β) κατηγοριοποιητής 5 κοντινότερων γειτόνων

Για το λόγο αυτό, έχουν αναπτυχθεί πολλές τεχνικές τόσο για το φιλτράρισμα των προτύπων του συνόλου εκπαίδευσης, όσο και για την μείωση των διαστάσεων του διανύσματος χαρακτηριστικών. Όσον αφορά στην εύρεση των κοντινότερων γειτόνων, ακολουθεί μία συνοπτική παρουσίαση κάποιων στρατηγικών που έχουν προταθεί για την επιτάχυνση της διαδικασίας αυτής:

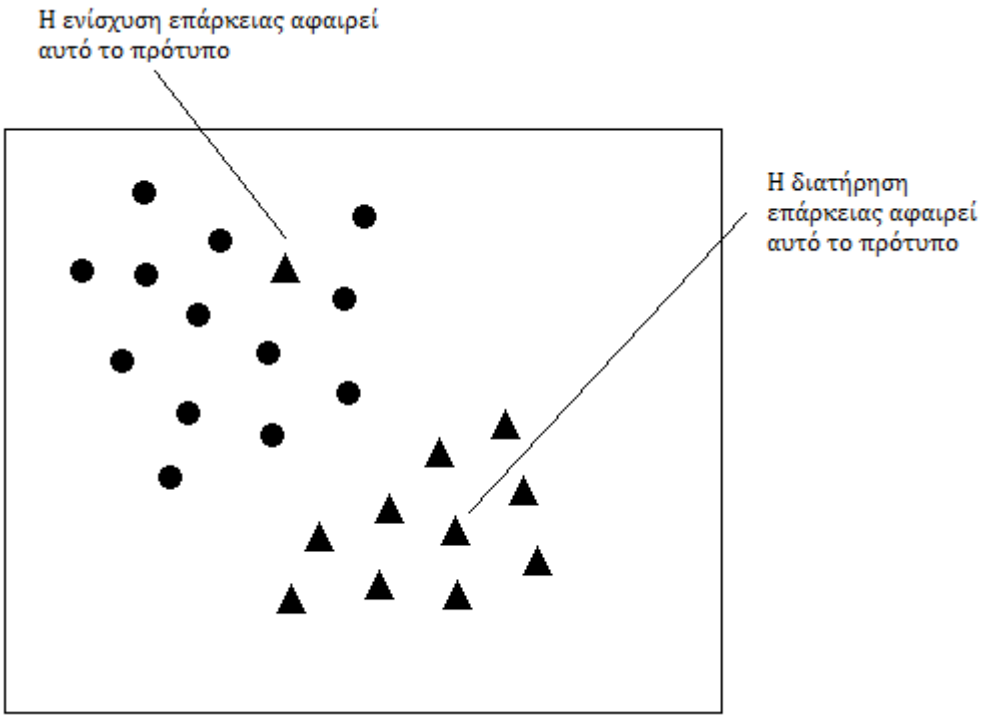
i. *Δίκτυο Ανάκτησης Περίπτωσεων (Case-Retrieval Net, CRN)*
   Η ανάκτηση kNN χρησιμοποιείται ευρέως στην Λογική Βάσει της Περίπτωσης (Case-Based Reasoning) και τα CRN είναι μία από τις δημοφιλέστερες τεχνικές για την βελτίωση του χρόνου ανάκτησης. Οι περιπτώσεις, δηλαδή τα πρότυπα εκπαίδευσης, υπόκεινται μία προεπεξεργασία ώστε να σχηματίσουν μία δομή δίκτυου η οποία χρησιμοποιείται κατά την εκτέλεση της κατηγοριοποίησης. Η διαδικασία της ανάκτησης γίνεται με *διαδιδόμενη ενεργοποίηση (spreading activation)* εντός της δικτυακής αυτής δομής. Είναι δυνατόν τα CRN να ρυθμιστούν έτσι ώστε να επιστρέφουν ακριβώς τόσες περιπτώσεις όσες οι k κοντινότεροι γείτονες [Lenz, Burkhard 1996].
ii. *Ανάκτηση Βάσει του Αποτυπώματος (Footprint-Based Retrieval)*
   Όπως όλες στρατηγικές του είδους, έτσι και αυτή περιλαμβάνει ένα στάδιο προεπεξεργασίας ώστε τα δεδομένα εκπαίδευσης να οργανωθούν σε μία ιεραρχία δύο επιπέδων στα οποία εφαρμόζεται μία διαδικασία ανάκτησης δύο σταδίων. Κατά την προεπεξεργασία κατασκευάζεται ένα μοντέλο επάρκειας το οποίο αναγνωρίζει τις περιπτώσεις-«απωτυπώματα» οι οποίες αποτελούν σημεία αναφοράς στο σύνολο εκπαίδευσης. Αυτή η διαδικασία δεν ανακτά απαραίτητα τα ίδια πρότυπα που θα ανακτούσε ο εξαντλητικός αλγόριθμος kNN, η βελτίωση όμως που προσφέρει στην ταχύτητα ανάκτησης σε συνδυασμό με την ποιότητα της είναι αρκετά εντυπωσιακή [Smyth 1999].
iii. *Ψάρεμα και Συρρίκνωση (Fish & Shrink)*

Αυτή η τεχνική απαιτεί το μέτρο της απόστασης που χρησιμοποιείται να είναι μετρικό με την αυστηρή έννοια και εκμεταλλεύεται την ιδιότητα της τριγωνικής ανισότητας ώστε να παράγει μία οργάνωση του συνόλου εκπαίδευσης σε πρότυπα που αποτελούν υποψήφιους γείτονες και πρότυπα που δεν λαμβάνονται υπ' όψιν. Πρότυπα που βρίσκονται μακριά από το υπό κατηγοριοποίηση πρότυπο μπορούν να αγνοηθούν και να μην συμπεριληφθούν στην διαδικασία ανάκτησης. Η μέθοδος αυτή έχει αποδειχθεί ότι είναι ισοδύναμη της ανάκτησης kNN [Schaaf, 1996].

iv. Δέντρα Κάλυψης (Cover Trees)

Αυτή η τεχνική μπορεί να θεωρηθεί η τελειότερη μέχρι στιγμής στην επιτάχυνση της ανάκτησης kNN. Για την οργάνωση των προτύπων εκπαίδευσης ώστε η ανάκτηση να είναι αποδοτική, χρησιμοποιείται μία δομή δεδομένων που ονομάζεται δέντρο κάλυψης. Η χρήση δέντρων κάλυψης προϋποθέτει τη χρήση μετρικού υπό την αυστηρή έννοια όμως οι απαιτήσεις σε χώρο καθώς και η βελτίωση της ταχύτητας καθιστούν την τεχνική αυτή πολύ ελκυστική. Οι απαιτήσεις σε χώρο είναι της τάξης του $O(N)$, ο χρόνος κατασκευής των δέντρων είναι της τάξης του $O(c^6 N \log N)$ ενώ ο χρόνος ανάκτησης $O(c^{12} N \log N)$, όπου c είναι ένα μέτρο της διαστατικότητας των δεδομένων. [Beygelzimer et al. 1006]

Οι παραπάνω τεχνικές περιλαμβάνουν επιπρόσθετη επεξεργασία των δεδομένων προκειμένου να κατασκευαστούν κατάλληλες δομές δεδομένων και παρουσιάζουν μεγαλύτερη δυσκολία στην υλοποίηση σε σχέση με την απλή μέθοδο kNN προσφέρουν όμως μεγάλη επιτάχυνση της ανάκτησης των κοντινότερων γειτόνων. Εναλλακτική επιλογή για την επιτάχυνση της κατηγοριοποίησης είναι η μείωση της διαστατικότητας των δεδομένων.



Σχήμα 16: Απεικόνιση του στόχου των δύο στρατηγικών επεξεργασίας και διόρθωσης του συνόλου εκπαίδευσης. Η ενίσχυση επάρκειας αφαιρεί πρότυπα με θόρυβο ή εξαιρέτικές περιπτώσεις, ενώ η

διατήρηση επάρκειας αφαιρεί πρότυπα που βρίσκονται στο εσωτερικό μίας συστάδας προτύπων της ίδιας κλάσης.

Η έρευνα για την μείωση των διαστάσεων των δεδομένων έχει δύο κατευθύνσεις. Το πρώτο πεδίο βελτιστοποιήσεων, η *επιλογή χαρακτηριστικών (feature selection)* εξηγείται σε ξεχωριστή ενότητα καθώς είναι μια τεχνική που εφαρμόζεται σε πολλούς κατηγοριοποιητές με επίβλεψη. Η άλλη πλευρά της μείωσης διαστάσεων είναι η διαγραφή περιττών προτύπων ή προτύπων με θόρυβο στο σύνολο εκπαίδευσης ώστε να βελτιώνεται η απόδοση της κατηγοριοποίησης.

Οι πρώτες τεχνικές που αναπτύχθηκαν στην κατεύθυνση της μείωσης των προτύπων στο σύνολο εκπαίδευσης χωρίζονται σε δύο υποκατηγορίες, τις τεχνικές *διατήρησης επάρκειας (competence preservation)* και τις τεχνικές *ενίσχυσης επάρκειας (competence enhancement)* (βλ. σχήμα 16). Η διατήρηση επάρκειας περιλαμβάνει τον εντοπισμό και την εξαίρεση από το σύνολο εκπαίδευσης των προτύπων που πλεονάζουν και δεν συνεισφέρουν στην επάρκεια της κατηγοριοποίησης. Αυτό επιτυγχάνεται αφαιρώντας πρότυπα που βρίσκονται εσωτερικά μίας συστάδας προτύπων της ίδιας κλάσης και όπως διαφαίνεται διατηρούν τα πρότυπα με θόρυβο ως εξαιρέσεις ή συνοριακά πρότυπα. Από την άλλη πλευρά, η ενίσχυση επάρκειας αποτελεί ουσιαστικά διαδικασία μείωσης θορύβου, αφαιρώντας πρότυπα θορύβου ή φθαρμένα πρότυπα. Υπάρχει περίπτωση όμως στα πλαίσια αυτής της προσπάθειας να αφαιρεθούν εξαιρετικές ή συνοριακές περιπτώσεις προτύπων που μπορεί να μην μπορούν να ξεχωρίσουν από τον πραγματικό θόρυβο. Συνεπώς, για την βέλτιστο αποτέλεσμα πρέπει να βρεθεί μια ισορροπία μεταξύ των δύο τεχνικών. Οι διάφορες τέτοιες στρατηγικές τροποποίησης του συνόλου εκπαίδευσης υλοποιούνται με δύο τρόπους, είτε επαυξητικά οπότε προστίθενται επιλεγμένα πρότυπα του συνόλου εκπαίδευσης σε ένα αρχικά κενό σύνολο διόρθωσης, είτε μειωτικά οπότε από το σύνολο εκπαίδευσης αφαιρούνται επιλεγμένα πρότυπα.

Μία πρώτη τεχνική διατήρησης επάρκειας είναι η επαυξητική τεχνική των λεγόμενων *Συνοπτικών Κοντινότερων Γειτόνων (Condensed Nearest Neighbor, CNN)* [Hart 1968]. Η μέθοδος CNN προβλέπει την χρήση ενός αρχικά κενού συνόλου διόρθωσης στο οποίο προστίθενται τα πρότυπα του συνόλου εκπαίδευσης που δεν είναι δυνατό να ταξινομηθούν σωστά από το σύνολο διόρθωσης. Αυτή η τεχνική είναι πολύ ευαίσθητη στο θόρυβο και μάλιστα τείνει εξ ορισμού να διατηρεί τα πρότυπα με θόρυβο. Για το λόγο αυτό προτάθηκε μία βελτιωμένη εκδοχή του CNN, η μέθοδος των *Επιλεκτικών Κοντινότερων Γειτόνων (Selective Nearest Neighbor, SNN)* [Ritter et al. 1975]. Η μέθοδος αυτή επιβάλει τον κανόνα ότι κάθε πρότυπο του συνόλου εκπαίδευσης πρέπει να βρίσκεται πιο κοντά σε ένα πρότυπο της ίδιας κλάσης στο σύνολο διόρθωσης απ' ότι σε κάποιο πρότυπο άλλης κλάσης. Μία άλλη τεχνική διατήρησης επάρκειας [Gates 1972], μειωτική όμως αυτή τη φορά, υπαγορεύει την ύπαρξη ενός συνόλου διόρθωσης αρχικά όμοιο με το σύνολο εκπαίδευσης. Στη συνέχεια, αφαιρούνται πρότυπα από το σύνολο διόρθωσης των οποίων η αφαίρεση δεν προκαλεί σφάλμα κατηγοριοποίησης των υπόλοιπων προτύπων εκπαίδευσης. Η τεχνική αυτή επιτυγχάνει σε ένα ποσοστό την αφαίρεση θορύβου, είναι όμως ευαίσθητη στη σειρά με την οποία παρουσιάζονται τα πρότυπα.

Όσον αφορά στις τεχνικές ενίσχυσης επάρκειας ή αλλιώς μείωσης θορύβου, η πρώτη που προτάθηκε ήταν ο αλγόριθμος *Επεξεργασμένων Κοντινότερων Γειτόνων (Edited*

Nearest Neighbor, ENN) [Wilson 1972]. Πρόκειται για μειωτική στρατηγική σύμφωνα με την οποία αφαιρούνται από το σύνολο εκπαίδευσης όσα πρότυπα δεν συμφωνούν με τους k κοντινότερους γείτονες τους καθώς θεωρούνται ως θόρυβος και δεν θα έπρεπε να υπάρχουν σε μία ομάδα προτύπων από την ίδια κλάση. Αυτός ο αλγόριθμος επεκτάθηκε στον *Επαναληπτικό ENN (RENN)* [Tomek 1976], ο οποίος διατρέχει κατ' επανάληψη το σύνολο εκπαίδευσης με τη λογική του ENN μέχρις ότου να μην μπορεί να ελαχιστοποιηθεί περεταίρω το σύνολο εκπαίδευσης. Αυτές οι στρατηγικές εστιάζουν στην εξάλειψη των προτύπων με θόρυβο ή άλλων εξαιρετικών περιπτώσεων, δεν έχουν όμως τόσο καλές επιδόσεις στην μείωση των απαιτήσεων μνήμης όσο οι τεχνικές διατήρησης επάρκειας.

Οι μεταγενέστερες τεχνικές επεξεργασίας του συνόλου εκπαίδευσης μπορούν να χαρακτηριστούν υβρίδια των δύο ειδών που αναφέρθηκαν παραπάνω καθώς ενσωματώνουν διαδικασίες τόσο διατήρησης επάρκειας όσο και ενίσχυσης επάρκειας. Ένα παράδειγμα είναι η σειρά αλγορίθμων *IBn (Instance Based n)* [Aha et al. 1991] οι οποίοι μειώνουν τις απαιτήσεις μνήμης της κατηγοριοποίησης ενώ παράλληλα αντιμετωπίζουν το θόρυβο. Ο αλγόριθμος IB2 είναι όμοιος με τον CNN με την προσθήκη ότι στο σύνολο διόρθωσης προστίθενται μόνο πρότυπα που δεν μπορούν να ταξινομηθούν σωστά από το μειωμένο σύνολο εκπαίδευσης. Η δυσκολία εξάλειψης του θορύβου του αλγορίθμου αυτού αντιμετωπίζεται στον αλγόριθμο IB3 ο οποίος καταγράφει πόσο σωστά ταξινομούνται τα πρότυπα που αφαιρούνται από το εναπομείναν σύνολο εκπαίδευσης και επιλέγει αυτά που ταξινομούνται σωστά σε σχέση με κάποιο στατιστικό μέτρο.

Οι πιο σύγχρονες προσεγγίσεις στην επεξεργασία του συνόλου εκπαίδευσης χρησιμοποιούν ένα μοντέλο επάρκειας των δεδομένων εκπαίδευσης και εισάγουν κάποιες ιδιότητες επάρκειας που καθορίζουν ποια πρότυπα θα συμπεριληφθούν τελικά στο σύνολο εκπαίδευσης. Οι πρώτες σημαντικές ιδιότητες που εισήχθησαν είναι το *σύνολο προσβασιμότητας (reachability set)* και το *σύνολο κάλυψης (coverage set)*. Το σύνολο προσβασιμότητας ενός προτύπου είναι το σύνολο των προτύπων τα οποία μπορούν να ταξινομήσουν σωστά το πρότυπο αυτό. Το σύνολο κάλυψης ενός προτύπου είναι το σύνολο των προτύπων που το πρότυπο αυτό μπορεί να ταξινομήσει επιτυχώς. Τα σύνολα προσβασιμότητας και κάλυψης αποτελούν τα χαρακτηριστικά τοπικής επάρκειας ενός προτύπου και αποτελούν τη βάση για έναν αριθμό τεχνικών επεξεργασίας του συνόλου εκπαίδευσης.

Οι αλγόριθμοι που αναπτύχθηκαν σε αυτή τη βάση έχουν τέσσερις βασικές λειτουργίες. Ανάλογα με τις επιλογές που γίνονται για την υλοποίηση αυτών των λειτουργιών σχηματίστηκαν διάφορες οικογένειες αλγορίθμων. Οι επιλογές αφορούν:

(i) μία *πολιτική διάταξης* για την παρουσίαση των προτύπων με βάση τα χαρακτηριστικά επάρκειας των προτύπων
(ii) ένας *κανόνας πρόσθεσης* για τον καθορισμό των προτύπων που θα συμπεριληφθούν στο διορθωμένο σύνολο.
(iii) ένας *κανόνας διαγραφής* για τον καθορισμό των προτύπων που θα αφαιρεθούν από το σύνολο εκπαίδευσης.
(iv) μία *πολιτική ενημέρωσης* ώστε να ελέγχεται αν το μοντέλο επάρκειας ενημερώνεται μετά από κάθε βήμα της επεξεργασίας.

Έτσι, αναπτύχθηκε ο αλγόριθμος *Επαναληπτικού Φιλτραρίσματος (Iterative Case Filtering, ICF)* [Brighton, Mellish 2002], ο οποίος ακολουθεί μειωτική στρατηγική και

αφαιρεί από το σύνολο εκπαίδευσης τα πρότυπα των οποίων το πλήθος των στοιχείων του σύνολου προσβασιμότητας είναι μεγαλύτερο από αυτό του συνόλου κάλυψης. Με αυτή τη μέθοδο αφαιρούνται πρότυπα που βρίσκονται μακριά από τα σύνορα μίας κλάσης. Μετά το πέρας κάθε επανάληψης το μοντέλο επάρκειας ενημερώνεται και η διαδικασία τερματίζεται όταν δεν μπορούν να αφαιρεθούν πλέον άλλα πρότυπα. Προβλέπεται επίσης και ένα στάδιο προεπεξεργασίας για τη μείωση του θορύβου, πρακτικά γίνεται χρήση του αλγορίθμου RENN.

Μία άλλη σειρά αλγορίθμων που προτάθηκε είναι οι αλγόριθμοι *Τεχνικής Μείωσης (Reduction Technique, RTn)* [Wilson, Martinez 1997]. Οι αλγόριθμοι αυτοί ακολουθούν την μειωτική στρατηγική και αφαιρούν κάποιο πρότυπο από το σύνολο εκπαίδευσης αν τουλάχιστον τόσα πρότυπα όσα υπάρχουν στο σύνολο κάλυψής του θα μπορούσαν να ταξινομηθούν σωστά αν δεν υπήρχε το πρότυπο αυτό. Πρακτικά, αφαιρεί πρότυπα με θόρυβο και πρότυπα που βρίσκονται στο κέντρο μίας συστάδας προτύπων της ίδιας κλάσης καθώς τα υπόλοιπα μπορούν να ταξινομηθούν και χωρίς αυτά.

Συμπερασματικά λοιπόν, η μέθοδος των k Κοντινότερων Γειτόνων είναι μία μέθοδος που είναι απλή, κατανοητή και εύκολη στην υλοποίηση, της οποίας μάλιστα οι επιδόσεις είναι δυνατόν να βελτιωθούν με χρήση τεχνικών για την αφαίρεση θορύβου αλλά και τη μείωση των απαιτήσεων μνήμης. Όμως, χρειάζεται να δοθεί και επιπλέον σπουδή προκειμένου να βελτιωθεί η χρονική πολυπλοκότητα, καθώς από τη φύση του αλγορίθμου kNN όλος ο φόρτος εργασίας αντιμετωπίζεται κατά την εκτέλεση. Επίσης, η μέθοδος είναι ευαίσθητη σε άσχετα ή περιττά χαρακτηριστικά καθώς όλα τα χαρακτηριστικά συμμετέχουν στο μέτρο της απόστασης, οπότε και απαιτούνται μέθοδοι για την επιλογή των κατάλληλων χαρακτηριστικών ή την στάθμιση των χαρακτηριστικών. Γενικά, σε δύσκολες εργασίες κατηγοριοποίησης η μέθοδος kNN υποσκελίζεται από άλλες τεχνικές όπως οι Μηχανές Διανυσμάτων Υποστήριξης ή τα Νευρωνικά Δίκτυα.

2.2. Μη Επιβλεπόμενη Μάθηση

Στην περίπτωση της μη επιβλεπόμενης μάθησης ή αλλιώς αυτο-οργανούμενης μάθησης δεν υπάρχει κάποιος εξωτερικός παράγοντας που να επεμβαίνει στη διαδικασία μάθησης. Το κενό αυτό καλύπτεται ορίζοντας κάποιο μέτρο ποιότητας της αναπαράστασης που καλείται να αναγνωρίσει ο κατηγοριοποιητής, ανεξάρτητο της εργασίας. Οι ελεύθερες παράμετροι του συστήματος θα πρέπει με την πάροδο της διαδικασίας μάθησης να βελτιστοποιούνται ως προς το μέτρο αυτό. Αν επιλεγεί ένα τέτοιο ανεξάρτητο από την εργασία μέτρο, τότε μετά από την τροφοδότηση του κατηγοριοποιητή με διανύσματα χαρακτηριστικών ώστε να παρουσιάζουν κάποια στατιστική κανονικότητα, τότε ο κατηγοριοποιητής αποκτά την ικανότητα να σχηματίζει εσωτερικές αναπαραστάσεις για την κωδικοποίηση χαρακτηριστικών των διανυσμάτων εισόδου και να δημιουργεί αυτόματα νέες κλάσεις/*συστάδες (clusters).*

Με άλλα λόγια, η μη επιβλεπόμενη μάθηση στηρίζεται σε αλγορίθμους που προσπαθούν να ανιχνεύσουν ομοιότητες μεταξύ των διανυσμάτων χαρακτηριστικών που δέχονται στην είσοδό τους και να κατατάξουν τα πρότυπα που ομοιάζουν στην ίδια ομάδα, ενώ ο αριθμός των ομάδων μπορεί να αυξηθεί δυναμικά ανάλογα με τα δεδομένα εισόδου. Αυτό επιτυγχάνεται μέσω της ανακάλυψης σημαντικών προτύπων ή χαρακτηριστικών

μεταξύ των δεδομένων εισόδου με χρήση των λεγόμενων *μη χαρακτηρισμένων παραδειγμάτων (unlabeled examples)*. Η μη επιβλεπόμενη μάθηση μπορεί να ακολουθήσει δύο στρατηγικές, την *αυτο-οργανούμενη μάθηση* και την *στατιστική θεωρία μάθησης*. Η υλοποίηση της αυτο-οργανούμενης μάθησης απορρέει από τη συμπεριφορά των βιολογικών νευρωνικών δικτύων, καθώς υιοθετείται η τοπικότητα των λειτουργιών που παρουσιάζεται στον εγκέφαλο. Από την άλλη πλευρά η στατιστική θεωρία μάθησης στηρίζεται περισσότερο στην χρήση καλά τεκμηριωμένων μαθηματικών εργαλείων.

Μερικοί κατηγοριοποιητές χωρίς επίβλεψη στους οποίους γίνει αναλυτική αναφορά στην εργασία αυτή είναι οι Χάρτες Αυτο-Οργάνωσης (Self-Organising Maps, SOM) και ο Ασαφής C-Μέσος (Fuzzy C-Means, FCM).

2.2.1. Χάρτες Αυτο-Οργάνωσης (Self-Organising Maps, SOM)

Οι χάρτες αυτο-οργάνωσης είναι μία ειδική κατηγορία τεχνητών νευρωνικών δικτύων και όπως είναι εμφανές, στηρίζονται στις αρχές της αυτο-οργανούμενης μάθησης. Στηρίζονται στην *ανταγωνιστική μάθηση*, υπό την έννοια ότι ένας μόνο νευρώνας (ή ένας νευρώνας ανά ομάδα) έχει δικαίωμα ενεργοποίησης στην έξοδο του δικτύου οπότε οι νευρώνες ανταγωνίζονται για το ποιος θα υπερισχύσει.

Αρχές Αυτο-οργάνωσης

Η αυτο-οργανούμενη μάθηση είναι μία προσέγγιση της μη επιβλεπόμενης μάθησης η οποία στηρίζεται στη λειτουργία των νευρώνων του εγκεφάλου. Συγκεκριμένα, ένας αλγόριθμος που ακολουθεί τις αρχές της αυτο-οργάνωσης εφοδιάζεται με ένα σύνολο κανόνων τοπικής συμπεριφοράς και ο στόχος είναι να χρησιμοποιηθούν οι κανόνες για να υπολογιστεί μια αντιστοίχιση εισόδου-εξόδου με τις επιθυμητές ιδιότητες. Η τοπικότητα έχει την έννοια ότι οι προσαρμογές που εφαρμόζονται στα συναπτικά βάρη κάθε νευρώνα του δικτύου περιορίζονται στην άμεση τοπική γειτονιά του νευρώνα. Έτσι, συγκροτούνται οι τέσσερις αρχές που ακολουθούν οι αλγόριθμοι αυτο-οργάνωσης.

1. Αυτο-ενίσχυση

Η πρώτη αρχή της αυτο-οργάνωσης βασίζεται στο επιχείρημα:

> *Οι τροποποιήσεις στα συναπτικά βάρη ενός νευρώνα τείνουν να αυτο-ενισχύονται σύμφωνα με το αίτημα μάθησης του Hebb, το οποίο εδράζεται στην πλαστικότητα των συνάψεων.*

Η διαδικασία της *αυτο-ενίσχυσης (self-amplification)*, δηλαδή, περιορίζεται από την απαίτηση οι τροποποιήσεις που γίνονται στα συναπτικά βάρη του νευρώνα να βασίζονται στα προσυναπτικά και μετασυναπτικά σήματα που είναι διαθέσιμα στο τοπικό επίπεδο. Με άλλα λόγια προβλέπεται ένας μηχανισμός ανάδρασης μέσω του οποίου μια ισχυρή σύναψη οδηγεί στη σύμπτωση των προσυναπτικών και μετασυναπτικών σημάτων και έτσι η σύναψη αυξάνει την ισχύ της λόγω αυτής της σύμπτωσης.

Το αίτημα μάθησης που έχει διατυπωθεί από τον Hebb το 1949 αφορούσε τους βιολογικούς νευρώνες. Το τροποποιημένο *αίτημα μάθησης του Hebb* για τα τεχνητά νευρωνικά δίκτυα διατυπώνεται ως ένας διμερής κανόνας:

i. *Εάν δύο νευρώνες στις δύο πλευρές μιας σύναψης ενεργοποιούνται ταυτόχρονα τότε η ισχύς αυτής της σύναψης αυξάνεται επιλεκτικά.*

ii. *Εάν δύο νευρώνες στις δύο πλευρές μιας σύναψης ενεργοποιούνται ασύγχρονα, τότε η σύναψη αποδυναμώνεται επιλεκτικά ή εξαφανίζεται.*

Μία σύναψη που υπακούει στον παραπάνω κανόνα ονομάζεται «Χεμπιανή». Πιο συγκεκριμένα μία σύναψη ορίζεται ως Χεμπιανή όταν χρησιμοποιεί έναν εξαρτώμενο από το χρόνο, εξαιρετικά τοπικό και έντονα αλληλεπιδραστικό μηχανισμό για την αύξηση της συναπτικής αποτελεσματικότητας σαν συνάρτηση της συσχέτισης μεταξύ της προσυναπτικής και μετασυναπτικής δραστηριότητας.

2. Ανταγωνισμός

Η δεύτερη αρχή της αυτο-οργάνωσης υπαγορεύει το εξής:

> *Η ύπαρξη περιορισμών στους διαθέσιμους πόρους, με οποιαδήποτε μορφή, οδηγεί στον ανταγωνσμό μεταξύ των συνάψεων ενός μεμονωμένου νευρώνα ή μεταξύ των νευρώνων μιας ομάδας νευρώνων, με αποτέλεσμα οι πιο έντονα αναπτυσσόμενες συνάψεις, ή νευρώνες, αντίστοιχα, να επιλέγονται εις βάρος των άλλων.*

Η ανταγωνιστική αυτή διαδικασία σε ένα δίκτυο εξελίσσεται ως εξής. Αρχικά, οι νευρώνες του δικτύου είναι όλοι στην ίδια κατάσταση εκτός από κάποια τυχαία κατανεμημένα συναπτικά βάρη, λόγω των οποίων οι νευρώνες αντιδρούν διαφορετικά σε ένα δεδομένο σύνολο προτύπων εισόδου. Επιβάλλεται, στη συνέχεια, ένα συγκεκριμένο όριο στην ισχύ κάθε νευρώνα του δικτύου. Οι νευρώνες ανταγωνίζονται με αυτό τον τρόπο σύμφωνα με ένα προκαθορισμένο κανόνα για το δικαίωμα να αποκριθούν σ' ένα δεδομένο υποσύνολο εισόδων. Έτσι, μόνο ένας νευρώνας εξόδου ή ένας νευρώνας ανά ομάδα, είναι ενεργός ανά πάσα στιγμή, ο οποίος επικρατεί. Οι μεμονωμένοι νευρώνες του δικτύου αναλαμβάνουν μέσω της *διαδικασίας ανταγωνιστικής μάθησης* το ρόλο των *ανιχνευτών χαρακτηριστικών (feature detectors)* για διαφορετικές κλάσεις προτύπων εισόδου.

3. Συνεργασία

Η τρίτη αρχή της αυτο-οργάνωσης συνίσταται στην επόμενη πρόταση

> *Οι τροποποιήσεις των συναπτικών βαρών σε επίπεδο μεμονωμένου νευρώνα και σε επίπεδο δικτύου τείνουν να συνεργάζονται μεταξύ τους.*

Ένας νευρώνας δεν είναι συνήθως δυνατόν να ενεργοποιηθεί από σήμα μίας μόνο σύναψης. Πρέπει επομένως να υπάρξει μία συνεργασία μεταξύ των συνάψεων του νευρώνα ώστε να είναι δυνατόν να παραχθούν σήματα επαρκούς ισχύος για την ενεργοποίηση του νευρώνα αυτού. Σε επίπεδο δικτύου η συνεργασία μπορεί να πάρει τη μορφή πλευρικής αλληλεπίδρασης μεταξύ μιας ομάδας διεγερμένων νευρώνων. Ένας νευρώνας που ενεργοποιείται τείνει να διεγείρει και νευρώνες που βρίσκονται στην άμεση γειτονιά του.

4. Δομική Πληροφορία

Η τέταρτη και τελευταία αρχή της αυτο-οργάνωσης αναφέρει:

> *Η υποκείμενη τάξη και η δομή που ενυπάρχουν σ' ένα σήμα εισόδου αντιπροσωπεύουν πλεονασματική πληροφορία, η οποία αποκτάται από ένα αυτο-οργανούμενο σύστημα με τη μορφή γνώσης.*

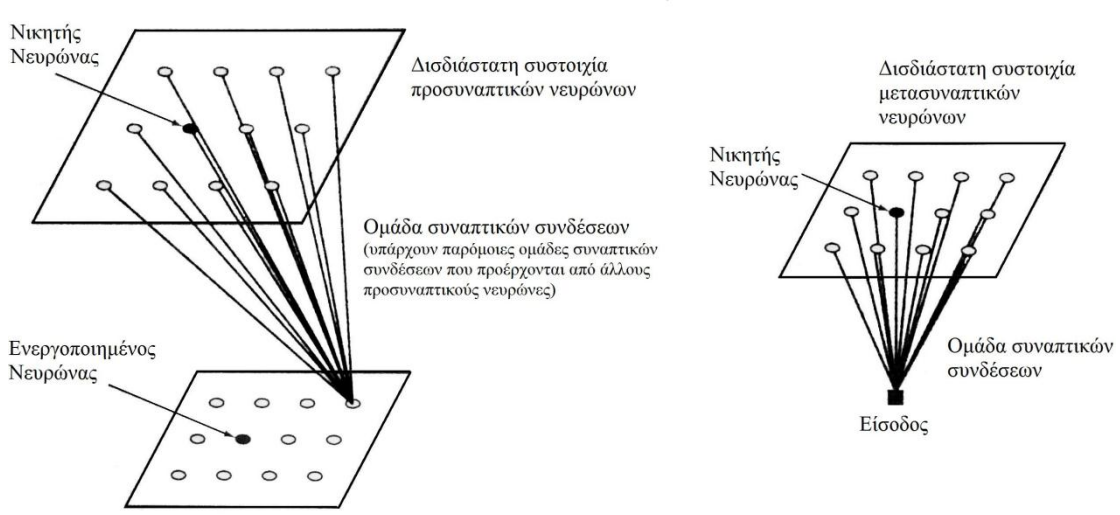
Η δομική πληροφορία που περιέχεται στα δεδομένα εισόδου είναι συνεπώς αναγκαία προϋπόθεση για την αυτο-οργανούμενη μάθηση. Η αυτο-ενίσχυση, ο ανταγωνισμός και η συνεργασία που αναφέρθηκαν προηγουμένως αφορούν τους νευρώνες ενώ η δομική πληροφορία είναι εγγενές χαρακτηριστικό του σήματος εισόδου. Αν αφαιρούνταν η δομική πληροφορία από ένα σήμα τότε το σήμα δεν θα ήταν διαχωρίσιμο από το θόρυβο και δεν θα είχε αξία σε οποιοδήποτε σύστημα μάθησης.

Χαρτογράφηση Χαρακτηριστικών

Στον εγκέφαλο διαφορετικές αισθητηριακές είσοδοι από τα διάφορα υποσυστήματα χαρτογραφούνται σε αντίστοιχες περιοχές του εγκεφαλικού φλοιού με διατεταγμένο τρόπο. Μια τέτοια δομή τοπογραφικού χάρτη που έχει την ικανότητα να μαθαίνει με τρόπο εμπνευσμένο από τη λειτουργία του εγκεφάλου είναι και το ζητούμενο στην κατασκευή των αυτο-οργανούμενων χαρτών. Ένας τέτοιος χάρτης πρέπει να υπακούει στην *αρχή σχηματισμού ενός τοπογραφικού χάρτη* [Kohonen, 1990]:

> *Η χωρική θέση ενός νευρώνα εξόδου σ' ένα τοπογραφικό χάρτη αντιστοιχεί σ' ένα συγκεκριμένο πεδίο ή χαρακτηριστικό των δεδομένων που αντλούνται από το χώρο εισόδου.*

Σύμφωνα με την αρχή αυτή αναπτύχθηκαν δύο διαφορετικά μοντέλα αντιστοίχισης χαρακτηριστικών (σχήμα 17).



(α) Μοντέλο Willshaw-von der Malsburg **(β) Μοντέλο Kohonen**

Σχήμα 17: Δύο μοντέλα αυτο-οργανούμενων χαρτών χαρακτηριστικών
(α) Το μοντέλο των Willshaw-von der Malsburg
(β) Το μοντέλο του Kohonen

Παρατηρώντας το σχήμα 17 φαίνεται πως και στις δύο περιπτώσεις οι νευρώνες εξόδου είναι διατεταγμένοι σε δισδιάστατο πλέγμα. Αυτό το είδος τοπολογίας εξασφαλίζει ότι κάθε νευρώνας έχει ένα σύνολο από γείτονες. Η διαφορά των δύο μοντέλων έγκειται στον τρόπο με τον οποίο καθορίζονται τα πρότυπα εισόδου.

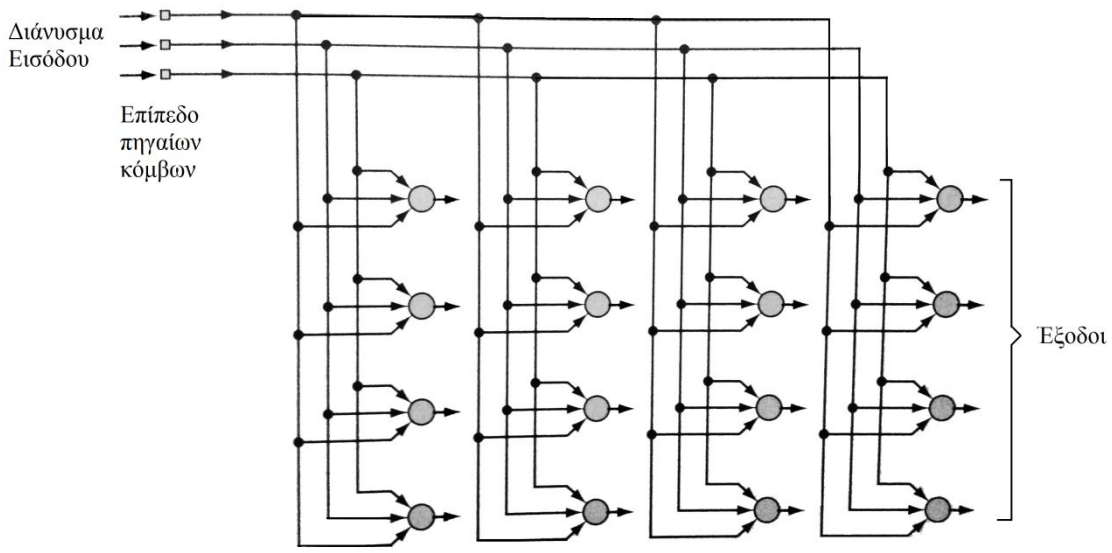
Το πρώτο μοντέλο που απεικονίζεται στην εικόνα 1α προτάθηκε από τους *Willshaw και von der Malsburg* (1976). Το μοντέλο αυτό προβλέπει την ύπαρξη δύο ξεχωριστά δισδιάστατα πλέγματα νευρώνων συνδεδεμένα μαζί, με το ένα να προβάλλεται στο άλλο. Το ένα πλέγμα αντιπροσωπεύει προσυναπτικούς νευρώνες (εισόδου) και το άλλο πλέγμα αντιπροσωπεύει μετασυναπτικούς νευρώνες (εξόδου). Το πλέγμα των μετασυναπτικών νευρώνων χρησιμοποιεί ένα διεγερτικό μηχανισμό μικρής εμβέλειας και έναν ανασταλτικό μηχανισμό μεγάλης εμβέλειας. Αυτοί οι δύο μηχανισμοί είναι τοπικής φύσεως και είναι μείζονος σημασίας για την αυτο-οργάνωση. Τα δύο πλέγματα συνδέονται μεταξύ τους με Χεμπιανές συνάψεις, των οποίων τα βάρη είναι τροποποιήσιμα. Έτσι, μόνο λίγοι μετασυναπτικοί νευρώνες ενεργοποιούνται κάθε στιγμή ενώ για να αποφευχθεί αστάθεια το συνολικό βάρος που αντιστοιχεί σε κάθε νευρώνα περιορίζεται από κάποιο άνω όριο. Με αυτόν τον τρόπο κάποια συναπτικά βάρη αυξάνονται και κάποια μειώνονται. Το μοντέλο αυτό ειδικεύεται σε προβλήματα στα οποία η διάσταση της εισόδου είναι ίδια με τη διάσταση της εξόδου.

Το δεύτερο μοντέλο που φαίνεται στο σχήμα 1β παρουσιάστηκε από τον Kohonen (1982) δεν σχετίζεται άμεσα με τις νευροβιολογικές λειτουργίες. Είναι πιο γενικό από το μοντέλο Willshaw-von der Malsburg αφού έχει τη δυνατότητα να εκτελεί συμπίεση δεδομένων, μειώνει δηλαδή τη διαστατικότητα της εισόδου. Το μοντέλο Kohonen παρέχει μία τοπολογική αντιστοίχιση η οποία τοποθετεί με βέλτιστο τρόπο ένα σταθερό αριθμό διανυσμάτων (κωδικών) σε έναν υψηλότερης διασταστικότητας χώρο εισόδου διευκολύνοντας τη συμπίεση δεδομένων, διαδικασία που αναφέρεται ως *διανυσματική κωδικοποίηση*. Επομένως, το μοντέλο αυτό μπορεί να προσεγγιστεί με δύο τρόπους, πρώτον, με την παραδοσιακή προσέγγιση η οποία βασίζεται στις αρχές της αυτο-οργάνωσης των βιολογικών νευρώνων και, δεύτερον, με μια εναλλακτική προσέγγιση, αυτή της διανυσματικής κβάντισης η οποία βασίζεται στους μηχανισμούς κωδικοποίησης και αποκωδικοποίησης όπως αυτοί ορίζονται στη θεωρία επικοινωνίας. Αυτό είναι και το μοντέλο που θα χρησιμοποιηθεί στην ανάλυση των αυτο-οργανούμενων χαρτών.

Χάρτες Αυτο-οργάνωσης

Ο στόχος ενός χάρτη αυτο-οργάνωσης (SOM) είναι να μετασχηματίζει ένα πρότυπο εισερχόμενου σήματος, τυχαίας διάστασης, σε ένα διακριτό χάρτη μίας ή δύο διαστάσεων και να εκτελεί αυτό το μετασχηματισμό προσαρμστικά με κάποιον τοπολογικά διατεταγμένο τρόπο. Στο σχήμα 2 απεικονίζεται ένα δισδιάστατο πλέγμα νευρώνων που χρησιμοποιείται συχνά ως διακριτός χάρτης. Ένας SOM μπορεί να θεωρηθεί ως ένα δίχτυ το οποίο απλώνεται στο χώρο των δεδομένων. Ο αλγόριθμος SOM μετακινεί τα διανύσματα βαρών ώστε να καλύπτει όλο το χώρο δεδομένων και ο χάρτης να είναι οργανωμένος, υπό την έννοια ότι γειτονικοί νευρώνες στο πλέγμα καταλήγουν να έχουν παραπλήσια διανύσματα βαρών.

Ο αλγόριθμος που είναι υπεύθυνος για το σχηματισμό του αυτο-οργανούμενου χάρτη ξεκινά αρχικοποιώντας τα συναπτικά βάρη στο δίκτυο. Τα συναπτικά βάρη αυτά αρχικοποιούνται με μικρές τιμές με χρήση γεννήτριας τυχαίων αριθμών ώστε να μην επιβληθεί κάποια αρχική προτεραιότητα σε κάποιους νευρώνες του χάρτη χαρακτηριστικών. Από το σημείο αυτό και έπειτα, λαμβάνουν χώρα τρεις βασικές διαδικασίες οι οποίες οδηγούν στο σχηματισμό του χάρτη:



Σχήμα 18: Δισδιάστατο πλέγμα νευρώνων με είσοδο τριών διαστάσεων και έξοδο 4x4 (οι έξοδοι συμβολίζονται με βέλη). Κάθε νευρώνας είναι πλήρως συνδεδεμένος με όλους τους πηγαίους κόμβους του επιπέδου εισόδου.

1. *Ανταγωνισμός*

Για κάθε πρότυπο εισόδου οι νευρώνες του δικτύου υπολογίζουν τις αντίστοιχες τιμές μιας συνάρτησης διάκρισης. Η συνάρτηση αυτή καθορίζει τους όρους με βάση τους οποίους διεξάγεται ο ανταγωνισμός μεταξύ των νευρώνων. Ο νευρώνας με τη μεγαλύτερη τιμή στη συνάρτηση διάκρισης είναι και αυτός που αναδεικνύεται νικητής του ανταγωνισμού.

Έστω m η διάσταση του χώρου εισόδου. Ένα πρότυπο εισόδου επομένως είναι το παρακάτω

$$\mathbf{x} = [x_1,\ x_2, \dots, x_m]^T \tag{86}$$

ενώ το διάνυσμα συναπτικών βαρών του νευρώνα j

$$\mathbf{w}_j = \left[w_{j1},\ w_{j2}, \dots, w_{jm}\right]^T, \qquad j = 1,2, \dots, l \tag{87}$$

όπου l ο συνολικός αριθμός νευρώνων του δικτύου. Η βέλτιστη ταύτιση του διανύσματος εισόδου $\mathbf{x}$ με τα διανύσματα συναπτικών βαρών προκύπτει επιλέγοντας το μεγαλύτερο από τα εσωτερικά γινόμενα $\mathbf{w}_j{}^T\mathbf{x}$ για $j = 1,2, \dots, l$. Η επιλογή αυτή καθορίζει ουσιαστικά σε ποιο σημείο τοποθετείται το κέντρο της τοπολογικής γειτονιάς των διεγερμένων νευρώνων. Η μεγιστοποίηση του εσωτερικού γινομένου $\mathbf{w}_j{}^T\mathbf{x}$ ανάγεται μαθηματικά στην ελαχιστοποίηση της Ευκλείδειας απόστασης μεταξύ των διανυσμάτων $\mathbf{x}$ και $\mathbf{w}_j$, δεδομένου ότι το $\mathbf{w}_j$ έχει μοναδιαίο μήκος για όλα τα j. Σύμφωνα με αυτή την προσέγγιση η διαδικασία ανταγωνισμού μετξύ των νευρώνων περιγράφεται από τη σχέση

$$i(\mathbf{x}) = \arg\min_j \left\|\mathbf{x} - \mathbf{w}_j\right\|, \qquad j \in \mathcal{A} \tag{88}$$

όπου $i(\mathbf{x})$ είναι ο δείκτης του νευρώνα που ταιριάζει καλύτερα με το διάνυσμα εισόδου $\mathbf{x}$ και $\mathcal{A}$ το πλέγμα των νευρώνων. Ο νευρώνας i που ικανοποιεί αυτή τη συνθήκη ονομάζεται *νευρώνας βέλτιστης ταύτισης* ή *νικητής νευρώνας* για το διάνυσμα εισόδου **x.** Έτσι, προκύπτει ότι ένας συνεχής χώρος εισόδου από πρότυπα ενεργοποίησης αντιστοιχίζεται σ'

ένα διακριτό χώρο εξόδου νευρώνων μέσω μιας διαδικασίας ανταγωνισμού μεταξύ των νευρώνων του δικτύου.

2. Συνεργασία

Ο νικητής νευρώνας i καθορίζει τη χωρική θέση μιας τοπολογικής γειτονιάς διεγερμένων νευρώνων θέτοντας έτσι τη βάση για τη συνεργασία τέτοιων γειτονικών νευρώνων. Η τοπολογική γειτονιά γύρω από τον νευρώνα i φθίνει ομαλά με την πλευρική απόσταση και συμβολίζεται με το γράμμα $h_{j,i}$ όπου j ένα τυπικό δείγμα του συνόλου των διεγερμένων συνεργαζόμενων νευρώνων γύρω από τον νικητή νευρώνα. Αν $d_{j,i}$ η πλευρική απόσταση μεταξύ του νικητή νευρώνα i και του διεγερμένου νευρώνα j, η τοπολογική γειτονιά $h_{j,i}$ θεωρείται ως μια μονοκόρυφη συνάρτηση της πλευρικής απόστασης $d_{j,i}$ τέτοια ώστε να ικανοποιούνται δύο απαιτήσεις:

i. Η τοπολογική γειτονιά $h_{j,i}$ είναι συμμετρική γύρω από το μέγιστο σημείο όπου ισχύει $d_{j,i} = 0$
ii. Το πλάτος της τοπολογικής γειτονιάς $h_{j,i}$ μειώνεται μονοτονικά με την αύξηση της πλευρικής απόστασης $d_{j,i}$ τείνοντας προς το 0 για $d_{j,i} \to \infty$, η οποία είναι και αναγκαία συνθήκη για τη σύγκλιση.

Μια συνηθισμένη επιλογή του $h_{j,i}$ που ικανοποιεί αυτές τις απαιτήσεις είναι η γκαουσιανή συνάρτηση

$$h_{j,i(\mathbf{x})} = \exp\left(-\frac{d_{j,i}^2}{2\sigma^2}\right), \qquad j \in \mathcal{A} \tag{89}$$

η οποία είναι αναλλοίωτη της μετατόπισης, δηλαδή ανεξάρτητη από τη θέση του νικητή νευρώνα. Η παράμετρος σ είναι το εύρος της τοπολογικής γειτονιάς και καθορίζει το βαθμό στον οποίο οι διεγερμένοι νευρώνες στην κοντινή περιοχή του νικητή νευρώνα συμμετέχουν στη διαδικασία μάθησης. Ο τρόπος υπολογισμού της απόστασης $d_{j,i}$ εξαρτάται από τη διαστατικότητα του πλέγματος.

Το σημαντικότερο χαρακτηριστικό του αλγορίθμου SOM, όμως, είναι ότι επιτρέπει το μέγεθος της τοπολογικής γειτονιάς να μειώνεται σε κάθε επανάληψη, το εύρος σ της σχέσης (4), δηλαδή, μειώνεται με το χρόνο. Μια συνήθης επιλογή για την εξάρτηση του σ από τον διακριτό χρόνο είναι η εκθετική μείωση σύμφωνα με την εξίσωση

$$\sigma(n) = \sigma_0 \exp\left(-\frac{n}{\tau_1}\right), \qquad n = 1, 2, \ldots \tag{90}$$

όπου σ_0 είναι η αρχική τιμή του σ κατά την έναρξη του αλγορίθμου SOM και τ_1 μια σταθερά χρόνου η οποία επιλέγεται από τον σχεδιαστή. Η σχέση (89) για την τοπολογική γειτονιά επομένως γίνεται

$$h_{j,i(\mathbf{x})}(n) = \exp\left(-\frac{d_{j,i}^2}{2\sigma^2(n)}\right), \qquad n = 1, 2, \ldots, n \tag{91}$$

Καθώς αυξάνεται ο διακριτός χρόνος, δηλαδή εκτελούνται επαναλήψεις του αλγορίθμου SOM το εύρος $\sigma(n)$ μειώνεται εκθετικά και η τοπολογική γειτονιά συρρικνώνεται ανάλογα. Η συνάρτηση τοπολογικής γειτονιάς έχει και πάλι τιμή 1 για το νικητή νευρώνα i.

3. Προσαρμογή Συναπτικών Βαρών

Η τελευταία φάση του αυτο-οργανούμενου σχηματισμού ενός χάρτη χαρακτηριστικών είναι η προσαρμοστική διαδικασία των συναπτικών βαρών, δηλαδή το διάνυσμα συναπτικών

βαρών $\mathbf{w}_j$ του νευρώνα j να προσαρμόζεται σε σχέση με το διάνυσμα εισόδου $\mathbf{x}$. Λόγω του γεγονότος ότι οι αλλαγές που γίνονται στα συναπτικά βάρη είναι προς μια κατεύθυνση και όχι προσυναπτικά και μετασυναπτικά όπως στην κλασική χεμπιανή θεώρηση, τα συναπτικά βάρη μπορεί να οδηγηθούν σε κορεσμό. Για το λόγο αυτό εισάγεται ένας *όρος λησμόνησης* $g(y_j)\mathbf{w}_j$ όπου $g(y_j)$ είναι κάποια θετική βαθμωτή συνάρτηση της απόκρισης y_j. Το ανάπτυγμα Taylor της συνάρτησης αυτή πρέπει να έχει μηδενικό σταθερό όρο, δηλαδή

$$g(y_j) = 0, \quad \text{για } y_j = 0 \tag{92}$$

Έτσι, η διόρθωση στο διάνυσμα βαρών του νευρώνα j του πλέγματος εκφράζεται:

$$\Delta\mathbf{w}_j = \eta y_j \mathbf{x} - g(y_j)\mathbf{w}_j \tag{93}$$

όπου η η παράμετρος ρυθμού μάθησης του αλγορίθμου. Ο πρώτος όρος είναι ο όρος που αντιστοιχεί στην χεμπιανή θεώρηση ενώ ο δεύτερος είναι ο όρος λησμόνησης. Για να ικανοποιείται η συνθήκη (92) επιλέγεται μια γραμμική συνάρτηση $g(y_j)$,

$$g(y_j) = \eta y_j \tag{94}$$

Για ένα νικητή νευρώνα $i(\mathbf{x})$ η απόκριση είναι

$$y_j = h_{j,i(\mathbf{x})} \tag{95}$$

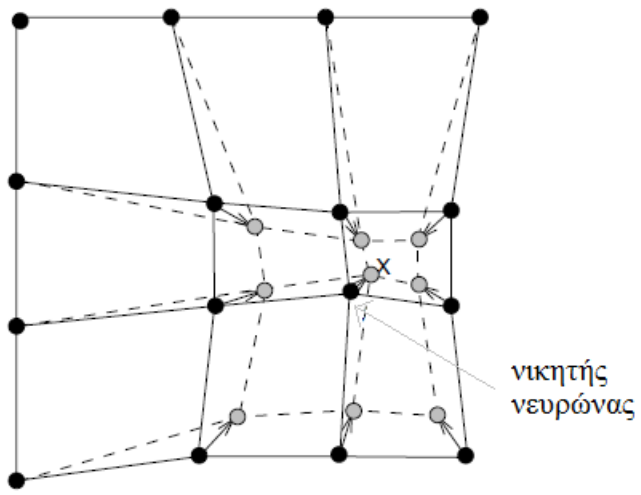
Άρα, η διόρθωση που υφίσταται το διάνυσμα βαρών του νευρώνα j από τις σχέσεις (93), (94) και (95) προκύπτει

$$\Delta\mathbf{w}_j = \eta h_{j,i(\mathbf{x})}(\mathbf{x} - \mathbf{w}_j) \tag{96}$$

όπου i είναι ο νικητής νευρώνας και j ο διεγερμένος νευρώνας. Έτσι, το ενημερωμένο διάνυσμα βαρών τη χρονική στιγμή $n + 1$ θα είναι

$$\mathbf{w}_j(n+1) = \mathbf{w}_j(n) + \eta(n)h_{j,i(\mathbf{x})}(n)\left(\mathbf{x}(n) - \mathbf{w}_j(n)\right) \tag{97}$$

Η σχέση αυτή εφαρμόζεται σε όλους τους νευρώνες του πλέγματος της τοπολογικής γειτονιάς του νικητή νευρώνα i. Αυτό που ουσιαστικά περιγράφει είναι η μετατόπιση του διανύσματος συναπτικών βαρών $\mathbf{w}_i$ του νικητή νευρώνα προς το διάνυσμα εισόδου $\mathbf{x}$, όπως φαίνεται στο σχήμα 19.



Σχήμα 19: Γραφική απεικόνιση της λειτουργίας του αλγορίθμου SOM στην τοπολογική γειτονιά του νικητή νευρώνα. Το πρότυπο εισόδου συμβολίζεται με x, ενώ η κανονική γραμμή αντιστοιχεί στην αρχική κατάσταση και η διακεκομμένη στην κατάσταση μετά την εμφάνιση του x.

Με αυτό τον τρόπο, μετά το πέρας της παρουσίασης των προτύπων εκπαίδευσης τα διανύσματα συναπτικών βαρών τείνουν να ακολουθήσουν την κατανομή των προτύπων εισόδου, οπότε τελικά ο αλγόριθμος οδηγεί στη δημιουργία μιας τοπολογικής διάταξης του χάρτη χαρακτηριστικών στο χώρο εισόδου. Αυτό δικαιολογείται από το γεγονός ότι οι νευρώνες που είναι γειτονικοί στο πλέγμα θα τείνουν να έχουν παρόμοια διανύσματα συναπτικών βαρών.

Η παράμετρος μάθησης όπως υποδεικνύεται από τη σχέση (97) είναι επίσης μεταβαλλόμενη στο χρόνο. Συγκεκριμένα υπάρχει η απαίτηση να ξεκινά από κάποια αρχική τιμή η_0 και στη συνέχεια να φθίνει βαθμιαία με το χρόνο. Έτσι, μπορεί να χρησιμοποιηθεί η εκθετική μείωση ως εξής:

$$\eta(n) = \eta_0 \exp\left(-\frac{n}{\tau_2}\right), \qquad n = 1, 2, \ldots \tag{98}$$

όπου τ_2 είναι μία ακόμα σταθερά του αλγορίθμου SOM. Η χρήση της εκθετικής μείωσης τόσο στη σχέση (90) όσο και στη σχέση (98) δεν είναι η βέλτιστη επιλογή, είναι όμως επαρκής για το σχηματισμό ενός αυτο-οργανούμενου χάρτη.

Πίνακας 3: Σύνοψη του Αλγορίθμου Χαρτών Αυτο-οργάνωσης (SOM)

1. ***Αρχικοποίηση***
   Επίλεξε τυχαίες τιμές για τα αρχικά διανύσματα συναπτικών βαρών $\mathrm{w}_j(0)$ υπό τον περιορισμό ότι οι τιμές $\mathrm{w}_j(0)$ πρέπει να είναι διαφορετικές για κάθε νευρώνα του πλέγματος και συνήθως μικρές τιμές.

2. ***Δειγματοληψία***
   Πάρε ένα δείγμα x από το χώρο εισόδου με μια συγκεκριμένη πιθανότητα. Το διάνυσμα x αντιπροσωπεύει το πρότυπο ενεργοποίησης που εφαρμόζεται στο πλέγμα.

3. ***Ταίριασμα Ομοιότητας***
   Βρες τον νικητή νευρώνα $i(\mathrm{x})$ στο χρονικό βήμα n χρησιμοποιώντας το κριτήριο ελάχιστης απόστασης
   $$i(\mathrm{x}) = \arg\min_j \|\mathrm{x} - \mathrm{w}_j\|, \qquad j \in \mathcal{A}$$

4. ***Ενημέρωση***
   Προσάρμοσε τα διανύσματα συναπτικών βαρών όλων των διεγερμένων νευρώνων χρησιμοποιώντας τον τύπο ενημέρωσης
   $$\mathrm{w}_j(n+1) = \mathrm{w}_j(n) + \eta(n) h_{j,i(\mathrm{x})}(n) \left(\mathrm{x}(n) - \mathrm{w}_j(n)\right)$$
   όπου $\eta(n)$ η παράμετρος του ρυθμού μάθησης και $h_{j,i(\mathrm{x})}(n)$ είναι η συνάρτηση γειτονιάς κεντραρισμένη γύρω από το νικητή νευρώνα $i(\mathrm{x})$. Οι $\eta(n)$ και $h_{j,i(\mathrm{x})}(n)$ μεταβάλλονται δυναμικά κατά τη διάρκεια της μάθησης.

5. ***Συνέχιση***
   Αύξησε το χρονικό βήμα n κατά 1 και επίστρεψε στο βήμα 2 μέχρι να μην παρατηρούνται πλέον ευδιάκριτες αλλαγές στο χάρτη χαρακτηριστικών.

Η προσαρμοστική διαδικασία που περιγράφηκε μπορεί να αναλυθεί σε δύο φάσεις, τη φάση διάταξης και στη συνέχεια τη φάση σύγκλισης. Η *φάση διάταξης* περιλαμβάνει τη διαδικασία της τοπολογικής διάταξης των διανυσμάτων βαρών. Η φάση αυτή μπορεί να απαιτήσει 1000 ή και περισσότερες επαναλήψεις, οπότε είναι κρίσιμης σημασίας να γίνει σωστή επιλογή των παραμέτρων του ρυθμού μάθησης και της συνάρτησης γειτονιάς. Η συνάρτηση γειτονιάς στην αρχή του αλγορίθμου πρέπει να περιλαμβάνει όλους σχεδόν τους νευρώνες του δικτύου κεντραρισμένους ως προς το νικητή νευρώνα και στη συνέχεια η γειτονιά σταδιακά να μειώνεται ώστε να περιλαμβάνει τελικά μόνο λίγους γειτονικούς νευρώνες γύρω από το νικητή ή και μόνο το νικητή νευρώνα. Η *φάση σύγκλισης* αφορά στην εκτέλεση μικρών προσαρμογών στο χάρτη χαρακτηριστικών ώστε να παρέχεται ακριβέστερη στατιστική ποσοτικοποίηση του χώρου εισόδου. Ο αριθμός των επαναλήψεων που απαιτείται σε αυτή τη φάση είναι ακόμα μεγαλύτερος, συγκεκριμένα πρέπει να είναι τουλάχιστον 500-πλάσιος του αριθμού των νευρώνων του δικτύου. Οι παράμετροι επιλέγονται ώστε ο ρυθμός μάθησης να είναι αρκετά μικρότερος της φάσης διάταξης και τοπολογική γειτονιά ν περιέχει μόνο τους πλησιέστερους γείτονες του νικητή νευρώνα. Έτσι, ο αλγόριθμός SOM μπορεί να συνοψιστεί όπως στον πίνακα 3.

2.2.2. Ασαφής C-Μέσος (Fuzzy C-Means, FCM)

Ο αλγόριθμος του Ασαφούς C-Μέσου ανήκει στην ευρύτερη οικογένεια των αλγορίθμων *ασαφούς συσταδοποίησης (fuzzy clustering)* η οποία βασίζεται στην *ασαφή λογική (fuzzy logic)*. Η ιδέα στην οποία βασίζεται η ασαφής συσταδοποίηση είναι ότι ένα πρότυπο μπορεί να ανήκει ταυτόχρονα σε περισσότερες από μία κλάσεις. Αυτή η θεώρηση αντιπροσωπεύει ουσιαστικά την κατάσταση όπου η γνώση είναι προσεγγιστική, όπως συμβαίνει πολύ συχνά και στην ανθρώπινη νόηση όταν κάποιο αντικείμενο του περιβάλλοντος *μοιάζει* να είναι κάποιο γνωστό αντικείμενο, μοιάζει δηλαδή να ανήκει σε κάποια γνωστή κατηγορία ή αλλιώς συστάδα.

Ασαφής Συσταδοποίηση και Μέτρα

Έστω ένα σύνολο δεδομένων $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$. Η ασαφής m-συσταδοποίηση του X ορίζεται ως η διαίρεση του X σε m το πλήθος σύνολα (συστάδες) τα οποία χαρακτηρίζονται από m συναρτήσεις u_j, οι οποίες ονομάζονται *συναρτήσεις συμμετοχής* και για τις οποίες ισχύει

$$u_j : X \to [0,1], \qquad j = 1,2,\dots,m \tag{99}$$

$$\sum_{j=1}^{m} u_j(\mathbf{x}_i) = 1, \qquad i = 1,2,\dots,N \tag{100α}$$

$$0 < \sum_{i=1}^{N} u_j(\mathbf{x}_i) < N, \qquad j = 1,2,\dots,m \tag{100β}$$

Η τιμή της ασαφούς συνάρτησης συμμετοχής είναι μία μαθηματική περιγραφή της συστάδας η οποία δεν είναι επακριβώς καθορισμένη. Με άλλα λόγια κάθε διάνυσμα $\mathbf{x}$

ανήκει «σε κάποιο βαθμό» ταυτόχρονα σε περισσότερες από μία συστάδες, γεγονός το οποίο ποσοτικοποιείται από την αντίστοιχη τιμή του u_j στο διάστημα [0,1]. Τιμές της συνάρτησης συμμετοχής κοντά στη μονάδα καταδεικνύουν μεγάλο βαθμό συμμετοχής στη συστάδα j, ενώ μικρές τιμές αντιστοιχούν σε μικρό βαθμό συμμετοχής. Οι τιμές των συναρτήσεων συμμετοχής προσφέρουν επίσης πληροφορίες για τη δομή του συνόλου δεδομένων υπό την έννοια ότι αν μία συνάρτηση συμμετοχής έχει περίπου την ίδια τιμή για για δύο διανύσματα του X τότε τα δύο αυτά διανύσματα θεωρούνται όμοια μεταξύ τους. Η συνθήκη (100β) εγγυάται ότι δεν υπάρχουν απλές περιπτώσεις όπου υπάρχουν συστάδες που δεν μοιράζονται κανένα διάνυσμα.

Έστω, τώρα, τα διανύσματα **x** και **y** με πραγματικές τιμές των οποίων οι συνιστώσες x_i και y_i ανήκουν στο διάστημα [0,1] όπου $i = 1, 2, \dots, l$. Αντίθετα με τις προηγούμενες περιπτώσεις οι τιμές των x_i δεν είναι αποτέλεσμα κάποιας συσκευής μέτρησης. Αντ' αυτού η τιμή x_i ανάλογα με το αν είναι κοντά στο 1 (ή κοντά στο 0) συμβολίζει την πιθανότητα το πρότυπο να διαθέτει (ή να μη διαθέτει) το i-οστο χαρακτηριστικό. Όταν το x_i πλησιάζει το $1/2$ γίνεται λιγότερο σαφές εάν το **x** να διαθέτει το i-οστο χαρακτηριστικό ενώ όταν $x_i = 1/2$ είναι τελείως άγνωστο. Όπως είναι εύκολο να διακρίνει κανείς, η ασαφής λογική είναι μια γενίκευση της δυαδικής λογικής. Στην ασαφή λογική τίποτα δεν συμβαίνει με απόλυτη σιγουριά. Η δυαδική λογική είναι μια ειδική περίπτωση της ασαφούς λογικής όπου το x_i παίρνει μόνο τις τιμές 0 και 1.

Στο πλαίσιο της σύνδεσης της ασαφούς με τη δυαδική λογική, προκύπτει ότι ο τελεστής AND της δυαδικής λογικής στην ασαφή λογική ανάγεται στον τελεστή min ενώ ο τελεστής OR της δυαδικής λογικής ανάγεται στον τελεστή max. Ακόμη, η λογική άρνηση (NOT) του x_i αντιστοιχεί στο $1 - x_i$. Έτσι, η παρακάτω σχέση ταυτότητας μεταξύ δύο δυαδικών μεταβλητών a και b

$$(a \equiv b) = \big((NOT\ a) AND\ (NOT\ b)\big)\ OR\ (a\ AND\ b\ ) \tag{101}$$

στην περίπτωση της ασαφούς λογικής μεταφράζεται σε σχέση ομοιότητας μεταξύ των πραγματικών μεταβλητών στο διάστημα [0,1] x_i και y_i ως εξής:

$$s(x_i, y_i) = \max(\min(1 - x_i, 1 - y_i), \min(x_i, y_i)) \tag{102}$$

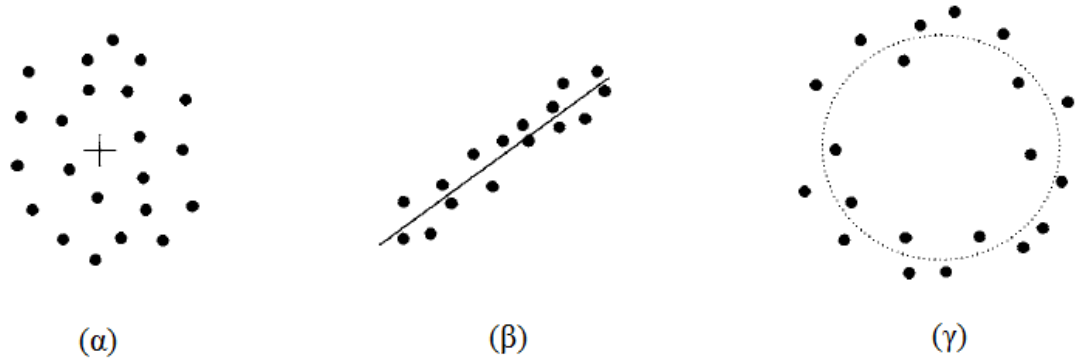
Ακόμη, προκειμένου για διανύσματα στον l-διάστατο χώρο, ο διανυσματικός χώρος είναι ο υπερκύβος H_l. Σύμφωνα με αυτή τη θεώρηση, όσο πιο κοντά βρίσκεται ένα διάνυσμα **x** στο κέντρο $(1/2, \dots, 1/2)$ του υπερκύβου τόσο πιο μεγάλη είναι η αβεβαιότητα για τα χαρακτηριστικά του **x**, ενώ, αντίθετα, όσο πιο κοντά βρίσκεται σε κάποια ακμή του υπερκύβου τόσο μεγαλύτερη είναι η βεβαιότητα. Με βάση, επομένως, την ομοιότητα s μεταξύ δύο μεταβλητών στο διάστημα [0,1] της σχέσης (102), ορίζεται το *μέτρο ομοιότητας μεταξύ δύο διανυσμάτων* **x** και **y:**

$$s_F^q(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^{l} s(x_i, y_i)^q \right)^{\frac{1}{q}} \tag{103}$$

Αλγόριθμος Ασαφούς c-Μέσου

Στη διατύπωση του αλγορίθμου FCM θεωρούμε ότι γνωρίζουμε εκ των προτέρων το πλήθος των συστάδων καθώς και το σχήμα τους. Το σχήμα των συστάδων χαρακτηρίζεται από τις παραμέτρους που υιοθετούνται, όπως φαίνεται στο σχήμα 20. Σε κάθε συστάδα

ανάλογα με το σχήμα της αντιστοιχίζεται ένα εκπρόσωπο οπότε η απόσταση ενός διανύσματος **x** από μία συστάδα ορίζεται ως η απόσταση του **x** από τον εκπρόσωπο.



Σχήμα 20: (α) Συμπαγής συστάδα με εκπρόσωπο ένα σημείο **(β)** Γραμμική συστάδα με εκπρόσωπο ένα υπερεπίπεδο **(γ)** Υπερσφαιρική συστάδα με εκπρόσωπο μία υπερσφαίρα

Έτσι, ανεξαρτήτως του σχήματος της συστάδας ορίζεται ο παρακάτω συμβολισμός:

$\boldsymbol{\theta}_j$: ο παραμετροποιημένος εκπρόσωπος της j-οστής συστάδας,

$$\boldsymbol{\theta} \equiv \left[\boldsymbol{\theta}_1{}^T, \boldsymbol{\theta}_2{}^T, \dots, \boldsymbol{\theta}_m{}^T\right]^T$$

U : ο πίνακας $N \times m$ του οποίου το (i,j) στοιχείο ισούται με το $u_j(\mathbf{x}_i)$

$d(\mathbf{x}_i, \boldsymbol{\theta}_j)$: η ανομοιότητα μεταξύ των $\mathbf{x}_i$ και $\boldsymbol{\theta}_j$, μετρούμενη με κάποιο μέτρο απόστασης όπως αναφέρθηκαν στην ενότητα 2.1.3

$q > 1$: μια παράμετρος που ονομάζεται *ασαφοποιητής (fuzzifier)*

Οι περισσότεροι αλγόριθμοι ασαφούς συσταδοποίησης προκύπτουν από την προσπάθεια ελαχιστοποίησης ως προς τα $\boldsymbol{\theta}$ και U μιας συνάρτησης κόστους της μορφής

$$J_q(\boldsymbol{\theta}, U) = \sum_{i=1}^{N}\sum_{j=1}^{m} u_{ij}^q d(\mathbf{x}_i, \boldsymbol{\theta}_j) \tag{104}$$

υπό τους περιορισμούς

$$\sum_{j=1}^{m} u_{ij} = 1\,, \qquad i = 1, 2, \dots, N \tag{105}$$

όπου

$$\begin{gathered} u_{ij} \in [0,1], \qquad i = 1, 2, \dots, N, \qquad j = 1,2, \dots, m \\ 0 < \sum_{i=1}^{N} u_{ij} < N, \qquad j = 1,2, \dots, m \end{gathered} \tag{106}$$

Ο όρος $d(\mathbf{x}_i, \boldsymbol{\theta}_j)$ αναπαριστά την απόσταση του $\mathbf{x}_i$ από το εκπρόσωπο της συστάδας, δηλαδή υπό μία έννοια το κέντρο μάζας της. Ο όρος $u_{ij}^q d(\mathbf{x}_i, \boldsymbol{\theta}_j)$ συμβολίζει το σφάλμα που υπεισέρχεται από τη χρήση του $\boldsymbol{\theta}_j$ ως εκπρόσωπο του $\mathbf{x}_i$ σταθμισμένο από το βαθμό συμμετοχής του $\mathbf{x}_i$ (υψωμένο σε κάποια δύναμη) στην κλάση j. Ο όρος $\sum_{j=1}^{m} u_{ij}^q d(\mathbf{x}_i, \boldsymbol{\theta}_j)$ αποτελεί το άθροισμα των σφαλμάτων λόγω της χρήσης εκπροσώπων σε όλες τις συστάδες, επομένως τελικά η συνάρτηση κόστους συνιστά το ολικό σταθμισμένο άθροισμα των γενικευμένων σφαλμάτων λόγω της αντικατάστασης του χώρου εισόδου από το $\boldsymbol{\theta}$.

Οι περιορισμοί της σχέσης (105) δηλώνουν ότι ο βαθμός συμμετοχής του $\mathbf{x}_i$ στη j-οστή συστάδα συνδέεται με το βαθμό συμμετοχής του $\mathbf{x}_i$ στις υπόλοιπες $m-1$ συστάδες. Η τιμή του q λειτουργεί ως πόλωση της $J_q(\boldsymbol{\theta}, U)$ είτε προς την ασαφή συσταδοποίηση είτε προς την αυστηρή συσταδοποίηση και κάθε τιμή του καθορίζει και έναν ασαφή αλγόριθμο. Δεν υπάρχει θεωρητικά ή υπολογιστικά κάποια βέλτιστη τιμή για το q οπότε η επιλογή του γίνεται πειραματικά

Για την ελαχιστοποίηση της συνάρτησης κόστους, θεωρείται κατ' αρχάς ότι κανένα $\mathbf{x}_i$ δεν συμπίπτει με κανέναν από τους εκπροσώπους των συστάδων. Πιο φορμαλιστικά, αν για τα $\mathbf{x}_i$ υπάρχει ένα σύνολο Z_i που περιλαμβάνει τους δείκτες των εκπροσώπων $\boldsymbol{\theta}_j$ για τους οποίους ισχύει $d(\mathbf{x}_i, \boldsymbol{\theta}_j) = 0$, τότε θα ισχύει $Z_i = \emptyset$ για όλα τα i. Η ελαχιστοποίηση της $J_q(\boldsymbol{\theta}, U)$ ως προς το U υπό τον περιορισμό (105) οδηγεί στην ακόλουθη Λαγκραντζιανή συνάρτηση:

$$\mathcal{J}(\boldsymbol{\theta}, U) = \sum_{i=1}^{N}\sum_{j=1}^{m} u_{ij}^{q} d(\mathbf{x}_i, \boldsymbol{\theta}_j) - \sum_{i=1}^{N} \lambda_i \left(\sum_{j=1}^{m} u_{ij} - 1 \right) \tag{107}$$

Η μερική παράγωγος του $\mathcal{J}(\boldsymbol{\theta}, U)$ ως προς το u_{rs} είναι

$$\frac{\partial \mathcal{J}(\boldsymbol{\theta}, U)}{\partial u_{rs}} = q {u_{rs}}^{q-1} d(\mathbf{x}_r, \boldsymbol{\theta}_s) - \lambda_r \tag{108}$$

Θέτοντας τη μερική παράγωγο αυτή ίση με 0 και λύνοντας ως προς u_{rs} προκύπτει

$$u_{rs} = \left(\frac{\lambda_r}{q d(\mathbf{x}_r, \boldsymbol{\theta}_s)} \right)^{\frac{1}{q-1}}, \qquad s = 1, 2, \ldots, m \tag{109}$$

Αντικαθιστώντας το u_{rs} της σχέσης (109) στη συνθήκη του περιορισμού (105) λαμβάνεται

$$\sum_{j=1}^{m} \left(\frac{\lambda_r}{q d(\mathbf{x}_r, \boldsymbol{\theta}_s)} \right)^{\frac{1}{q-1}} = 1$$

ή αλλιώς

$$\lambda_r = \frac{q}{\left(\sum_{j=1}^{m} \left(\frac{1}{q d(\mathbf{x}_r, \boldsymbol{\theta}_s)} \right)^{\frac{1}{q-1}} \right)^{q-1}} \tag{110}$$

Από τις σχέσεις (109) και (110) μετά από πράξεις προκύπτει

$$u_{rs} = \frac{1}{\sum_{j=1}^{m} \left(\frac{d(\mathbf{x}_r, \boldsymbol{\theta}_s)}{d(\mathbf{x}_r, \boldsymbol{\theta}_j)} \right)^{\frac{1}{q-1}}}, \qquad r = 1, 2, \ldots, N, \qquad s = 1, 2, \ldots, m \tag{111}$$

Επαναλαμβάνοντας τη διαδικασία για το διάνυσμα παραμέτρων $\boldsymbol{\theta}_j$, υπολογίζεται η κλίση της $\mathcal{J}(\boldsymbol{\theta}, U)$ ως προς το $\boldsymbol{\theta}_j$ και τίθεται ίση με 0 οπότε προκύπτει

$$\frac{\partial \mathcal{J}(\boldsymbol{\theta}, U)}{\partial \boldsymbol{\theta}_j} = \sum_{i=1}^{N} u_{ij}^{q} \frac{\partial d(\mathbf{x}_i, \boldsymbol{\theta}_j)}{\partial \boldsymbol{\theta}_j} = \mathbf{0}, \qquad j = 1,2, \ldots, m \tag{112}$$

Οι εξισώσεις (111) και (112) είναι συζευγμένες και γενικά δεν δίνουν λύσεις κλειστού τύπου. Ένας τρόπος να παρακαμφθεί αυτή η δυσκολία είναι να χρησιμοποιηθεί ο

αλγόριθμος του πίνακα 4 ώστε να βρεθούν προσεγγίσεις του U και του $\boldsymbol{\theta}$. Πρόκειται για μια γενικής χρήσης αλγοριθμική τεχνική στην ασαφή συσταδοποίηση η οποία αναφέρεται ως *Γενικό Αλγοριθμικό Σχέδιο Ασάφειας (Generalised Fuzzy Algorithmic Scheme, GFAS).*

Πίνακας 4: Περιγραφή του Αλγορίθμου GFAS

1. Διάλεξε τα $\theta_j(0)$ ως αρχικές εκτιμήσεις των θ_j για $j = 1,2,\dots,m$

2. Επανάλαβε

Για i=1 **μέχρι** N

Για j=1 **μέχρι** m

$$u_{rs}(n) = \frac{1}{\sum_{j=1}^{m}\left(\frac{d(\mathrm{x}_r, \theta_s)}{d(\mathrm{x}_r, \theta_j)}\right)^{\frac{1}{q-1}}}$$

Τέλος {Για- j}

Τέλος {Για- i}

n=n+1

Για j=1 **μέχρι** m

Βρες το $\theta_j(n)$ από την εξίσωση

$$\sum_{i=1}^{N} u_{ij}^{q}(n-1)\frac{\partial d(\mathrm{x}_i, \theta_j)}{\partial \theta_j} = 0$$

Τέλος {Για- j}

Μέχρις ότου ικανοποιηθεί το κριτήριο τερματισμού

Ως κριτήριο τερματισμού μπορεί να χρησιμοποιηθεί μία σχέση της μορφής $\|\boldsymbol{\theta}(n) - \boldsymbol{\theta}(n-1)\| < \varepsilon$, όπου το ε είναι μία σχετικά μικρή σταθερά που καθορίζεται από το σχεδιαστή.

Καταλήγοντας τώρα στον ζητούμενο αλγόριθμο FCM, επιλέγεται ως μέτρο της ανομοιότητας, δηλαδή της απόστασης του $\mathbf{x}_i$ από το κέντρο της συστάδας, το μετρικό

$$\boldsymbol{d}(\mathbf{x}_i, \boldsymbol{\theta}_j) = (\mathbf{x}_i - \boldsymbol{\theta}_j)^T A(\mathbf{x}_i - \boldsymbol{\theta}_j) = \|\mathbf{x}_i - \boldsymbol{\theta}_j\|_A^2 \tag{113}$$

δηλαδή μια A-νόρμα όπου A ένας συμμετρικός θετικός πίνακας βαρών. Ο πίνακας αυτός ελέγχει το σχήμα των βέλτιστων συστάδων και υπάρχουν άπειρες επιλογές για τις τιμές που μπορεί να πάρει. Οι πιο συνήθεις περιπτώσεις είναι η Ευκλείδεια νόρμα με $A = I$, η διαγώνια νόρμα με $A = D_x^{-1}$ και η νόρμα Mahalonobis με $A = C_x^{-1}$, όπου

$$c_x = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i \tag{114α}$$

$$C_x = \sum_{i=1}^{N}(\mathbf{x}_i - c_x)(\mathbf{x}_i - c_x)^t \tag{114β}$$

και D_x ο διαγώνιος πίνακας με στοιχεία της διαγωνίου ίσα με τις ιδιοτιμές του C_x.
Έτσι, έπεται ότι:

$$\frac{\partial d(\mathbf{x}_i, \boldsymbol{\theta}_j)}{\partial \boldsymbol{\theta}_j} = 2A(\boldsymbol{\theta}_j - \mathbf{x}_i) \tag{115}$$

και αντικαθιστώντας την εξίσωση (115) στη σχέση (112) υπολογίζεται ότι:

$$\sum_{i=1}^{N} u_{ij}^{q}\, 2A(\boldsymbol{\theta}_j - \mathbf{x}_i) = \mathbf{0} \tag{116}$$

Τελικά, λύνοντας ως προς το $\boldsymbol{\theta}_j$ προκύπτει η βασική σχέση του αλγορίθμου FCM ως μια υποπερίπτωση του GFAS

$$\boxed{\boldsymbol{\theta}_j(n) = \frac{\sum_{i=1}^{N} u_{ij}^{q}(n-1)\mathbf{x}_i}{\sum_{i=1}^{N} u_{ij}^{q}(n-1)}} \tag{117}$$

Παρ' όλο που ο αλγόριθμος FCM περιλαμβάνει απλά μία εργασία ελαχιστοποίησης μιας συνάρτησης κόστους, δεν υπάρχει καλή γνώση για τη συμπεριφορά του ως προς τη σύγκλιση. Συγκεκριμένα, έχει αποδειχθεί ότι όταν χρησιμοποιείται η απόσταση Mahalanobis η επαναληπτική σειρά που παράγει ο αλγόριθμος FCM είτε συγκλίνει σε ένα σταθερό σημείο της συνάρτησης κόστους σε πεπερασμένο αριθμό επαναλήψεων είτε περιέχει τουλάχιστον μία υποσειρά η οποία συγκλίνει σε ένα σταθερό σημείο της συνάρτησης κόστους. Αυτό το σημείο μπορεί να είναι τοπικό ή ολικό ελάχιστο οπότε πρέπει να μελετηθεί η φύση του σημείου σύγκλισης.

3. ΕΠΙΛΟΓΗ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ

Η εργασία της αναγνώρισης προτύπων μπορεί να αφορά δεδομένα τα οποία αποτελούνται ακόμα και από εκατοντάδες χαρακτηριστικά. Το γεγονός αυτό, το οποίο αναφέρεται συχνά και ως «*κατάρα της διαστατικότητας*» (*dimensionality curse*), πέραν του ότι δημιουργεί μεγάλο υπολογιστικό φορτίο, δημιουργεί και μία σειρά άλλων δυσκολιών στη διαδικασία της κατηγοριοποίησης. Επίσης, υπάρχει περίπτωση δύο χαρακτηριστικά όταν χρησιμοποιηθούν ξεχωριστά να περιέχουν ιδιαίτερα χρήσιμη πληροφορία για την κατηγοριοποίηση, όταν όμως χρησιμοποιηθούν από κοινού στο διάνυσμα χαρακτηριστικών να μην προσφέρουν αξιόλογο κέρδος ενώ παράλληλα όμως προσθέτουν πολυπλοκότητα στη διαδικασία. Αυτό συμβαίνει όταν τα δύο αυτά χαρακτηριστικά εμφανίζουν μεγάλη συσχέτιση.

Επιπλέον, προκειμένου για επιβλεπόμενη εκπαίδευση, όσο μεγαλύτερος είναι ο λόγος του πλήθους των προτύπων εκπαίδευσης N προς τις ελεύθερες παραμέτρους της κατηγοριοποίησης l τόσο καλύτερη είναι η ικανότητα γενίκευσης του κατηγοριοποιητή που προκύπτει. Ένας μεγάλος αριθμός χαρακτηριστικών μεταφράζεται σε μεγάλο αριθμό παραμέτρων κατηγοριοποίησης (συναπτικά βάρη στην περίπτωση νευρωνικών δικτύων, βάρη στην περίπτωση γραμμικού κατηγοριοποιητή). Άρα, είναι πολύ σημαντικό για την καλή απόδοση στην γενίκευση του κατηγοριοποιητή να μειωθεί το πλήθος των χαρακτηριστικών. Ένα ακόμα σημείο όπου ο λόγος N/l παίζει σημαντικό ρόλο είναι στο στάδιο της αξιολόγησης του κατηγοριοποιητή. Όσο μεγαλύτερος είναι ο λόγος αυτός τόσο βελτιώνεται το σφάλμα κατηγοριοποίησης.

Το πρόβλημα λοιπόν που απαιτείται να λυθεί είναι το εξής: Με δεδομένο ένα σύνολο χαρακτηριστικών D, πώς μπορεί κανείς να επιλέξει ένα υποσύνολο d με τα σημαντικότερα από αυτά έτσι ώστε να μειωθεί ο αριθμός τους αλλά ταυτοχρόνως να διατηρηθεί όσο το δυνατόν περισσότερη διακριτική πληροφορία του αρχικού σύνολου. Η διαδικασία επίλυσης ενός τέτοιου προβλήματος ονομάζεται *επιλογή χαρακτηριστικών ή μείωση χαρακτηριστικών (feature selection/reduction)*. Το στάδιο αυτό, όπως γίνεται εμφανές, είναι πολύ κρίσιμο για την απόδοση του κατηγοριοποιητή που θα προκύψει. Η εξαντλητική αναζήτηση του βέλτιστου υποσυνόλου όπως είναι εύκολα κατανοητό είναι απαγορευτική στην περίπτωση διανύσματος χαρακτηριστικών μεγάλης διαστατικότητας οπότε είναι απαραίτητη η χρήση πιο αποδοτικών μεθόδων. Οι δύο βασικότερες τεχνικές που χρησιμοποιούνται είναι η εξέταση κάθε χαρακτηριστικού ξεχωριστά ως προς τη διακριτική του ικανότητα και η εξέταση των χαρακτηριστικών σε συνδυασμούς, ενώ μπορεί επί του διανύσματος χαρακτηριστικών να χρησιμοποιηθούν και γραμμικοί ή μη γραμμικοί μετασχηματισμοί ώστε να προκύψει ένα καινούριο με καλύτερες διακριτικές ιδιότητες.

Στη συνέχεια, θα μελετηθούν δύο αποδοτικές μέθοδοι επιλογής χαρακτηριστικών, η μέθοδος της *Σειριακής Εμπρόσθιας Μεταβλητής Επιλογής (Sequential Forward Floating Selection, SFFS)* και η μέθοδος *Επιλογής Αμοιβαίας Πληροφορίας (Mutual Information).*

3.1. Σειριακή Εμπρόσθια Μεταβλητή Επιλογή (Sequential Forward Floating Selection, SFFS)

Η τεχνική της Σειριακής Εμπρόσθιας Μεταβλητής Επιλογής (Sequential Forward Floating Selection, SFFS) αποτελεί βελτιστοποίηση της τεχνικής Σειριακής Εμπρόσθιας Επιλογής (Sequential Forward Selection, SFS). Η τεχνική SFS είναι προβληματική, αποτέλεσε όμως βάση για την ανάπτυξη της SFFS. Σύμφωνα με αυτή την προσέγγιση, το βέλτιστο σύνολο χαρακτηριστικών X_d με πληθικότητα d προκύπτει από ένα αρχικά κενό σύνολο X_0 στο οποίο σταδιακά προστίθενται χαρακτηριστικά από το αρχικό σύνολο Y σύμφωνα με κάποιο μέτρο, ακολουθείται δηλαδή η στρατηγική «bottom up». Η διαδικασία επιλογής της SFS αποτελείται από τα παρακάτω βήματα:

1. *Όρισε ένα κριτήριο διαχωρισιμότητας των κλάσεων C*
2. *Υπολόγισε την τιμή του C για κάθε χαρακτηριστικό και*
3. *Επίλεξε το χαρακτηριστικό με την καλύτερη τιμή.*
4. $k = 2$
5. *Σχημάτισε όλα τα πιθανά k-διάστατα διανύσματα που περιέχουν το διάνυσμα-νικητή που προέκυψε από το προηγούμενο βήμα*
6. *Υπολόγισε την τιμή του C για κάθε διάνυσμα του προηγούμενου βήματος*
7. *Επίλεξε το διάνυσμα με την καλύτερη τιμή*
8. *Αύξησε το k κατά 1 και συνέχισε στο βήμα 5 μέχρι $k = d$*
9. *Το διάνυσμα που προκύπτει περιέχει τα στοιχεία του X_d*

Στη γενική περίπτωση όπου το αρχικό σύνολο χαρακτηριστικών περιλαμβάνει m χαρακτηριστικά και μειώνεται σε ένα σύνολο d χαρακτηριστικών ελέγχονται $dm - d(d-1)/2$ συνδυασμοί.

Προφανώς η μέθοδος αυτή δεν υπακούει στην αρχή της βελτιστότητας αφού άπαξ και ένα στοιχείο του Y επιλεχθεί για το υποσύνολο δεν υπάρχει τρόπος να βγει ώστε να προκύψει κάποιος καλύτερος συνδιασμός. Η συμπεριφορά αυτή του αλγορίθμου αναφέρεται και ως *φαινόμενο φωλιάσματος (nesting effect)*. Έτσι, προκειμένου να προκύψει ένα βέλτιστο υποσύνολο είναι απαραίτητο να προστεθεί ένας μηχανισμός επανεξέτασης όλων των στοιχείων του υποσυνόλου σε κάθε επανάληψη ώστε να αφαιρεθεί κάποιο αν αυτό κριθεί απαραίτητο. Η μέθοδος που επεκτείνει την μέθοδο SFS προς αυτή την κατεύθυνση είναι η SFFS.

Πριν την φορμαλιστική διατύπωση του αλγορίθμου SFFS, πρέπει να δοθούν οι ακόλουθοι ορισμοί. Έστω $X_k = \{x_i : 1 \le i \le k,\ x_i \in Y\}$ υποσύνολο k χαρακτηριστικών από το σύνολο $Y = \{y_i : 1 \le i \le D\}$ των D διαθέσιμων χαρακτηριστικών. Η τιμή $J(y_i)$ της συναρτησης-κριτηρίου επιλογής χαρακτηριστικών C, στην περίπτωση που χρησιμοποιείται μόνο το i-οστό χαρακτηριστικό $y_i,\ i = 1,2,\dots,D$, ονομάζεται *ατομική σημαντικότητα (individual significance)* $S_0(y_i)$ του χαρακτηριστικού. Η *σημαντικότητα* $S_{k-1}(x_j)$ *του χαρακτηριστικού* $x_j, j = 1,2,\dots,k$, *στο σύνολο* X_k ορίζεται ως

$$S_{k-1}(x_j) = J(X_k) - J(X_k - x_j) \tag{1}$$

Η *σημαντικότητα* $S_{k+1}(f_j)$ *του χαρακτηριστικού* f_j *του σύνολου* $Y - X_k = \{f_i : 1 \le i \le D - k,\ f_i \in Y,\ f_i \ne x_l\ \ \forall x_l \in X_k\}$ *ως προς το σύνολο* X_k ορίζεται ως

$$S_{k+1}(f_j) = J(X_k + f_j) - J(X_k) \tag{2}$$

Όπως είναι εμφανές, για $k = 1$ ο όρος της σημαντικότητας του χαρακτηριστικού ενός συνόλου της σχέσης (1) συμπίπτει με την ατομική σημαντικότητα. Ένα χαρακτηριστικό x_j του συνόλου X_k ονομάζεται

(α) το *σημαντικότερο* (καλύτερο) χαρακτηριστικό *στο σύνολο* X_k εάν

$$S_{k-1}(x_j) = \max_{1 \le i \le k} S_{k-1}(x_i) \Longrightarrow J(X_k - x_j) = \min_{1 \le i \le k} J(X_k - x_i) \tag{3}$$

(β) το *λιγότερο σημαντικό* (χειρότερο) χαρακτηριστικό *στο σύνολο* X_k εάν

$$S_{k-1}(x_j) = \min_{1 \le i \le k} S_{k-1}(x_i) \Longrightarrow J(X_k - x_j) = \max_{1 \le i \le k} J(X_k - x_i) \tag{4}$$

Ένα χαρακτηριστικό f_j του συνόλου $Y - X_k$ ονομάζεται

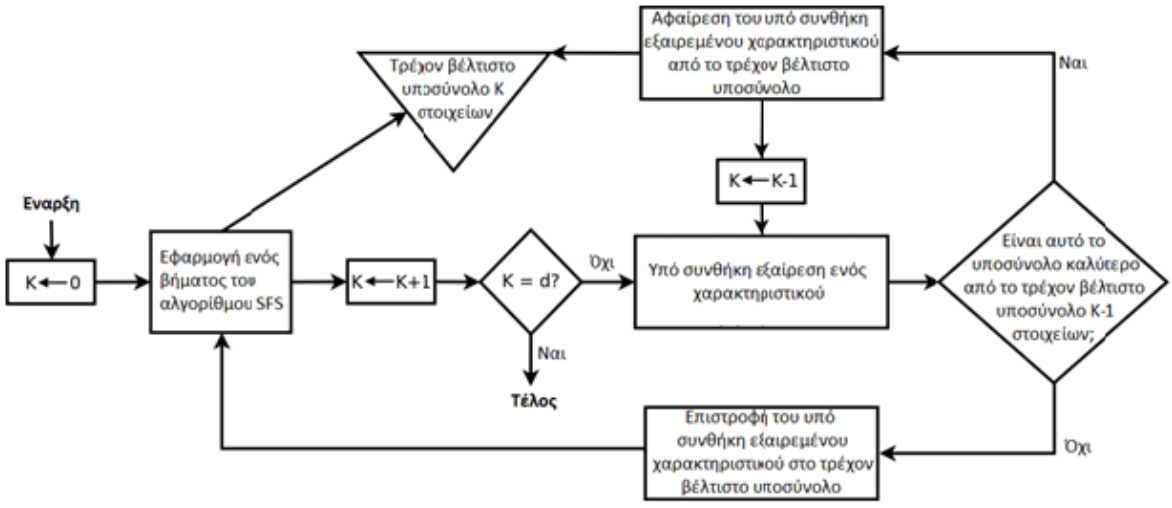
(α) το *σημαντικότερο* (καλύτερο) χαρακτηριστικό *ως προς το σύνολο* X_k εάν

$$S_{k+1}(f_j) = \max_{1 \le i \le D-k} S_{k+1}(f_i) \Longrightarrow J(X_k + f_j) = \min_{1 \le i \le D-k} J(X_k + f_i) \tag{5}$$

(β) το *λιγότερο σημαντικό* (χειρότερο) χαρακτηριστικό *ως προς το σύνολο* X_k εάν

$$S_{k+1}(f_j) = \min_{1 \le i \le k} S_{k+1}(f_i) \Longrightarrow J(X_k + f_j) = \max_{1 \le i \le k} J(X_k + f_i) \tag{6}$$

Έτσι, με τη βοήθεια των παραπάνω ορισμών είναι πλέον δυνατή η διατύπωση του αλγορίθμου SFFS. Η βασική ιδέα του αλγορίθμου είναι ότι επιλέγονται τα νέα χαρακτηριστικά σύμφωνα με τη διαδικασία SFS ξεκινώντας από το τρέχον σύνολο, στη συνέχεια όμως ακολουθείται μία σειρά από διαδοχικούς αποκλεισμούς του χειρότερου χαρακτηριστικού του μόλις ανανεωμένου υποσυνόλου. Ο αλγόριθμος SFFS παρουσιάζεται στον πίνακα 1 ενώ το αντίστοιχο διάγραμμα ροής φαίνεται στην εικόνα 1.



Σχήμα 1: Το διάγραμμα ροής του αλγορίθμου επιλογής χαρακτηριστικών SFFS

Πίνακας 1: Σύνοψη του Αλγορίθμου SFFS

1. ***Αρχική Επιλογή***

Χρησιμοποιώντας τη βασική μέθοδο SFS επίλεξε το σημαντικότερο χαρακτηριστικό x_{k+1} ως προς το σύνολο X_k από το σύνολο των διαθέσιμων μετρήσεων $Y - X_k$, ώστε να σχηματιστεί το σύνολο $X_{k+1} = X_k + x_{k+1}$.

2. ***Εξαίρεση υπό Συνθήκη***

Βρες το λιγότερο σημαντικό χαρακτηριστικό στο σύνολο X_{k+1}.

2.1. Αν το x_{k+1} είναι το λιγότερο σημαντικό χαρακτηριστικό του X_k δηλαδή $J(X_{k+1} - x_{k+1}) \geq J(X_{k+1} - x_j)$, $\forall j = 1,2, \dots, k$

Θέσε $k = k + 1$ και επίστρεψε στο βήμα 1.

2.2. Αν κάποιο x_r, $1 \leq r \leq k$ είναι το λιγότερο σημαντικό χαρακτηριστικό στο σύνολο X_{k+1}, δηλαδή $J(X_{k+1} - x_r) > J(X_k)$ τότε αφαίρεσε το x_r από το X_{k+1} οπότε προκύπτει το σύνολο $X'_k = X_{k+1} - x_r$.

(Ισχύει $J(X'_k) > J(X_k)$)

2.2.1. Αν $k = 2$ τότε θέσε $X_k = X'_k$ και $J(X'_k) = J(X_k)$ και επίστρεψε στο βήμα 1.

Αλλιώς πήγαινε στο βήμα 3.

3. ***Συνέχιση Εξαίρεσης υπό Συνθήκη***

Βρες το λιγότερο σημαντικό χαρακτηριστικό x_s στο σύνολο X'_k.

3.1. Αν $J(X'_k - x_s) \leq J(X'_{k-1})$ τότε θέσε $X_k = X'_k$ και $J(X'_k) = J(X_k)$ και επίστρεψε στο βήμα 1.

3.2. Αν $J(X'_k - x_s) > J(X'_{k-1})$ τότε αφαίρεσε το x_s από το X'_k οπότε προκύπτει το σύνολο $X'_{k-1} = X'_k - x_s$.

Θέσε $k = k - 1$.

3.2.1. Αν $k = 2$ τότε θέσε $X_k = X'_k$ και $J(X'_k) = J(X_k)$ και επίστρεψε στο βήμα 1.

Αλλιώς πήγαινε στο βήμα 3.

Ο αλγόριθμος SFFS του πίνακα 1 αρχικοποιείται με $k = 0$ και $X_0 = \emptyset$ και στη συνέχεια χρησιμοποιείται ο αλγόριθμος SFS μέχρι να προκύψει το X_2, να προκύψει δηλαδή ένα σύνολο με δύο χαρακτηριστικά και στη συνέχεια μεταβαίνει στο βήμα 1. Τερματίζεται όταν προκύψει σύνολο με τον επιθυμητό αριθμό χαρακτηριστικών. Παρ' όλο που ο SFFS δεν εγγυάται την εύρεση όλων των βέλτιστων υποσυνόλων, αποδίδει πολύ καλύτερα σε σύγκριση με τον απλό SFS υπό το κόστος βέβαια της αυξημένης πολυπλοκότητας.

3.2. Αμοιβαία Πληροφορία (Mutual Information)

Μία άλλη προσέγγιση στην επιλογή χαρακτηριστικών στα συστήματα αναγνώρισης προτύπων είναι το κριτήριο μέγιστης στατιστικής εξάρτησης που βασίζεται στην *αμοιβαία πληροφορία*. Η άμεση υλοποίηση της συνθήκης μέγιστης εξάρτησης παρουσιάζει όμως σημαντική δυσκολία. Για το λόγο αυτό εισάγεται αρχικά ένα ισοδύναμο κριτήριο, το *κριτήριο ελάχιστου πλεονασμού-μέγιστης σχέσης (minimal redundancy-maximal relevance, mRMR)*, ενώ στη συνέχεια χρησιμοποιείται ένας αλγόριθμος επιλογής χαρακτηριστικών δύο σταδίων συνδυάζοντας το mMRM μαζί με άλλες, πιο σύνθετες μεθόδους επιλογής χαρακτηριστικών. Επιτυγχάνεται έτσι ένα συμπαγές σύνολο με τα επικρατέστερα χαρακτηριστικά με μικρό κόστος.

Έστω ότι διατίθεται ένα σύνολο δεδομένων με D χαρακτηριστικά στο διάνυσμα χαρακτηριστικών τους, τα οποία συγκροτούν το σύνολο $X = \{x_i,\ i = 1,2,\dots,M\}$ και c η μεταβλητή-στόχος της κατηγοριοποίησης. Ο σκοπός της επιλογής χαρακτηριστικών, όπως έχει προαναφερθεί, είναι να βρεθεί το υποσύνολο του X με m στοιχεία που χαρακτηρίζει την c με το βέλτιστο τρόπο. Η εργασία αυτή μεταφράζεται στην μείωση του ελάχιστου σφάλματος κατηγοριοποίησης. Στην περίπτωση της μη επιβλεπόμενης κατηγοριοποίησης όπου οι κατηγοριοποιητές δεν είναι καθορισμένοι, το ελάχιστο σφάλμα απαιτεί τη μέγιστη στατιστική εξάρτηση της κλάσης c από την κατανομή των δεδομένων στον υποχώρο $\mathbb{R}^m$, εισάγεται δηλαδή το σχήμα της *μέγιστης εξάρτησης* (Max-Dependency). Η μέγιστη εξάρτηση τώρα μπορεί να υλοποιηθεί σύμφωνα με την προσέγγιση της επιλογής χαρακτηριστικών μέγιστης σχέσης (Max-Relevance), δηλαδή επιλέγονται τα χαρακτηριστικά με τη μεγαλύτερη σχέση με την κλάση-στόχο c. Η έννοια της «σχέσης» μεταφράζεται συνήθως στη συσχέτιση ή την αμοιβαία πληροφορία, από τις οποίες εδώ ενδιαφέρει η δεύτερη.

Δεδομένων δύο τυχαίων μεταβλητών x και y, η αμοιβαία πληροφορία τους ορίζεται σε σχέση με τις συναρτήσεις πυκνότητας πιθανότητας $p(x)$, $p(y)$ και $p(x,y)$ ως

$$I(x;y) = \int\int p(x,y)\log\frac{p(x,y)}{p(x)p(y)}dxdy \tag{7}$$

Η σχέση αυτή εκφράζει την αβεβαιότητα σχετικά με την είσοδο x του συστήματος που εξαλείφεται όταν παρατηρηθεί στο σύστημα η έξοδος y. Για να ικανοποιείται το κριτήριο μέγιστης σχέσης τα χαρακτηριστικά x_i που επιλέγονται απαιτείται το καθένα ξεχωριστά να έχουν τη μέγιστη αμοιβαία πληροφορία $I(x_i\,;c)$ με την κλάση-στόχο c. Με όρους σειριακής αναζήτησης, κατατάσσοντας τα χαρακτηριστικά σε φθίνουσα σειρά με βάση το $I(x_i\,;c)$, τα πρώτα m χαρακτηριστικά, είναι αυτά που επιλέγονται. Το κριτήριο του ελάχιστου πλεονασμού αφορά στην προσπάθεια να μειωθεί η πλεονάζουσα πληροφορία μεταξύ των χαρακτηριστικών, αφού ο συνδυασμός ατομικά «καλών» χαρακτηριστικών δεν οδηγεί απαραίτητα σε κατηγοριοποίηση με καλές επιδόσεις.

Στο πλαίσιο της αμοιβαίας πληροφορίας, ο σκοπός της επιλογής χαρακτηριστικών είναι η εύρεση ενός συνόλου χαρακτηριστικών X_m με m χαρακτηριστικά $\{x_i\}$ τα οποία συνδυασμένα έχουν την *μέγιστη εξάρτηση* από την κλάση-στόχο c, όπως εκφράζει η σχέση

$$\max D(X,c),\ D = I(\{x_i,\ i = 1,2,\dots,m\}\,;c) \tag{8}$$

Για d=1 η λύση είναι το χαρακτηριστικό που μεγιστοποιεί την $I(x_j\,;c)$, $1 \le j \le M$. Για $m > 1$ ακολουθείται η επαυξητική λογική σύμφωνα με την οποία σε κάθε επανάληψη προστίθεται ένα χαρακτηριστικό ως εξής: αν το τρέχον σύνολο X_{m-1} διαθέτει $m-1$ χαρακτηριστικά, το m-οστο χαρακτηριστικό προσδιορίζεται από το κατά πόσο συνεισφέρει στη μεγαλύτερη αύξηση του $I(X\,;c)$, δηλαδή

$$\begin{aligned} I(X_m\,;c) &= \iint p(X_m,c)\log\frac{p(X_m,c)}{p(X_m)p(c)}\,dX_m dc \\ &= \iint p(X_{m-1},x_m,c)\log\frac{p(X_{m-1},x_m,c)}{p(X_{m-1},x_m)p(c)}\,dX_{m-1}dx_m dc \\ &= \int\dots\int p(x_1,\dots,x_m,c)\log\frac{p(x_1,\dots,x_m,c)}{p(x_1,\dots,x_m)p(c)}\,dx_1\dots dx_m dc \end{aligned} \tag{9}$$

Οι συναρτήσεις πυκνότητας πιθανότητας της σχέσης (9) με τόσες πολλές μεταβλητές όμως, είναι πολύ δύσκολο να υπολογιστούν σε χώρους πολλών διαστάσεων καθώς απαιτείται μεγάλος αριθμός δειγμάτων ενώ οι υπολογισμοί είναι πολύ επίπονοι υπολογιστικά. Για το λόγο αυτό, η χρήση του κριτηρίου μέγιστης εξάρτησης δεν είναι πρακτικά εφαρμόσιμη παρά μόνο σε περιορισμένο αριθμό περιπτώσεων.

Ως εναλλακτική του κριτηρίου μέγιστης εξάρτησης εισάγεται το κριτήριο *μέγιστης σχέσης*. Για το σκοπό αυτό η $D(X,c)$ της σχέσης (8) προσεγγίζεται από τη μέση τιμή όλων των τιμών αμοιβαίας πληροφορίας μεταξύ της κλάσης c και του κάθε χαρακτηριστικού x_i ξεχωριστά, δηλαδή

$$\max D(X,c),\quad D=\frac{1}{|X|}\sum_{x_i\in X} I(x_i\,;c) \tag{10}$$

Τα χαρακτηριστικά που επιλέγονται σύμφωνα με το κριτήριο αυτό όμως, είναι πιθανό να περιλαμβάνουν πλεονασμούς, δηλαδή μερικά χαρακτηριστικά να παρουσιάζουν συσχέτιση προσθέτοντας υπολογιστική πολυπλοκότητα χωρίς ιδιαίτερο κερδος για την κατηγοριοποίηση. Για το σκοπό αυτό προστίθεται και η συνθήκη *ελάχιστου πλεονασμού* ώστε να επιλεχθούν χαρακτηριστικά χωρίς μεγάλη αμοιβαία πληροφορία ως εξής:

$$\min R(X),\quad R=\frac{1}{|X|^2}\sum_{x_i,x_j\in X} I(x_i\,;x_j) \tag{11}$$

Ο συνδυασμός των δύο περιορισμών είναι που συγκροτεί το *κριτήριο ελάχιστου πλεονασμού-μέγιστης σχέσης* (mRMR). Για το συνδυασμό αυτό των D και R εισάγεται και ο τελεστής $\Phi(D,R)$ και για να βελτιστοποιούνται αυτά ταυτόχρονα ισχύει

$$\max \Phi(D,R)\,,\qquad \Phi = D-R \tag{12}$$

Μία κοινή πρακτική είναι να χρησιμοποιηθεί κάποια επαυξητική μέθοδος, η οποία όμως δίνει ένα *σχεδόν* βέλτιστο σύνολο χαρακτηριστικών όπως καθορίζει η $\Phi(\cdot)$. Αν διαθέτουμε το σύνολο X_{m-1} με επιλεγμένα $m-1$ χαρακτηριστικά. Το m-οστό χαρακτηριστικό επιλέγεται από το σύνολο $Y-X_{m-1}$ μεγιστοποιώντας την $\Phi(\cdot)$. Ο αντίστοιχος επαυξητικός αλγόριθμος, επομένως, θα πρέπει να βελτιστοποιεί την ακόλουθη συνθήκη

$$\max_{x_j \in Y - X_{m-1}} \left[I(x_j ; c) - \frac{1}{m-1} \sum_{x_i \in X_{m-1}} I(x_j ; x_i) \right] \quad (13)$$

Η πολυπλοκότητα αυτής της μεθόδου επαυξητικής αναζήτησης είναι $O(|S|M)$. Αποδεικνύεται ότι για κατηγοριοποίηση πρώτης τάξης, δηλαδή στην περίπτωση που διατίθεται το σύνολο X_{m-1} και σε κάθε βήμα προστίθεται ένα χαρακτηριστικό, το κριτήριο mRMR είναι *ισοδύναμο* με το κριτήριο μέγιστης εξάρτησης. Έτσι, ο υπολογισμός των συναρτήσεων πυκνότητας πιθανότητας πολλών μεταβλητών $p(x_1, \dots, x_m, c)$ αντικαθίσταται από τον υπολογισμό των συναρτήσεων δύο μεταβλητών $p(x_i, c)$ και $p(x_j, x_i)$, ο οποίος είναι πολύ απλούστερος και ακριβέστερος.

Μετά την παρουσίαση του τρόπου επιλογής των καλύτερων χαρακτηριστικών από το διαθέσιμο σύνολο, είναι απαραίτητο να προσδιοριστεί το βέλτιστο πλήθος m των χαρακτηριστικών που πρέπει να επιλεχθούν. Δεδομένου ότι η επαυξητική επιλογή δεν περιλαμβάνει μηχανισμό αφαίρεσης τυχόν πλεοναζόντων χαρακτηριστικών όπως υπαγορεύει το κριτήριο mRMR, πρέπει γίνει μία διαλογή των χαρακτηριστικών του συνόλου που προκύπτει. Επομένως, ο συνολικός αλγόριθμος επιλογής χαρακτηριστικών περιλαμβάνει ένα στάδιο όπου χρησιμοποιείται ο επαυξητικός αλγόριθμος απ' όπου προκύπτει ένα υποψήφιο σύνολο και ένα δεύτερο στάδιο αναζητείται ένα συμπαγές υποσύνολο του υποψήφιου συνόλου το οποίο να είναι απαλλαγμένο από πλεονασμούς.

Για το πρώτο στάδιο, την επιλογή του υποψήφιου συνόλου, υπολογίζεται το σφάλμα κατηγοριοποίησης με χρήση διασταυρωμένης επικύρωσης για ένα μεγάλο αριθμό χαρακτηριστικών. Στη συνέχεια καθορίζεται ένα σχετικά ευσταθές εύρος μικρού σφάλματος, το οποίο συμβολίζεται ως Ω. Το βέλτιστο πλήθος χαρακτηριστικών n^* του υποψήφιου συνόλου καθορίζεται εντός του Ω. Μια σύνοψη της διαδικασίας παρουσιάζεται στον πίνακα 2.

Πίνακας 2: Επιλογή Υποψήφιου Συνόλου Χαρακτηριστικών

1. Επίλεξε n διαδοχικά χαρακτηριστικά από το σύνολο Y χρησιμοποιώντας το κριτήριο (n ένας μεγάλος προκαθορισμένος αριθμός)
$$\max_{x_j \in Y - X_{m-1}} \left[I(x_j ; c) - \frac{1}{m-1} \sum_{x_i \in X_{m-1}} I(x_j ; x_i) \right]$$
Έτσι προκύπτουν n διαδοχικά σύνολα χαρακτηριστικών $X_1 \subset X_2 \subset \cdots \subset X_{n-1} \subset X_n$.

2. Μεταξύ των n διαδοχικών συνόλων χαρακτηριστικών $X_1, \dots, X_k, \dots, X_n$ βρες το εύρος Ω του συνόλου k $(1 \le k \le n)$ για το οποίο το αντίστοιχο σφάλμα κατηγοριοποίησης e_k είναι σταθερά μικρό (π.χ. έχει και μικρή μέση τιμή και μικρή διακύμανση)

3. Εντός του Ω, βρες το μικρότερο σφάλμα κατηγοριοποίησης $e^* = \min e_k$.
Επίλεξε το βέλτιστο μέγεθος του υποψήφιου συνόλου n^* ως το μικρότερο k που αντιστοιχεί στο e^*.

4. ΑΞΙΟΛΟΓΗΣΗ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ

Η κατηγοριοποίηση των προτύπων αξιολογείται με τη βοήθεια διαφόρων στατιστκών μέτρων κάνοντας χρήση των πραγματικών ετικετών των διανυσμάτων εισόδου και της κλάσης εξόδου που υπολογίζει ο κατηγοριοποιητής. Τα μέτρα που θα χρησιμοποιηθούν είναι η *ακρίβεια (accuracy)*, η *ευαισθησία (sensitivity)* και η *προσδιοριστικότητα (specificity)* τα οποία αναφέρονται σε δυαδική κατηγοριοποίηση, δηλαδή κατηγοριοποίηση όπου τα πρότυπα κατατάσσονται σε δύο κλάσεις 0 και 1. Με τα μέτρα αυτά σχετίζεται επίσης και ο *πίνακας σύγχυσης (confusion matrix)* ο οποίος συνιστά μια οπτικοποίηση του αποτελέσματος της ταξινόμησης. Επιπλέον, θα παρουσιαστεί και η *Χαρακτηριστική Καμπύλη Λειτουργίας (ROC curve)* η οποία σχεδιάζεται με βάση την ευαισθησία και την προσδιοριστικότητα και τον τρόπο που αυτές μεταβάλλονται αν μεταβληθεί κάποια παράμετρος της κατηγοριοποίησης.

Χρήσιμες έννοιες για την κατανόηση των παραπάνω μέτρων είναι οι όροι ορθό θετικό, λανθασμένο θετικό, ορθό αρνητικό και λανθασμένο αρνητικό. Ο όρος *ορθό θετικό (true positive)* αναφέρεται στα πρότυπα τα οποία είναι γνωστό από τα δεδομένα εισόδου ότι ανήκουν στην κλάση 1 αλλά και ο κατηγοριοποιητής τα κατατάσσει επίσης στην κλάση 1. Πρόκειται δηλαδή για μία περίπτωση ενός ατόμου που είναι ασθενής και η κατηγοριοποίηση αποφαίνεται ότι είναι ασθενής επίσης. Αντίστοιχα ορίζονται και οι όροι *λανθασμένο θετικό (false positive)* για τα πρότυπα που ο κατηγοριοποιητής κατατάσσει στην κατηγορία 1 ενώ στην πραγματικότητα ανήκουν στην κλάση 0, *Ορθό Αρνητικό (true negative)* που αναφέρεται σε πρότυπα που ανήκουν στην κλάση 0 και από τον κατηγοριοποιητή κατατάσσονται επίσης στην κλάση αυτή και, τέλος, *λανθασμένο αρνητικό (false negative)* για τα πρότυπα τα οποία που ανήκουν στην κλάση 1 αλλά κατατάσσονται λανθασμένα στην κλάση 0.

Ευαισθησία (Sensitivity)

Η ευαισθησία εκφράζει την ικανότητα του κατηγοριοποιητή να εντοπίζει σωστά τα πρότυπα της κλάσης 1 καθώς ορίζεται το ποσοστό των ορθών θετικών στο σύνολο των προτύπων που ανήκουν στην κλάση 1 ως εξής

$$\text{ευαισθησία} = \frac{\#\ \text{ορθών θετικών}}{\#\ \text{ορθών θετικών} + \#\ \text{λανθασμένων αρνητικών}}$$

Υψηλή τιμή ευαισθησίας δηλώνει ότι αν ένα πρότυπο ταξινομηθεί στην κλάση 0 τότε υπάρχει μεγάλη πιθανότητα να είναι ταξινομημένο σωστά. Στην περίπτωση δεδομένων από μία εξέταση, η υψηλή τιμή ευαισθησίας υποδηλώνει ότι αν ο κατηγοριοποιητής αποφανθεί ότι ο ασθενής είναι υγιής η πιθανότητα να είναι και στην πραγματικότητα υγιής είναι μεγάλη.

Προσδιοριστικότητα (Specificity)

Η προσδιοριστικότητα εκφράζει την ικανότητα του κατηγοριοποιητή να εντοπίζει σωστά τα πρότυπα της κλάσης 0. Ορίζεται ως το ποσοστό των ορθών αρνητικών στο σύνολο των προτύπων που ανήκουν στην κλάση 0 ως εξής

$$\text{προσδιοριστικότητα} = \frac{\#\ \text{ορθών αρνητικών}}{\#\ \text{ορθών αρνητικών} + \#\ \text{λανθασμένων θετικών}}$$

Υψηλή τιμή προσδιοριστικότητας δηλώνει ότι αν ένα πρότυπο ταξινομηθεί στην κλάση 1 τότε υπάρχει μεγάλη πιθανότητα να είναι ταξινομημένο σωστά. Στην περίπτωση δεδομένων από μία εξέταση, η υψηλή τιμή ευαισθησίας υποδηλώνει ότι αν ο κατηγοριοποιητής αποφανθεί ότι ο ασθενής είναι ασθενής η πιθανότητα να είναι και στην πραγματικότητα ασθενής είναι μεγάλη.

Ακρίβεια (Accuracy)

Η ακρίβεια αποτελεί το μέτρο που εκφράζει το ποσοστό των σωστών ταξινομήσεων που εκτελεί ο κατηγοριοποιητής ως προς το σύνολο των ταξινομήσεων, ορίζεται δηλαδή ως

$$\text{ακρίβεια} = \frac{\#\ \text{ορθών θετικών} + \#\ \text{ορθών αρνητικών}}{\#\ \text{ορθών θετικών} + \#\ \text{ορθών αρνητικών} + \#\ \text{λανθ. θετικών} + \#\ \text{λανθ. αρνητικών}}$$

Η ακρίβεια είναι ουσιαστικά το συνολικότερο μέτρο απόδοσης ενός κατηγοριοποιητή καθώς όσο μεγαλύτερη είναι η ακρίβεια τόσο μεγαλύτερη είναι η πιθανότητα να καταταχθούν τα πρότυπα στη σωστή κλάση.

Πίνακας Σύγχυσης (Confusion Matrix)

Ο Πίνακας Σύγχυσης ορίζεται κατά βάση για τις μεθόδους επιβλεπόμενης μάθησης, εύκολα όμως η χρήση του μπορεί να επεκταθεί και στην περίπτωση της μη επιβλεπόμενης μάθησης. Πρόκειται για έναν πίνακα με 2 γραμμές και 2 στήλες ο οποίος στη θέση (1,1) περιλαμβάνει το πλήθος των ορθών θετικών, στη θέση (1,2) το πλήθος των λανθασμένων αρνητικών, στη θέση (2,1) το πλήθος των λανθασμένων θετικών και στη θέση (2,2) το πλήθος των ορθών αρνητικών. Έχει δηλαδή την παρακάτω μορφή:

# ορθών θετικών	# λανθασμένων αρνητικών
# λανθασμένων θετικών	# ορθών αρνητικών

Έχοντας την εικόνα του πίνακα σύγχυσης είναι δυνατή μία εποπτική αξιολόγηση του κατηγοριοποιητή καθώς όσο μεγαλύτερο είναι το άθροισμα της διαγωνίου του πίνακα τόσο καλύτερη είναι και η απόδοση του κατηγοριοποιητή. Παρατηρώντας τους ορισμούς της

ακρίβειας, της ευαισθησίας και της προσδιοριστικότητας είναι εμφανές ότι αν διατίθεται ο πίνακας σύγχυσης τα μέτρα αυτά υπολογίζονται εύκολα.

Χαρακτηριστική Καμπύλη Λειτουργίας (ROC curve)

Στις περισσότερες περιπτώσεις δεν είναι δυνατόν να επιτευχθεί μεγάλη ακρίβεια στην απόδοση του κατηγοριοποιητή και κατά συνέπεια ο στόχος της κατηγοριοποίησης μεταφέρεται στην επίτευξη καλής επίδοσης στην ανίχνευση της μίας από τις δύο κλάσεις, δηλαδή είτε υψηλής τιμής ευαισθησίας είτε υψηλής τιμής προσδιοριστικότητας. Συνήθως τα δύο αυτά μεγέθη είναι αλληλοσυγκρουόμενα και επομένως χρειάζεται να βρεθεί ένα σημείο συμβιβασμού. Τη συσχέτιση αυτή των δύο μεγεθών απεικονίζει η χαρακτηριστική καμπύλη λειτουργίας. Η καμπύλη ROC αποτελεί τη γραφική παράσταση του ρυθμού ορθών θετικών, δηλαδή της ευαισθησίας, συναρτήσει του ρυθμού λανθασμένων θετικών, δηλαδή του μεγέθους 1-προσδιοριστικότητα. Από την καμπύλη ROC ενός κατηγοριοποιητή η οποία αποτελείται από διάφορους συνδυασμούς ευαισθησίας και προσδιοριστικότητας για διάφορες τιμές κάποιας παραμέτρου του κατηγοριοποιητή είναι δυνατό να επιλεγεί το σημείο εκείνο με τις επιθυμητές τιμές και ο κατηγοριοποιητής να έχει την κατάλληλη συμπεριφορά.

5. ΕΦΑΡΜΟΓΗ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ ΣΕ ΙΑΤΡΙΚΑ ΔΕΔΟΜΕΝΑ

Στο σημείο αυτό, θα ελεγχθεί και θα παρουσιαστεί η απόδοση πέντε κατηγοριοποιητών, BK, SVM, kNN, SOM και FCM, σε έξι διαφορετικά σύνολα δεδομένων που σχετίζονται με κάποια ασθένεια ή διαγνωστική εξέταση. Τα σύνολα δεδομένων προέρχονται από τη βάση δεδομένων UCI για χρήση σε εφαρμογές μηχανικής μάθησης. Σε όλες τις περιπτώσεις τα διανύσματα είναι χωρισμένα σε δύο κλάσεις, την κλάση 0 για το υγιές άτομο και την κλάση 1 για τον ασθενή.

Αρχικά χρησιμοποιώντας τον αλγόριθμο SFFS πραγματοποιείται επιλογή χαρακτηριστικών με σκοπό να προκύψει ένα σύνολο δεδομένων με λιγότερα χαρακτηριστικά ώστε να μειώνεται ο υπολογιστικός φόρτος αλλά και να αποκλείονται χαρακτηριστικά που εμποδίζουν τη σωστή κατηγοριοποίηση. Ως κριτήριο για την επιλογή των χαρακτηριστικών επιλέγεται η ακρίβεια του κατηγοριοποιητή FCM καθώς ο αλγόριθμος FCM προσφέρει αρκετά γρήγορη κατηγοριοποίηση με δεδομένο το μεγάλο υπολογιστικό φορτίο που συνεπάγεται η επιλογή χαρακτηριστικών. Για το σκοπό αυτό χρησιμοποιείται το 30% των διαθέσιμων προτύπων. Στη συνέχεια, εξετάζεται η απόδοση των κατηγοριοποιητών κάνοντας χρήση πρώτα όλων των χαρακτηριστικών του αρχικού συνόλου δεδομένων ενώ στη συνέχεια επαναλαμβάνεται η κατηγοριοποίηση κάνοντας χρήση αυτή τη φορά μόνο των χαρακτηριστικών που προέκυψαν από την επιλογή χαρακτηριστικών. Και στις δύο περιπτώσεις λαμβάνονται η ακρίβεια, η ευαισθησία, η προσδιοριστικότητα, ο πίνακας σύγχυσης και η καμπύλη ROC.

Στις περιπτώσεις όπου έχουμε επιβλεπόμενη μάθηση, δηλαδή στους αλγορίθμους SVM, kNN και BK, η κατηγοριοποίηση πραγματοποιείται 10 φορές σε κάθε σύνολο δεδομένων με διαφορετικό σύνολο εκπαίδευσης κάθε φορά προκειμένου να ληφθεί αντιπροσωπευτικότερο αποτέλεσμα. Το σύνολο εκπαίδευσης κυμαίνεται στο 30-40% των διαθέσιμων προτύπων ανάλογα με το πλήθος τους ώστε να είναι όσο το δυνατόν πιο αξιόπιστα τα αποτελέσματα του ελέγχου που ακολουθεί. Σ' αυτή την περίπτωση τα στατιστικά μέτρα προκύπτουν ως η μέση τιμή των 10 επαναλήψεων, ενώ υπολογίζεται και η τυπική τους απόκλιση. Ο πίνακας σύγχυσης και η καμπύλη ROC που παρουσιάζονται αντιστοιχούν σε μία μέση περίπτωση.

Όσον αφορά τις λεπτομέρειες στην εφαρμογή του κάθε αλγορίθμου, για την εκτέλεση του αλγορίθμου Οπισθοδιάδοσης (BK) χρησιμοποιείται νευρωνικό δίκτυο δύο επιπέδων. Οι είσοδοί του είναι όσες τα χαρακτηριστικά του υπό κατηγοριοποίηση του συνόλου προτύπων, το κρυφό επίπεδο περιλαμβάνει 10 νευρώνες και οι έξοδοι είναι δύο οι οποίες ενεργοποιούνται όταν το πρότυπο ταξινομείται στην αντίστοιχη κλάση. Ως συνάρτηση ενεργοποίησης στο κρυφό επίπεδο χρησιμοποιείται η σιγμοειδής συνάρτηση ενώ για το επίπεδο εξόδου, χρησιμοποιείται γραμμική συνάρτηση. Για την εκπαίδευση χρησιμοποιείται η μέθοδος της βαθμωτής κατάβασης (gradient descent) με παράμτερο εκπαίδευσης 0.05. Το 15% των δεδομένων χρησιμοποιείται για επαλήθευση σε περίπτωση που ο αλγόριθμος δεν συγκλίνει. Κατά την εφαρμογή της Μηχανής Διανυσμάτων Υποστήριξης (SVM) επιλέγεται ως συνάρτηση πυρήνα η καλύτερη μεταξύ του γραμμικού

πυρήνα (dot product) και του πολυωνυμικού πυρήνα 3^{ης} τάξης. Για τη μέθοδο προσδιορισμού του διαχωριστκού υπερεπιπέδου χρησιμοποιείται ελαστικό περιθώριο διαχωρισμού δεδομένου ότι οι κλάσεις σπάνια είναι γραμμικά διαχωρίσιμες. Επίσης, ο κατηγοριοποιητής k Κοντινότερων Γειτόνων (kNN) εξετάζονται 3 εκδοχές του, οι 3NN, 5NN και 7NN όπου για την ταξινόμηση χρησιμοποιούνται οι 3,5 και 7 κοντινότεροι γείτονες αντίστοιχα. Τέλος, όσον αφορά στη χρήση των Χαρτών Αυτο-Οργάνωσης, δημιουργείται ένα νευρωνικό δίκτυο 5x5 εξαγωνικών νευρώνων, δηλαδή 25 εξόδων, του οποίου οι είσοδοι είναι όσα και τα χαρακτηριστικά των προτύπων που ταξινομούνται.

5.1. Διάγνωση Αρρυθμίας

Το διαθέσιμο σύνολο δεδομένων περιλαμβάνει 395 διανύσματα με 134 χαρακτηριστικά. Μετά την εφαρμογή της μεθόδου SFFS για την επιλογή χαρακτηριστικών τα χαρακτηριστικά μειώνονται σε 15.

5.1.1. Κατηγοριοποιητής BK

- Με όλα τα χαρακτηριστικά:

Πίνακας Σύγχυσης:

210	67
19	99

Ακρίβεια : 0.7506 ± 0.0344
Ευαισθησία : 0.8864 ± 0.0455
Προσδιοριστικότητα : 0.5636 ± 0.0371

- Μετά την επιλογή χαρακτηριστικών:

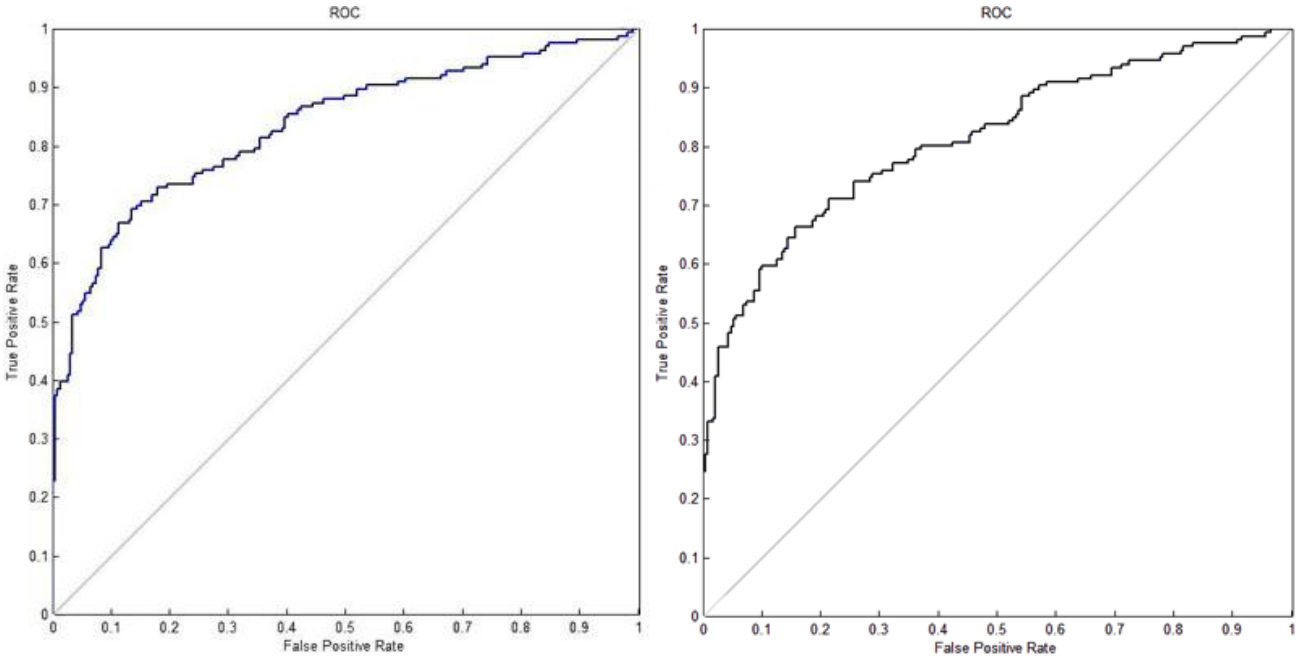
Πίνακας Σύγχυσης:

217	82
12	84

Ακρίβεια : 0.7518 ± 0.0143
Ευαισθησία : 0.9010 ± 0.0330
Προσδιοριστικότητα : 0.5456 ± 0.0286

Σύμφωνα με τα παραπάνω στατιστικά δεδομένα και το σχήμα 1, ο αλγόριθμος BK έχει ικανοποιητική ακρίβεια. Η ακρίβεια όμως αυτή οφείλεται στη μεγάλη ευαισθησία, δηλαδή στην δυνατότητα του κατηγοριοποιητή να διακρίνει τις υγιείς περιπτώσεις. Η ευαισθησία αυτή αυξάνεται μάλιστα μετά την επιλογή χαρακτηριστικών, με αντίστοιχη μείωση της

προσδιοριστικότητας, η οποία όμως ήταν ούτως ή άλλως αρκετά χαμηλή ώστε να μην είναι δυνατή η ασφαλής αναγνώριση των ασθενών. Η ακρίβεια μετά την επιλογή χαρακτηριστικών συμπαρασύρεται από τη μείωση της προσδιοριστικότητας όπως καταδεικνύεται και από τις καμπύλες ROC.



Σχήμα 1: Καμπύλη ROC του κατηγοριοποιητή ΒΚ για τη διάγνωση αρρυθμίας, αριστερά με χρήση όλων των χαρακτηριστικών και δεξιά μετά την επιλογή χαρακτηριστικών.

5.1.2. Κατηγοριοποιητής SVM

Για την κατηγοριοποίηση χρησιμοποιήθηκε ο γραμμικός πυρήνας.

- Με όλα τα χαρακτηριστικά:

  Πίνακας Σύγχυσης:

82	31
32	52

Ακρίβεια : 0.6721 ± 0.0334
Ευαισθησία : 0.7325 ± 0.0667
Προσδιοριστικότητα : 0.5892 ± 0.0552

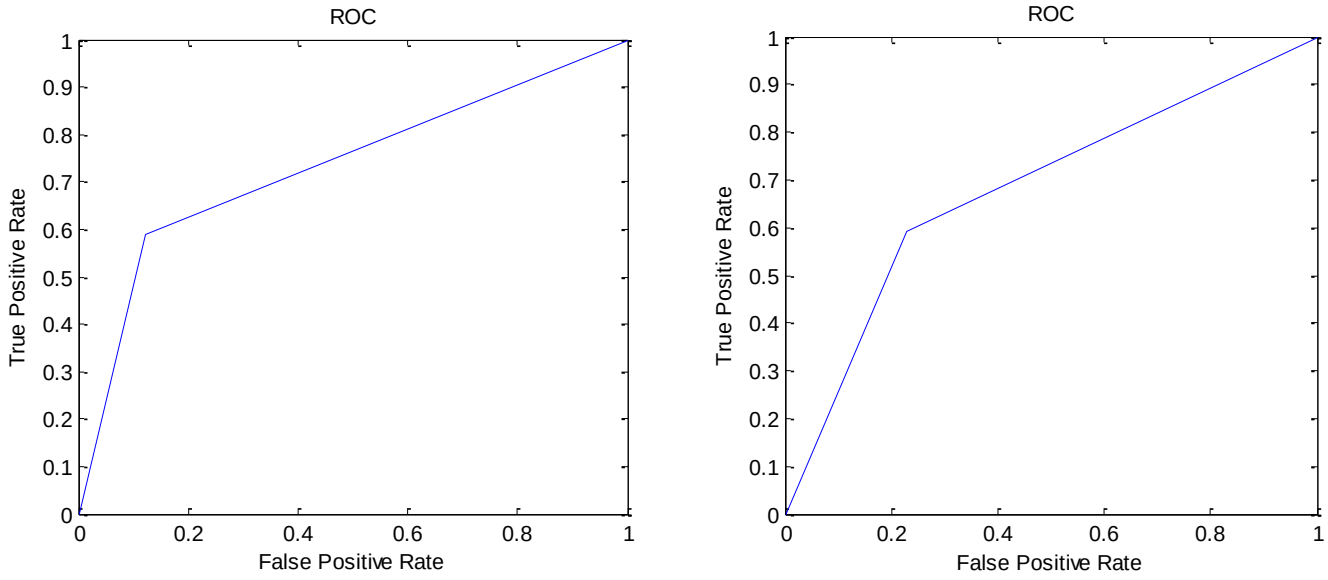
- Μετά την επιλογή χαρακτηριστικών:

  Πίνακας Σύγχυσης:

101	33

13	50

Ακρίβεια : 0.7548 ± 0.0303
Ευαισθησία : 0.8860 ± 0.0336
Προσδιοριστικότητα : 0.5747 ± 0.0848



Σχήμα 2: Καμπύλη ROC του κατηγοριοποιητή SVM για τη διάγνωση αρρυθμίας, αριστερά με χρήση όλων των χαρακτηριστικών και δεξιά μετά την επιλογή χαρακτηριστικών.

Στην περίπτωση αυτή η απόδοση της κατηγοριοποίησης είναι μέτρια καθώς η ακρίβεια κινείται σε χαμηλά επίπεδα με κάποια βελτίωση μετά την επιλογή χαρακτηριστικών όπως φαίνεται και από τις καμπύλες ROC του σχήματος 2. Η προσδιοριστικότητα είναι κακή, η ευαισθησία όμως είναι αρκετά καλή και ιδιαίτερα μετά την επιλογή χαρακτηριστικών.

5.1.3. Κατηγοριοποιητής kNN

5.1.3.1. Κατηγοριοποιητής 3NN

– Με όλα τα χαρακτηριστικά:
Πίνακας Σύγχυσης:

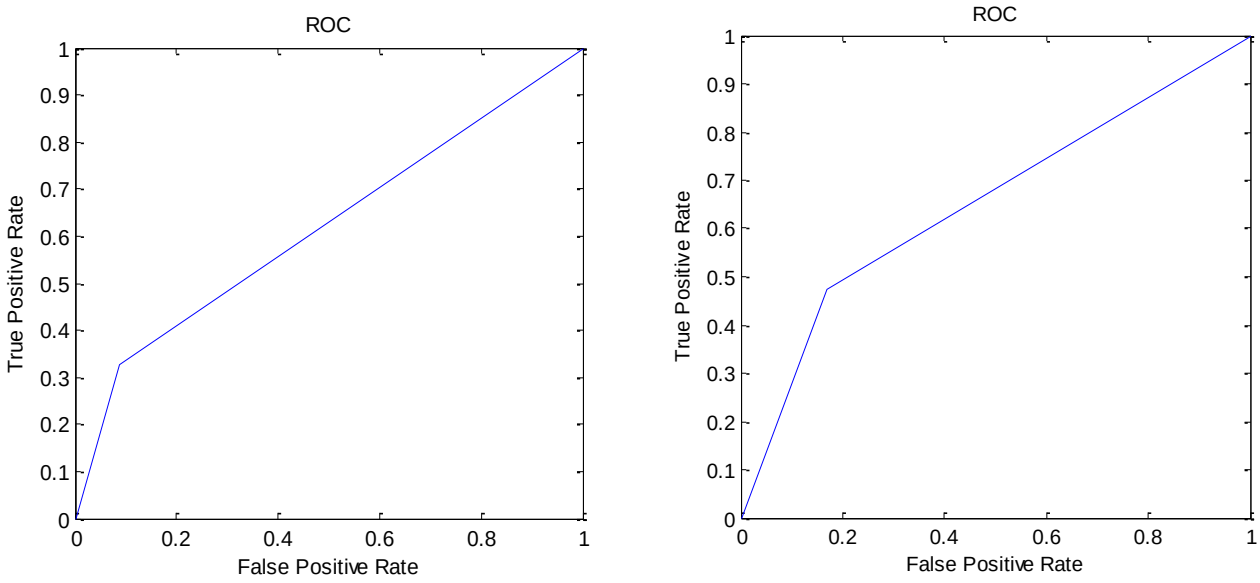
145	80
15	36

Ακρίβεια : 0.6489 ± 0.0173
Ευαισθησία : 0.8825 ± 0.3267
Προσδιοριστικότητα : 0.5892 ± 0.0252

– Μετά την επιλογή χαρακτηριστικών:
Πίνακας Σύγχυσης:

144	67
16	49

Ακρίβεια : 0.6953 ± 0.0204
Ευαισθησία : 0.8625 ± 0.0276
Προσδιοριστικότητα : 0.4647 ± 0.0521



Σχήμα 3: Καμπύλη ROC του κατηγοριοποιητή 3NN για τη διάγνωση αρρυθμίας, αριστερά με χρήση όλων των χαρακτηριστικών και δεξιά μετά την επιλογή χαρακτηριστικών.

Ο κατηγοριοποιητής 3NN έχει, επομένως, μη ικανοποιητική απόδοση παρά τη μικρή αύξηση της ακρίβειας μετά την επιλογή χαρακτηριστικών (βλ. σχήμα 3). Στον αντίποδα της μικρής προσδιοριστικότητας βρίσκεται η αρκετά ικανοποιητική ευαισθησία η οποία όμως είναι καλύτερη όταν χρησιμοποιούνται όλα τα χαρακτηριστικά.

5.1.3.2. Κατηγοριοποιητής 5NN

– Με όλα τα χαρακτηριστικά:
Πίνακας Σύγχυσης:

143	78
17	38

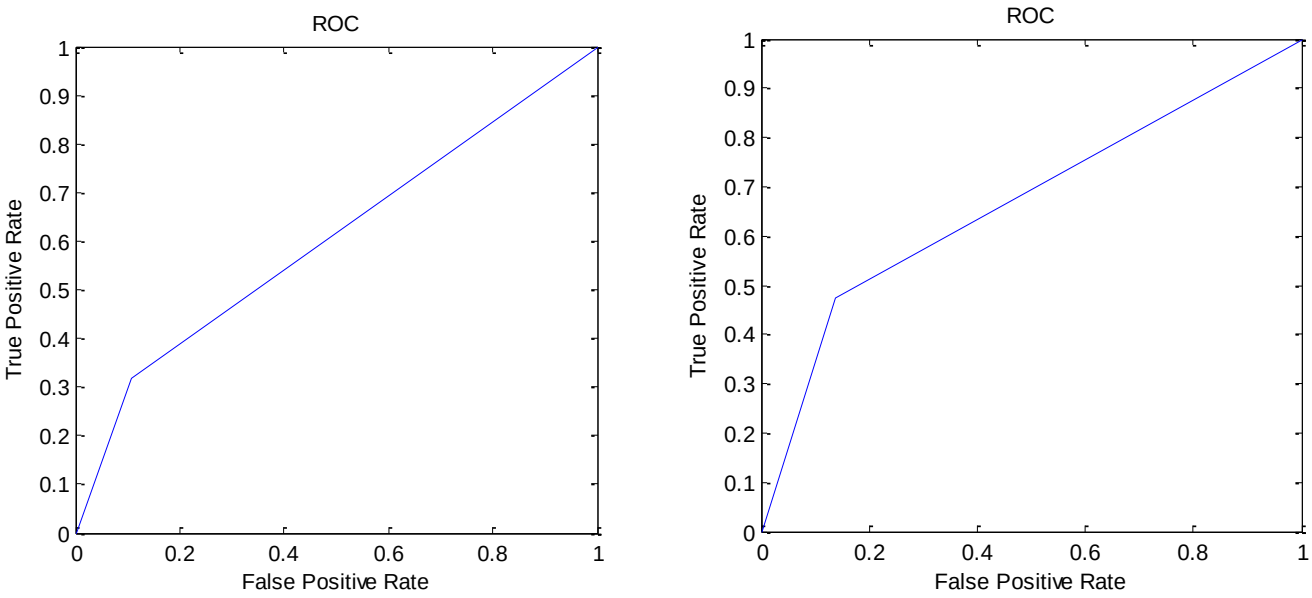
Ακρίβεια : 0.6543 ± 0.0227

Ευαισθησία : 0.8919 ± 0.0340
Προσδιοριστικότητα : 0.3267 ± 0.0405

– Μετά την επιλογή χαρακτηριστικών:
Πίνακας Σύγχυσης:

136	58
24	58

Ακρίβεια : 0.7036 ± 0.0247
Ευαισθησία : 0.8913 ± 0.0354
Προσδιοριστικότητα : 0.4448 ± 0.0623



Σχήμα 4: Καμπύλη ROC του κατηγοριοποιητή 5NN για τη διάγνωση αρρυθμίας, αριστερά με χρήση όλων των χαρακτηριστικών και δεξιά μετά την επιλογή χαρακτηριστικών.

Ο κατηγοριοποιητής 5NN έχει αποκτά ικανοποιητική ακρίβεια μετά την επιλογή χαρακτηριστικών (βλ. σχήμα 4), η οποία όμως οφείλεται στην μεγάλη ευαισθησία αφού η προσδιοριστικότητα είναι κακή τόσο με χρήση όλων των χαρακτηριστικών όσο και μετά την επιλογή χαρακτηριστικών. Επομένως, η χρήση του θα είχε πρακτικά εφαρμογή στην ανίχνευση των υγιών περιπτώσεων, αφού ένα αρνητικό αποτέλεσμα είναι με καλή πιθανότητα στην πραγματικότητα αρνητικό αφού τα πραγματικά θετικά ταξινομούνται με μεγάλη πιθανότητα ως θετικά.

5.1.3.3. Κατηγοριοποιητής 7NN

– Με όλα τα χαρακτηριστικά:
Πίνακας Σύγχυσης:

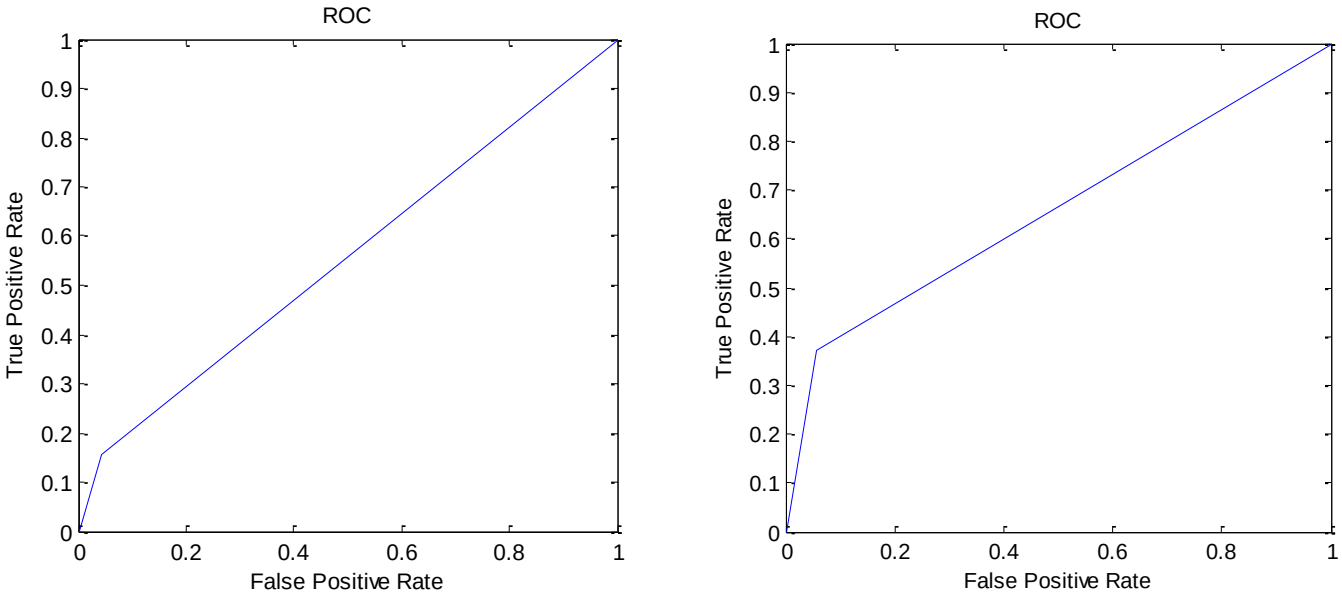
152	95
8	21

Ακρίβεια : 0.6217 ± 0.0144
Ευαισθησία : 0.9569 ± 0.0225
Προσδιοριστικότητα : 0.1595 ± 0.0474

– Μετά την επιλογή χαρακτηριστικών:
Πίνακας Σύγχυσης:

147	70
13	46

Ακρίβεια : 0.6929 ± 0.0100
Ευαισθησία : 0.9206 ± 0.0406
Προσδιοριστικότητα : 0.3767 ± 0.0491



Σχήμα 5: Καμπύλη ROC του κατηγοριοποιητή 7NN για τη διάγνωση αρρυθμίας, αριστερά με χρήση όλων των χαρακτηριστικών και δεξιά μετά την επιλογή χαρακτηριστικών.

Η ακρίβεια του κατηγοριοποιητή αυτού είναι τέτοια που δεν επιτρέπει τη χρήση του για ταξινόμηση όλων των προτύπων είτε πριν είτε μετά την επιλογή χαρακτηριστικών παρά τη μικρή βελτίωση που παρουσιάζεται όπως φαίνεται και στο σχήμα 5. Η μεγάλη τιμή ευαισθησίας που επιτυγχάνεται όμως, ειδικά με τη χρήση όλων των χαρακτηριστικών καθιστά δυνατή την καλή διάκριση των υγιών περιπτώσεων. Σ'αυτή την περίπτωση η επιλογή χαρακτηριστικών αυξάνει μεν την ακρίβεια του κατηγοριοποιητή όμως όχι σε

ικανοποιητικό βαθμό ενώ μειώνει και την ευαισθησία οπότε προτιμάται η κατηγοριοποίηση με όλα τα χαρακτηριστικά.

5.1.4. Κατηγοριοποιητής SOM

– Με όλα τα χαρακτηριστικά:
Πίνακας Σύγχυσης:

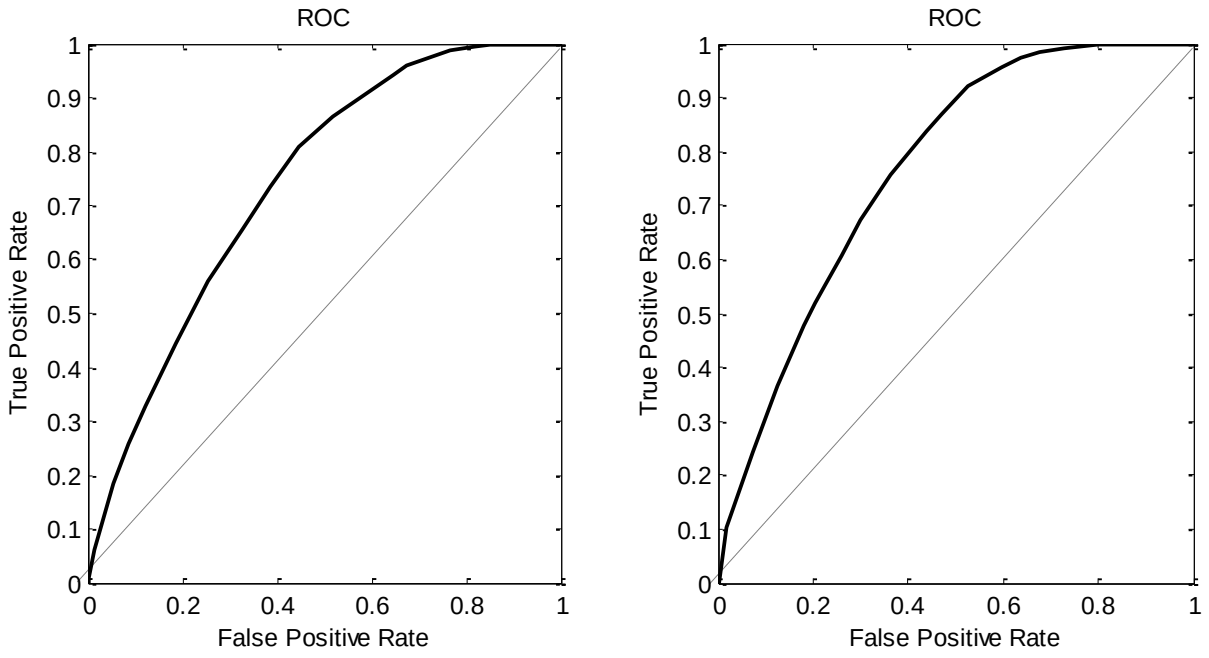
209	100
20	66

Ακρίβεια : 0.6962
Ευαισθησία : 0.9127
Προσδιοριστικότητα : 0.3976

– Μετά την επιλογή χαρακτηριστικών:
Πίνακας Σύγχυσης:

208	84
21	82

Ακρίβεια : 0.7342
Ευαισθησία : 0.9083
Προσδιοριστικότητα : 0.4940



Σχήμα 6: Καμπύλη ROC του κατηγοριοποιητή SOM για τη διάγνωση αρρυθμίας, αριστερά με χρήση όλων των χαρακτηριστικών και δεξιά μετά την επιλογή χαρακτηριστικών.

Στην περίπτωση του κατηγοριοποιητή SOM επιτυγχάνεται μεγάλη τιμή ευαισθησίας αλλά πολύ μικρή τιμή προσδιοριστικότητας και επομένως μέτρια ακρίβεια. Η προσδιοριστικότητα αυξάνεται μετά την επιλογή των χαρακτηριστικών όχι όμως σε επίπεδο που να επιτρέπει την ταξινόμηση των υγιών περιπτώσεων στη σωστή κλάση. Η αύξηση αυτή συμπαρασύρει και την ακρίβεια, όπως φαίνεται στο σχήμα 6, όμως επέρχεται μικρή μείωση στην ευαισθησία οπότε προτιμάται η κατηγοριοποίηση με όλα τα χαρακτηριστικά.

5.1.5. Κατηγοριοποιητής FCM

- Με όλα τα χαρακτηριστικά:

Πίνακας Σύγχυσης:

96	79
133	87

Ακρίβεια : 0.4633
Ευαισθησία : 0.4192
Προσδιοριστικότητα : 0.5241

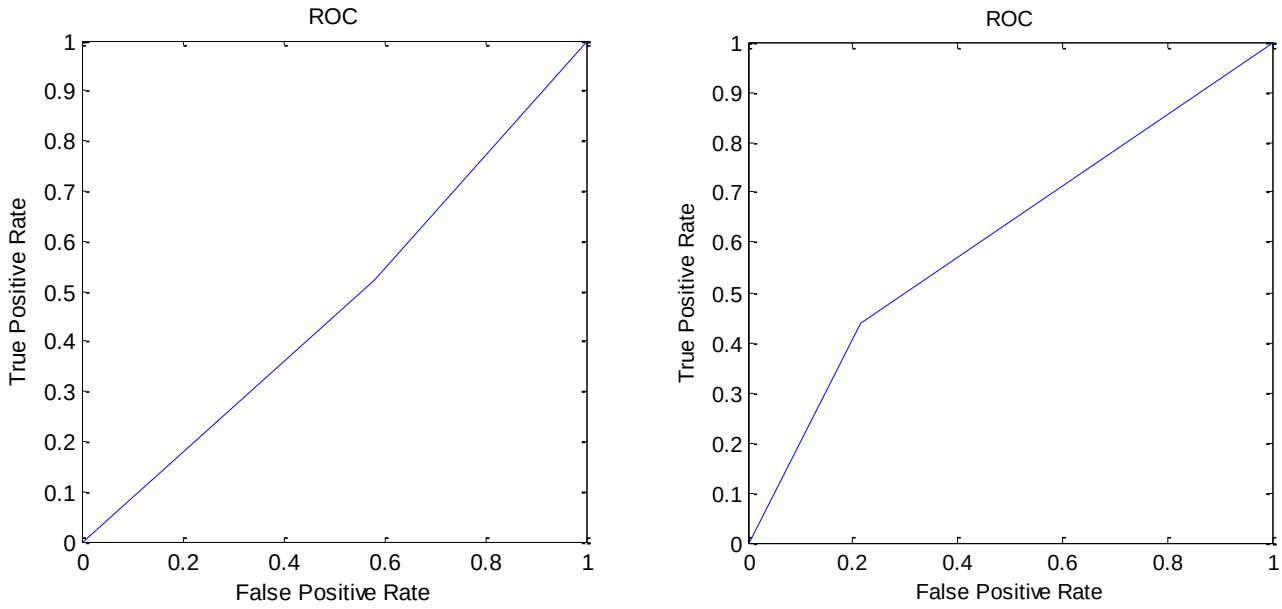
- Μετά την επιλογή χαρακτηριστικών:

Πίνακας Σύγχυσης:

180	93
49	73

Ακρίβεια : 0.6405
Ευαισθησία : 0.7860
Προσδιοριστικότητα : 0.4398

Στην περίπτωση του FCM η απόδοση του κατηγοριοποιητή δεν είναι καλή. Όπως φαίνεται και στο σχήμα 7, μετά την επιλογή χαρακτηριστικών ο κατηγιοριοποιητής ακολουθεί την τάση των άλλων κατηγοριοποιητών για αυξημένη ευαισθησία αλλά και οι γεικότερες επιδόσεις της κατηγοριοποίησης βελτιώνονται αισθητά όμως όχι σε βαθμό ώστε τα αποτελέσματα να είναι πρακτικά αξιοποιήσιμα.



Σχήμα 7: Καμπύλη ROC του κατηγοριοποιητή FCM για τη διάγνωση αρρυθμίας, αριστερά με χρήση όλων των χαρακτηριστικών και δεξιά μετά την επιλογή χαρακτηριστικών

5.1.6. Αποτίμηση Κατηγοριοποιητών

Κανένας από τους κατηγοριοποιητές δεν εμφανίζει ικανοποιητική προσδιοριστικότητα γεγονός που συνεπάγεται ότι κατά πάσα πιθανότητα πολλά από τα πρότυπα της κλάσης των υγιών βρίσκονται εντός της περιοχής των ασθενών η οποία είναι καλά καθορισμένη. Το ενδιαφέρον επομένως στρέφεται στην αξιολόγηση των κατηγοριοποιητών με βάση την ευαισθησία αλλά και τη συνολική ακρίβεια. Τη χειρότερη απόδοση έχει καθαρά η μέθοδος FCM, καθώς η ασαφής λογική σε συνδυασμό με την ιδιαιτερότητα του συνόλου δεδομένων φαίνεται πως δεν προσφέρονται για την επιτυχημένη κατηγοριοποίηση. Αμέσως καλύτερος κρίνεται ο κατηγοριοποιητής SVM ο οποίος φαίνεται ότι επηρεάζεται επίσης από την κακή διαχωρισιμότητα των κλάσεων αν και λιγότερο απ' ό,τι στον FCM. Πολύ καλή ευαισθησία αποδίδει ο αλγόριθμός SOM ενώ από τους κατηγοριοποιητές kNN καλύτερος κρίνεται ο 7NN καθώς η χρήση περισσότερων γειτόνων βελτιώνει τη κατηγοριοποίηση σ' αυτή την περίπτωση και παρουσιάζει την μεγαλύτερη ευαισθησία από όλους τους κατηγοριοποιητές που δοκιμάστηκαν. Όσον αφορά στον καλύτερο συνδυασμό ευαισθησίας και ακρίβειας αυτός επιτυγχάνεται με τη χρήση του αλγορίμθου BK.

5.2. Διάγνωση Καρκίνου του Στήθους

Το διαθέσιμο σύνολο δεδομένων περιλαμβάνει 683 διανύσματα με 9 χαρακτηριστικά, τα οποία μετά την επιλογή χαρακτηριστικών μπορούν να μειωθούν σε 7.

5.2.1. Κατηγοριοποιητής BK

– Με όλα τα χαρακτηριστικά:

Πίνακας Σύγχυσης:

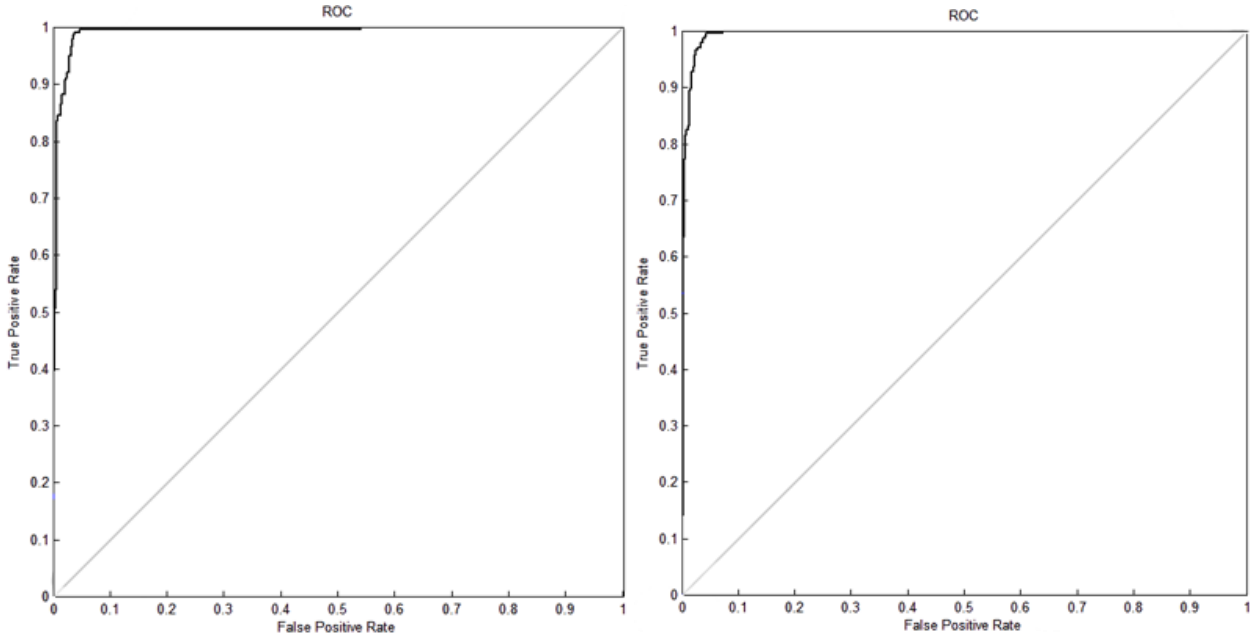
429	8
15	231

Ακρίβεια : 0.9672 ± 0.0046
Ευαισθησία : 0.9760 ± 0.0026
Προσδιοριστικότητα : 0.9508 ± 0.0138

– Μετά την επιλογή χαρακτηριστικών:
Πίνακας Σύγχυσης:

433	9
11	230

Ακρίβεια : 0.9642 ± 0.0049
Ευαισθησία : 0.9728 ± 0.0051
Προσδιοριστικότητα : 0.9458 ± 0.0074



Σχήμα 8: Καμπύλη ROC του κατηγοριοποιητή ΒΚ για τη διάγνωση καρκίνου του στήθους αριστερά με χρήση όλων των χαρακτηριστικών και δεξιά μετά την επιλογή χαρακτηριστικών.

Στην περίπτωση αυτή έχουμε εξαιρετικές επιδόσεις του κατηγοριοποιητή σε όλους τους δείκτες. Μετά την επιλογή χαρακτηριστικών παρατηρείται μικρή περεταίρω αύξηση της ευαισθησίας αλλά λόγω της μικρής μείωσης της προσδιοριστικότητας η ακρίβεια παραμένει σταθερή. Η πολύ καλή αυτή επίδοση καταδεικνύεται και από τις καμπύλες ROC του σχήματος 8 καθώς οι αιχμές τους προσεγγίζουν την πάνω αριστερή γωνία του διαγράμματος.

5.2.2. Κατηγοριοποιητής SVM

Για την κατηγοριοποίηση χρησιμοποιήθηκε γραμμική συνάρτηση πυρήνα.

– Με όλα τα χαρακτηριστικά:

Πίνακας Σύγχυσης:

216	5
6	114

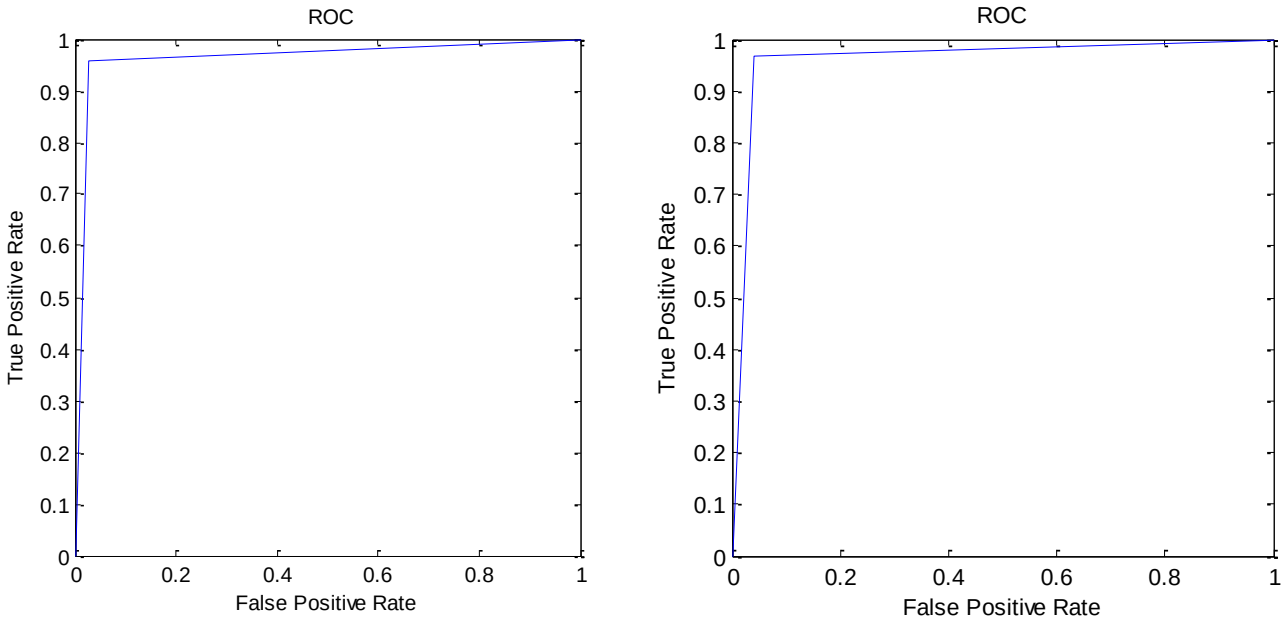
Ακρίβεια : 0.9674 ± 0.0040
Ευαισθησία : 0.9752 ± 0.0038
Προσδιοριστικότητα : 0.9529 ± 0.0127

– Μετά την επιλογή χαρακτηριστικών:

Πίνακας Σύγχυσης:

216	6
6	113

Ακρίβεια : 0.9669 ± 0.0052
Ευαισθησία : 0.9739 ± 0.0097
Προσδιοριστικότητα : 0.9538 ± 0.0082



Σχήμα 9: Καμπύλη ROC του κατηγοριοποιητή SVM για τη διάγνωση καρκίνου του στήθους αριστερά με χρήση όλων των χαρακτηριστικών και δεξιά μετά την επιλογή χαρακτηριστικών.

Η επιδόσεις του κατηγοριοποιητή SVM είναι εξαιρετικές και κυμαίνονται στα ίδια επίπεδα τόσο πριν όσο και μετά από την επιλογή χαρακτηριστικών, ενώ παρουσιάζεται και ιδιαίτερη σταθερότητα στους δείκτες καθώς η τυπική τους απόκλιση μεταξύ των διαφόρων τρεξιμάτων είναι μικρή. Οι καμπύλες ROC του σχήματος 9 επιβεβαιώνουν την καλή επίδοση αυτή.

5.2.3. Κατηγοριοποιητής kNN

5.2.3.1. Κατηγοριοποιητής 3NN

- Με όλα τα χαρακτηριστικά:

Πίνακας Σύγχυσης:

303	11
7	156

Ακρίβεια : 0.9619 ± 0.0045
Ευαισθησία : 0.9784 ± 0.0070
Προσδιοριστικότητα : 0.9305 ± 0.0186

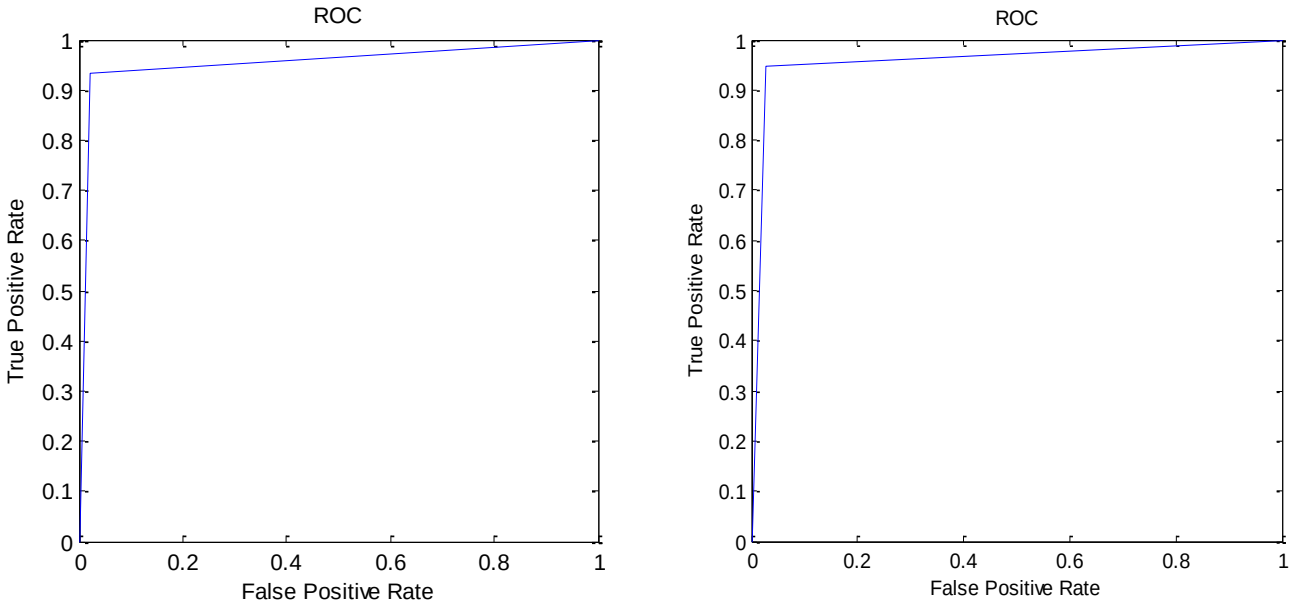
- Μετά την επιλογή χαρακτηριστικών:

Πίνακας Σύγχυσης:

302	9
8	158

Ακρίβεια : 0.9644 ± 0.0077
Ευαισθησία : 0.9710 ± 0.0076
Προσδιοριστικότητα : 0.9521 ± 0.0185

Τα αποτελέσματα της κατηγοριοποίησης είναι και σε αυτή την περίπτωση πάρα πολύ κοντά στην πραγματικότητα (βλ σχήμα 10). Η λίγο μικρότερη από τα άλλα μεγέθη προσδιοριστικότητα αυξάνεται μετά την επιλογή χαρακτηριστικών στο επίπεδο των άλλων δεικτών.



Σχήμα 10: Καμπύλη ROC του κατηγοριοποιητή 3NN για τη διάγνωση καρκίνου του στήθους αριστερά με χρήση όλων των χαρακτηριστικών και δεξιά μετά την επιλογή χαρακτηριστικών.

5.2.3.2. Κατηγοριοποιητής 5NN

– Με όλα τα χαρακτηριστικά:

Πίνακας Σύγχυσης:

301	6
9	161

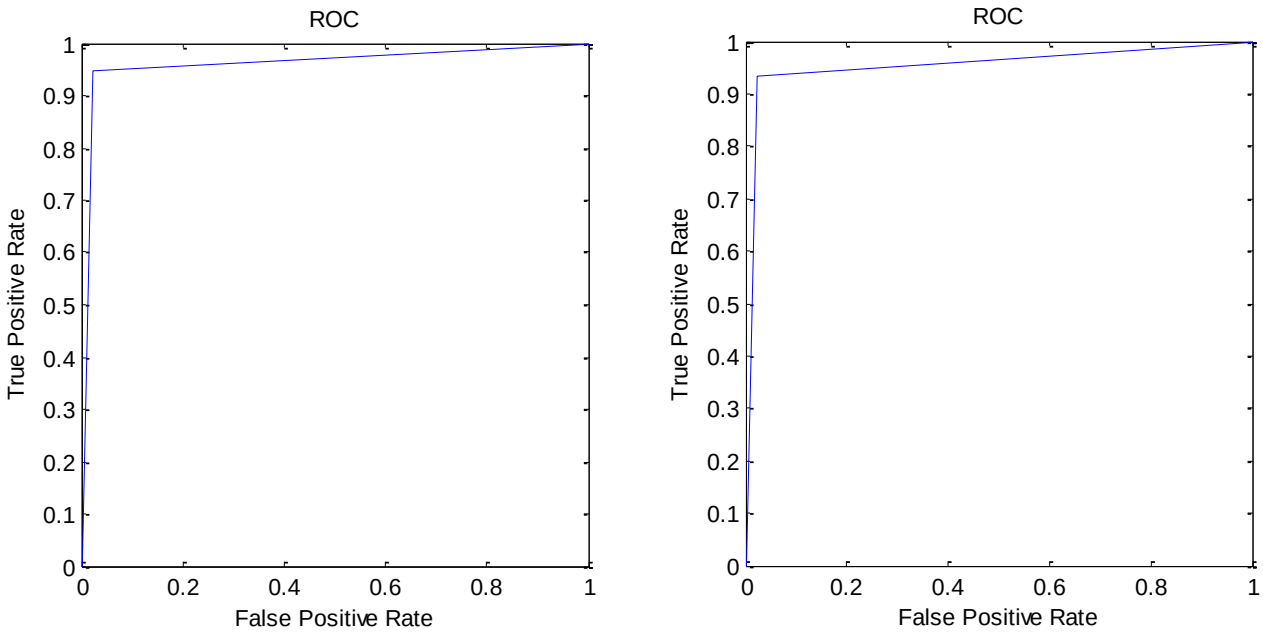
Ακρίβεια : 0.9671 ± 0.0072
Ευαισθησία : 0.9755 ± 0.0072
Προσδιοριστικότητα : 0.9515 ± 0.0253

– Μετά την επιλογή χαρακτηριστικών:

Πίνακας Σύγχυσης:

301	7
9	160

Ακρίβεια : 0.9681 ± 0.0079
Ευαισθησία : 0.9745 ± 0.0058
Προσδιοριστικότητα : 0.9563 ± 0.0217



Σχήμα 11: Καμπύλη ROC του κατηγοριοποιητή 5NN για τη διάγνωση καρκίνου του στήθους αριστερά με χρήση όλων των χαρακτηριστικών και δεξιά μετά την επιλογή χαρακτηριστικών

Και στην περίπτωση του κατηγοριοποιητή 5NN η κατηγοριοποίηση γίνεται με μεγάλη επιτυχία (βλ. σχήμα 11). Η επιλογή χαρακτηριστικών δεν επιφέρει καμία ουσιαστική διαφορά στην επίδοση του κατηγοριοποιητή.

5.2.3.3. Κατηγοριοποιητής 7NN

– Με όλα τα χαρακτηριστικά:
Πίνακας Σύγχυσης:

305	11
5	156

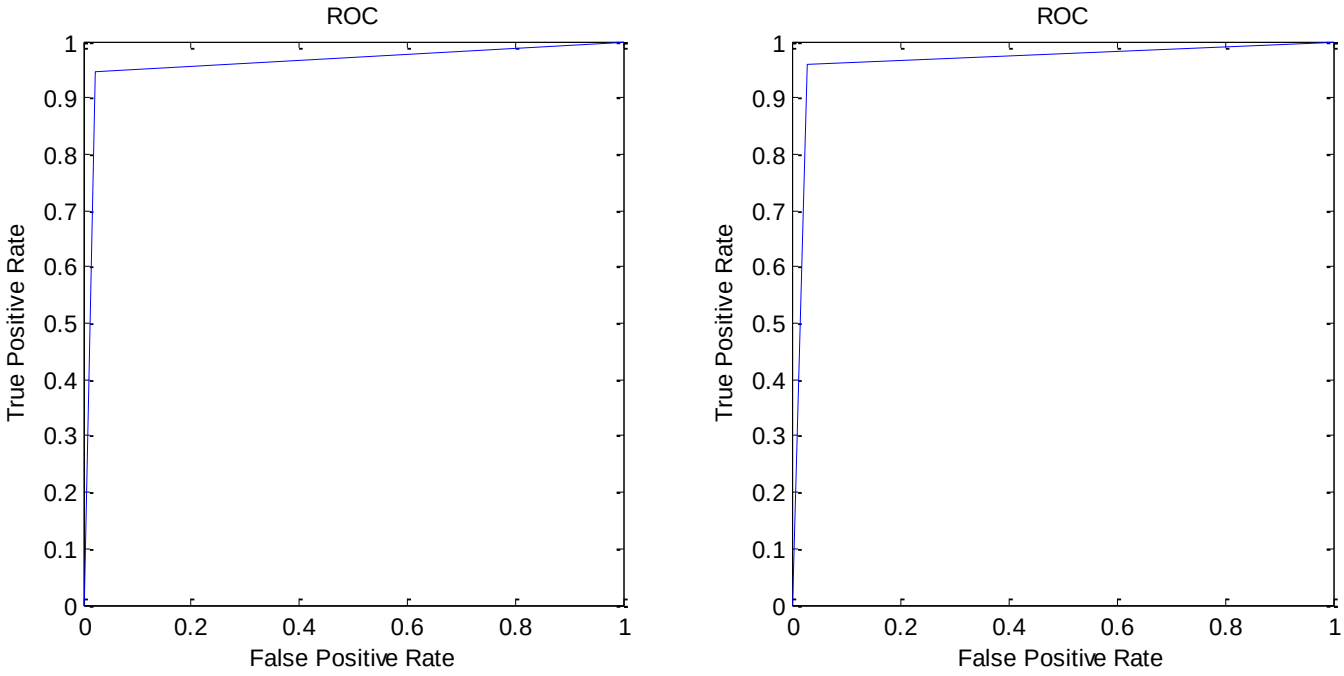
Ακρίβεια : 0.9637 ± 0.0075
Ευαισθησία : 0.9781 ± 0.0048
Προσδιοριστικότητα : 0.9371 ± 0.0224

– Μετά την επιλογή χαρακτηριστικών:
Πίνακας Σύγχυσης:

301	9
9	158

Ακρίβεια : 0.9652 ± 0.0081
Ευαισθησία : 0.9765 ± 0.0076

Προσδιοριστικότητα : 0.9443 ± 0.0259



Σχήμα 12: Καμπύλη ROC του κατηγοριοποιητή 7NN για τη διάγνωση καρκίνου του στήθους αριστερά με χρήση όλων των χαρακτηριστικών και δεξιά μετά την επιλογή χαρακτηριστικών

Και σε αυτή την περίπτωση ο κατηγοριοποιητής κατατάσσει τα πρότυπα στις σωστές κλάσεις με πολύ μεγάλη πιθανότητα, όπως καταδεικνύουν και οι καμπύλες ROC του σχήματος 12. Η επιλογή χαρακτηριστικών βελτιώνει λίγο την προσδιοριστικότητα.

5.2.4. Κατηγοριοποιητής SOM

– Με όλα τα χαρακτηριστικά:
Πίνακας Σύγχυσης:

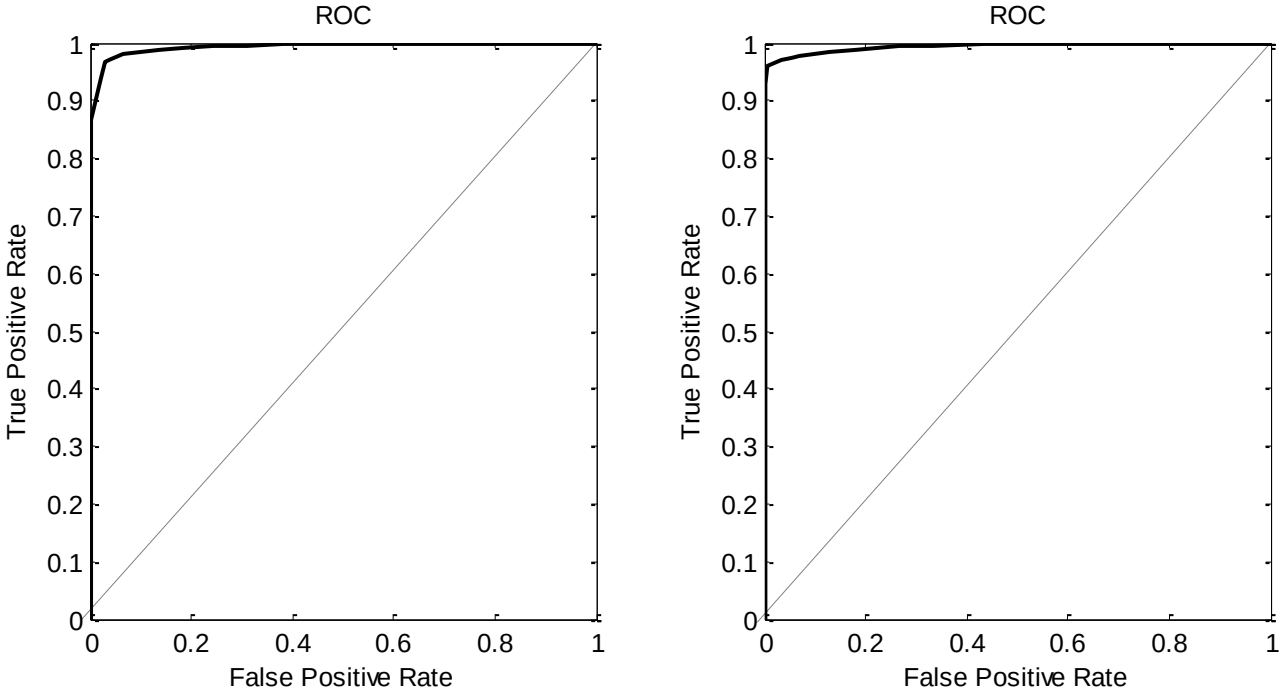
435	11
9	228

Ακρίβεια : 0.9707
Ευαισθησία : 0.9797
Προσδιοριστικότητα : 0.9540

– Μετά την επιλογή χαρακτηριστικών:
Πίνακας Σύγχυσης:

427	2
7	237

Ακρίβεια : 0.9722
Ευαισθησία : 0.9617
Προσδιοριστικότητα : 0.9916



Σχήμα 13: Καμπύλη ROC του κατηγοριοποιητή SOM για τη διάγνωση καρκίνου του στήθους αριστερά με χρήση όλων των χαρακτηριστικών και δεξιά μετά την επιλογή χαρακτηριστικών

Και στην περίπτωση της κατηγοριοποίησης με τη χρήση της μεθόδου SOM τα αποτελέσματα της κατηγοριοποίησης είναι πάρα πολύ καλά (βλ. σχήμα 13). Μετά την επιλογή χαρακτηριστικών μάλιστα η τιμή της προσδιοριστικότητα αγγίζει το τέλειο.

5.2.5. Κατηγοριοποιητής FCM

– Με όλα τα χαρακτηριστικά**:**
Πίνακας Σύγχυσης:

436	22
8	217

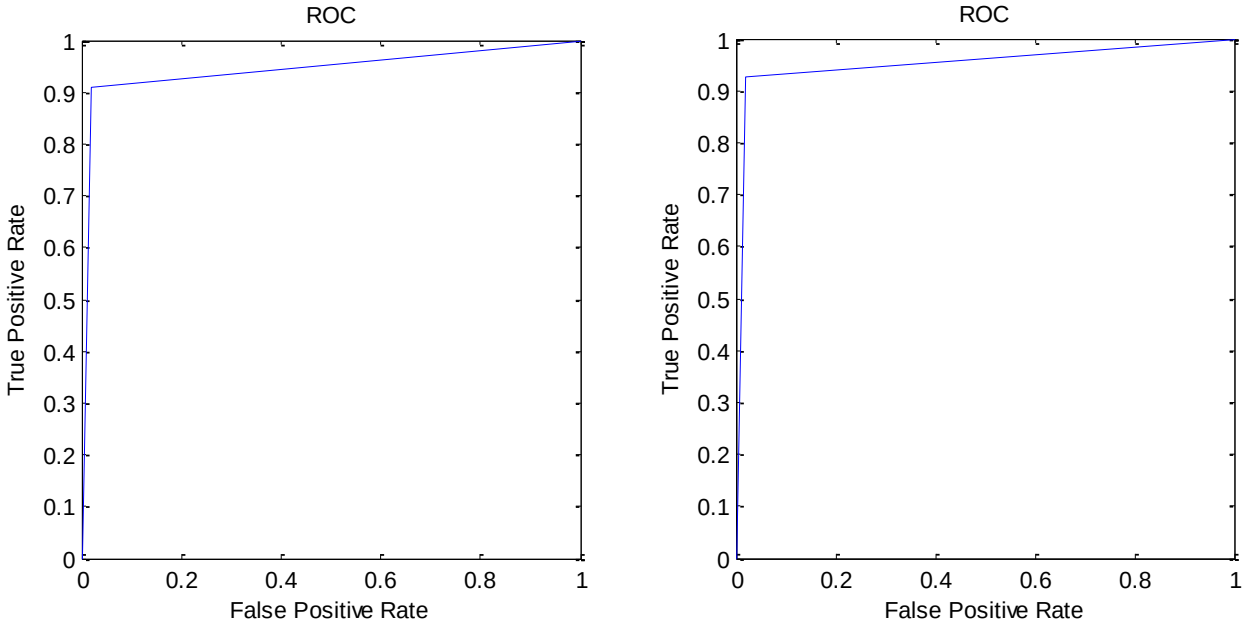
Ακρίβεια : 0.9561
Ευαισθησία : 0.9820
Προσδιοριστικότητα : 0.9079

– Μετά την επιλογή χαρακτηριστικών**:**
Πίνακας Σύγχυσης:

435	18
9	221

Ακρίβεια : 0.9605
Ευαισθησία : 0.9797
Προσδιοριστικότητα : 0.9247

Ο αλγόριθμος FCM ταξινομεί τα πρότυπα με μεγάλη επιτυχία, με την τιμή της προσδιοριστικότητας να υστερεί λίγο και την ευαισθησία να είναι πολύ μεγάλη. Η επιλογή χαρακτηριστικών, όπως απεικονίζεται στις καμπύλες του σχήματος 14, αυξάνει ελαφρά την ακρίβεια της κατηγοριοποίησης σε συνδυασμό μια μικρή μείωση της ευαισθησίας και μια μικρή αύξηση της προσδιοριστικότητας.



Σχήμα 14: Καμπύλη ROC του κατηγοριοποιητή FCM για τη διάγνωση καρκίνου του στήθους αριστερά με χρήση όλων των χαρακτηριστικών και δεξιά μετά την επιλογή χαρακτηριστικών

5.2.6. Αποτίμηση Κατηγοριοποιητών

Οι κλάσεις στις οποίες κατατάσσονται τα προτύπα του υπό εξέταση σύνολου δεδομένων φαίνεται ότι είναι καλά διαχωρίσιμες μεταξύ τους καθώς όλοι οι κατηγοριοποιητές αποδίδουν πολύ ακριβή αποτελέσματα. Μεταξύ των πολύ καλών αυτών επιδόσεων ξεχωρίζει ο FCM ως ο λιγότερο αποδοτικός. Όλοι οι υπόλοιποι αλγόριθμοι έχουν παρεμφερή αποτελέσματα με τον αλγόριθμο SOM να παρουσιάζει ελαφρώς καλύτερους δείκτες.

5.3. Διάγνωση Καρδιακής Νόσου

Το διαθέσιμο σύνολο δεδομένων περιλαμβάνει 87 διανύσματα με 13 χαρακτηριστικά, τα οποία για να βελτιωθεί η απόδοση της κατηγοριοποίησης μπορούν να μειωθούν σε μόλις 3.

5.3.1. Κατηγοριοποιητής BK

- Με όλα τα χαρακτηριστικά:

  Πίνακας Σύγχυσης:

46	8
2	31

Ακρίβεια : 0.8370 ± 0.0525
Ευαισθησία : 0.8958 ± 0.0532
Προσδιοριστικότητα : 0.7640 ± 0.0879

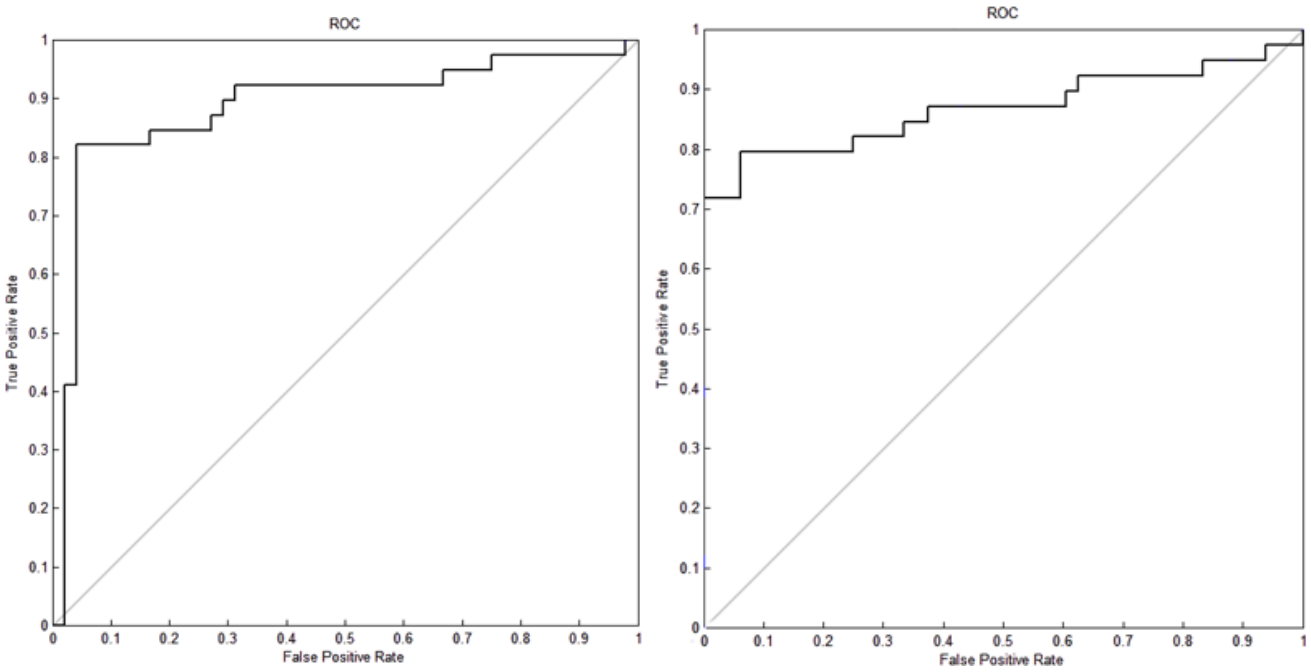
- Μετά την επιλογή χαρακτηριστικών:

  Πίνακας Σύγχυσης:

46	11
2	28

Ακρίβεια : 0.8334 ± 0.0246
Ευαισθησία : 0.8794 ± 0.0708
Προσδιοριστικότητα : 0.7948 ± 0.0599

Ο κατηγοριοποιητής BK λειτουργεί πολύ καλά με αιχμή τη μεγάλη τιμή ευαισθησίας. Όπως καταγράφεται και στις καμπύλες ROC του σχήματος 15, μετά την επιλογή χαρακτηριστικών η ακρίβεια και η προσδιοριστικότητα μειώνονται ελαφρά, όμως η ευαισθησία αυξάνεται φτάνοντας το 100%. Αν το πλήθος των προτύπων ήταν μεγαλύτερο θα ήταν δυνατό να ισχυριστούμε ότι ο κατηγοριοποιητής ταξινομεί σωστά όλα τα διανύσματα που ανήκουν στην κλάση του ασθενούς και κατά συνέπεια αν το αποτέλεσμα της κατηγοριοποίησης ενός προτύπου είναι αρνητικό να μπορούμε να είμαστε σίγουροι ότι το πρότυπο αυτό είναι στην πραγματικότητα αρνητικό.



Σχήμα 15: Καμπύλη ROC του κατηγοριοποιητή ΒΚ για τη διάγνωση Καρδιακής Νόσου αριστερά με χρήση όλων των χαρακτηριστικών και δεξιά μετά την επιλογή χαρακτηριστικών μετά την επιλογή χαρακτηριστικών.

5.3.2. Κατηγοριοποιητής SVM

Η κατηγοριοποίηση έγινε με χρήση του πολυωνυμικού πυρήνα.

– Με όλα τα χαρακτηριστικά:

Πίνακας Σύγχυσης:

20	4
4	15

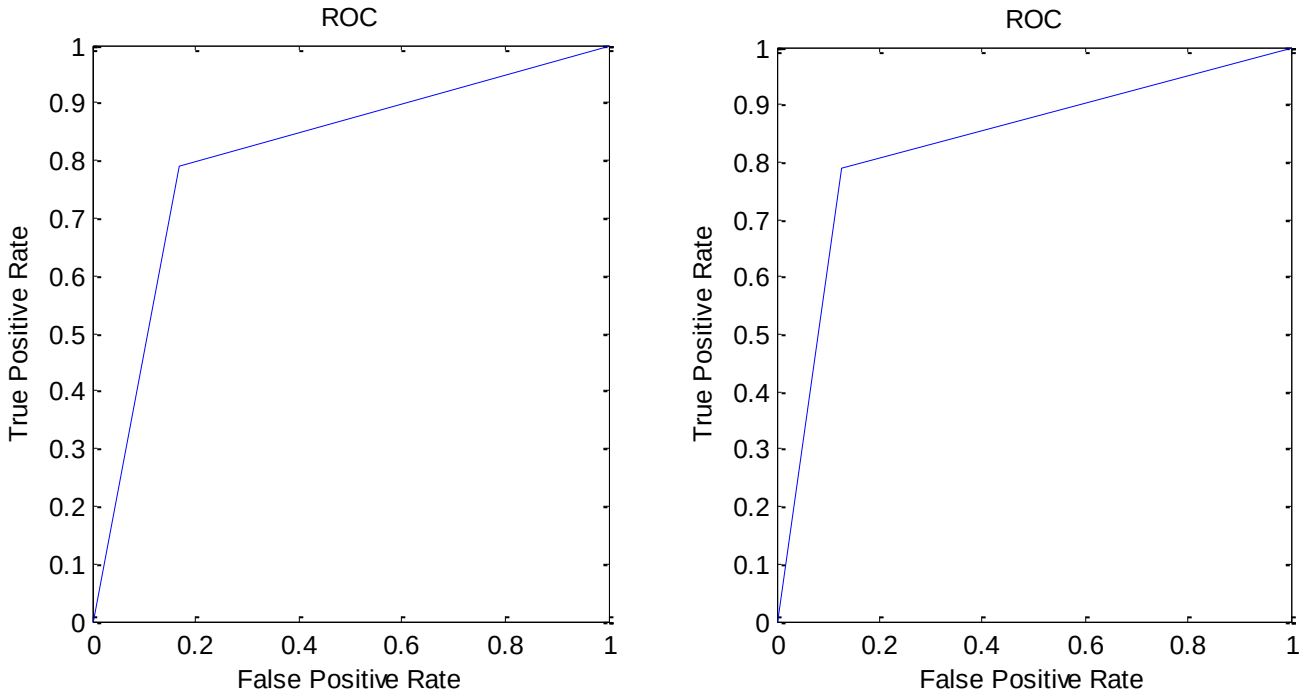
Ακρίβεια : 0.8279 ± 0.0539
Ευαισθησία : 0.8792 ± 0.1119
Προσδιοριστικότητα : 0.7632 ± 0.0794

– Μετά την επιλογή χαρακτηριστικών:

Πίνακας Σύγχυσης:

22	4
2	15

Ακρίβεια : 0.8442 ± 0.0560
Ευαισθησία : 0.9042 ± 0.0591
Προσδιοριστικότητα : 0.7684 ± 0.0901



Σχήμα 16: Καμπύλη ROC του κατηγοριοποιητή SVM για τη διάγνωση Καρδιακής Νόσου αριστερά με χρήση όλων των χαρακτηριστικών και δεξιά μετά την επιλογή χαρακτηριστικών μετά την επιλογή χαρακτηριστικών.

Η απόδοση του κατηγοριοποιητή είναι αρκετά ικανοποιητική, παρουσιάζεται όμως μια αστάθεια ως προς την ευαισθησία καθώς παρουσιάζει μεγάλη τυπική απόκλιση. Μετά την επιλογή χαρακτηριστικών (βλ. σχήμα 16) οι δείκτες βελτιώνονται ενώ οι τιμές τους σταθεροποιούνται σε σχέση με πριν.

5.3.3. Κατηγοριοποιητής kNN

5.3.3.1. Κατηγοριοποιητής 3NN

– <u>Με όλα τα χαρακτηριστικά</u>**:**
Πίνακας Σύγχυσης:

21	13
12	14

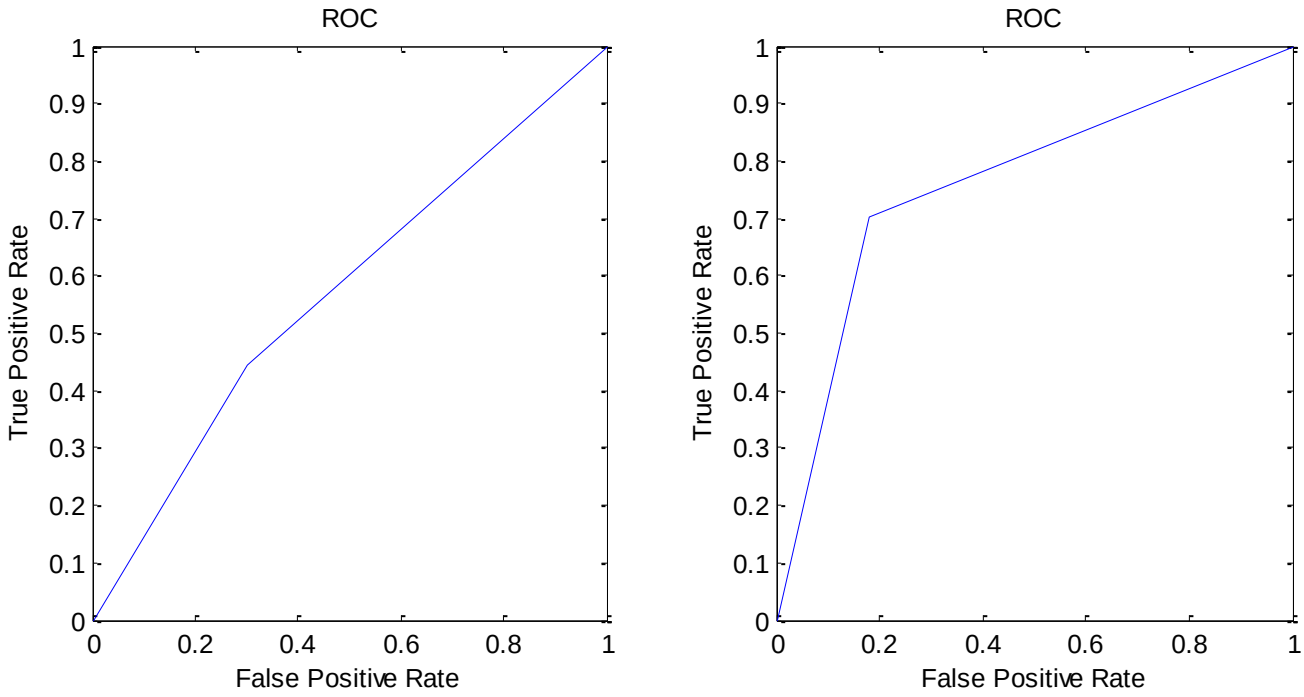
Ακρίβεια : 0.5742 ± 0.0427
Ευαισθησία : 0.6636 ± 0.1348
Προσδιοριστικότητα : 0.4648 ± 0.1286

– <u>Μετά την επιλογή χαρακτηριστικών</u>:
Πίνακας Σύγχυσης:

26	8
7	19

Ακρίβεια : 0.7483 ± 0.0319
Ευαισθησία : 0.7788 ± 0.0918
Προσδιοριστικότητα : 0.7111 ± 0.1059

Ο κατηγοριοποιητής 3NN έχει πολύ ασταθή συμπεριφορά, όπως φίνεται το σχήμα 17. Οι μέσες τιμές των δεικτών απόδοσης να είναι πολύ χαμηλές και επομένως η συμπεριφορά του είναι πολύ απρόβλεπτη. Μετά την επιλογή χαρακτηριστικών οι μέσες τιμές των δεικτών αυξάνονται και η αστάθεια μειώνεται όχι όμως αρκετά ώστε να θεωρηθεί ο κατηγοριοποιητής αξιόπιστος.



Σχήμα 17: Καμπύλη ROC του κατηγοριοποιητή 3NN για τη διάγνωση Καρδιακής Νόσου αριστερά με χρήση όλων των χαρακτηριστικών και δεξιά μετά την επιλογή χαρακτηριστικών μετά την επιλογή χαρακτηριστικών.

5.3.3.2. Κατηγοριοποιητής 5NN

– <u>Με όλα τα χαρακτηριστικά</u>**:**
Πίνακας Σύγχυσης:

23	16
10	11

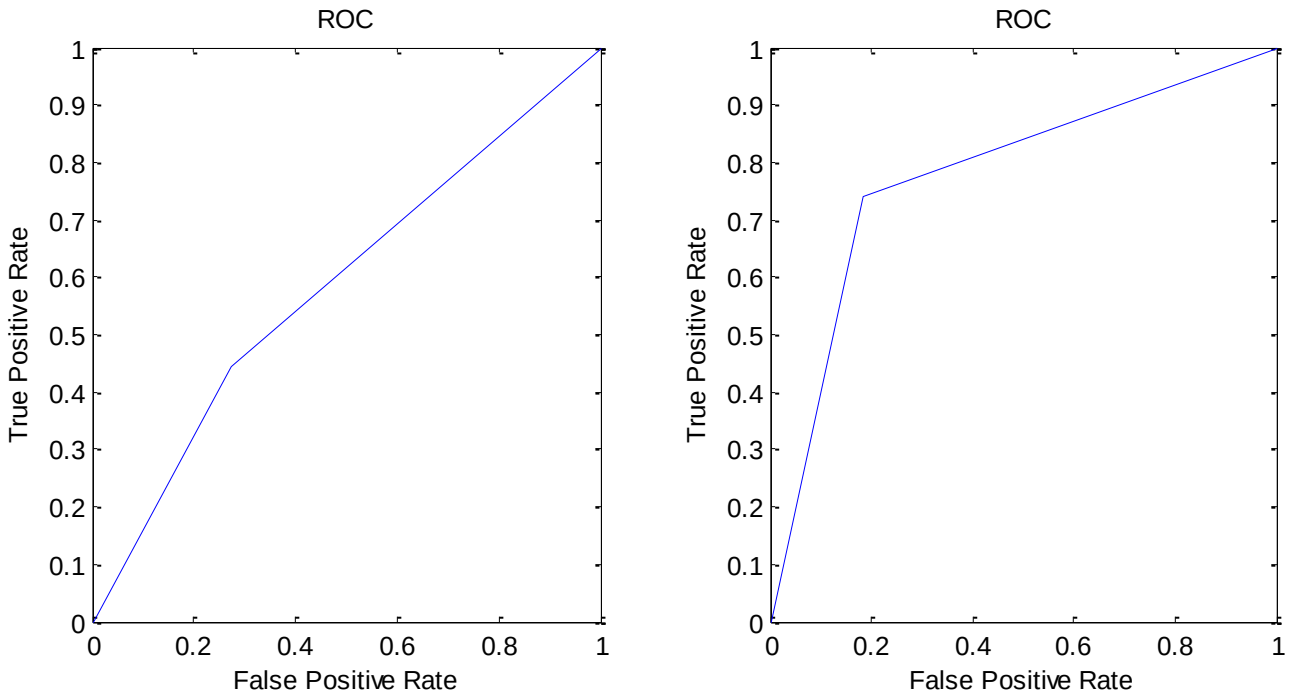
Ακρίβεια : 0.5708 ± 0.0518
Ευαισθησία : 0.6894 ± 0.1196
Προσδιοριστικότητα : 0.4259 ± 0.1555

– Μετά την επιλογή χαρακτηριστικών:

Πίνακας Σύγχυσης:

26	6
7	21

Ακρίβεια : 0.7825 ± 0.0327
Ευαισθησία : 0.8015 ± 0.0871
Προσδιοριστικότητα : 0.7593 ± 0.0765



Σχήμα 18: Καμπύλη ROC του κατηγοριοποιητή 5NN για τη διάγνωση Καρδιακής Νόσου αριστερά με χρήση όλων των χαρακτηριστικών και δεξιά μετά την επιλογή χαρακτηριστικών μετά την επιλογή χαρακτηριστικών.

Όπως καταδεικνύεται από το σχήμα 17, ο κατηγοριοποιητής 5NN έχει πολύ ασταθή συμπεριφορά, με τις μέσες τιμές των δεικτών απόδοσης να είναι πολύ χαμηλές και επομένως η χρήση του δεν προσφέρεται καθόλου για την κατηγοριοποίηση στο συγκεκριμένο σύνολο δεδομένων. Μετά την επιλογή χαρακτηριστικών οι μέσες τιμές των δεικτών αυξάνονται όμως η αστάθεια εξακολουθεί να υπάρχει καθιστώντας αναξιόπιστη τη χρήση του κατηγοριοποιητή.

5.3.3.3. Κατηγοριοποιητής 7NN

– Με όλα τα χαρακτηριστικά:

Πίνακας Σύγχυσης:

22	14
11	13

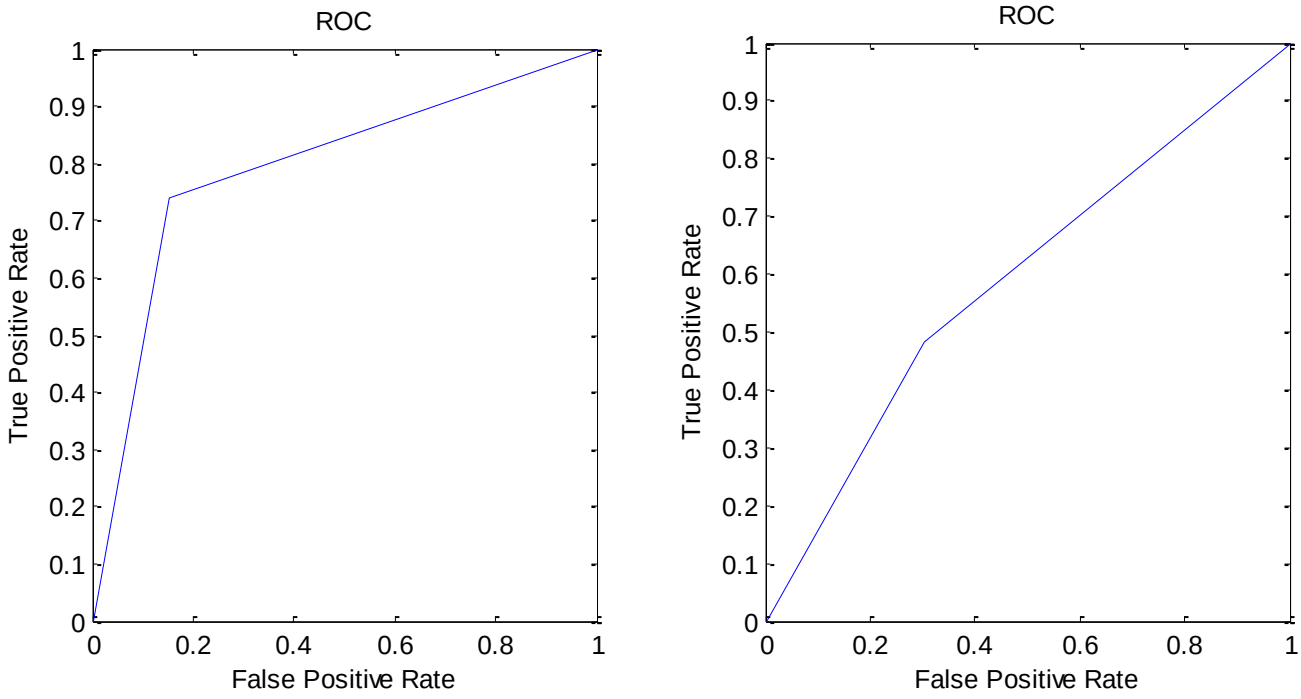
Ακρίβεια : 0.5775 ± 0.0531
Ευαισθησία : 0.7076 ± 0.1646
Προσδιοριστικότητα : 0.4185 ± 0.1525

– Μετά την επιλογή χαρακτηριστικών:

Πίνακας Σύγχυσης:

27	6
6	21

Ακρίβεια : 0.8060 ± 0.0242
Ευαισθησία : 0.8530 ± 0.0420
Προσδιοριστικότητα : 0.7460 ± 0.0592



Σχήμα 19: Καμπύλη ROC του κατηγοριοποιητή 7NN για τη διάγνωση Καρδιακής Νόσου αριστερά με χρήση όλων των χαρακτηριστικών και δεξιά μετά την επιλογή χαρακτηριστικών μετά την επιλογή χαρακτηριστικών.

Η συμπεριφορά του κατηγοριοποιητή αυτού όταν χρησιμοποιούνται όλα τα χαρακτηριστικά είναι πολύ ασταθής και ακόμα και στη μέση περίπτωση οι δείκτες δεν είναι ικανοποιητικοί (βλ. σχήμα 19). Μετά την επιλογή χαρακτηριστικών όμως οι επιδόσεις του κατηγοριοποιητή βελτιώνονται σημαντικά και φτάνουν σε αρκετά ικανοποιητικά επίπεδα με την τιμή της ευαισθησίας να είναι σε πολύ καλο επίπεδο σε ενώ η συμπεριφορά του παύει να έχει την αστάθεια που παρατηρείται όταν χρησιμοποιούνται όλα τα χαρακτηριστικά.

5.3.4. Κατηγοριοποιητής SOM

- Με όλα τα χαρακτηριστικά:

  Πίνακας Σύγχυσης:

38	11
10	28

Ακρίβεια : 0.7586
Ευαισθησία : 0.7919
Προσδιοριστικότητα : 0.7179

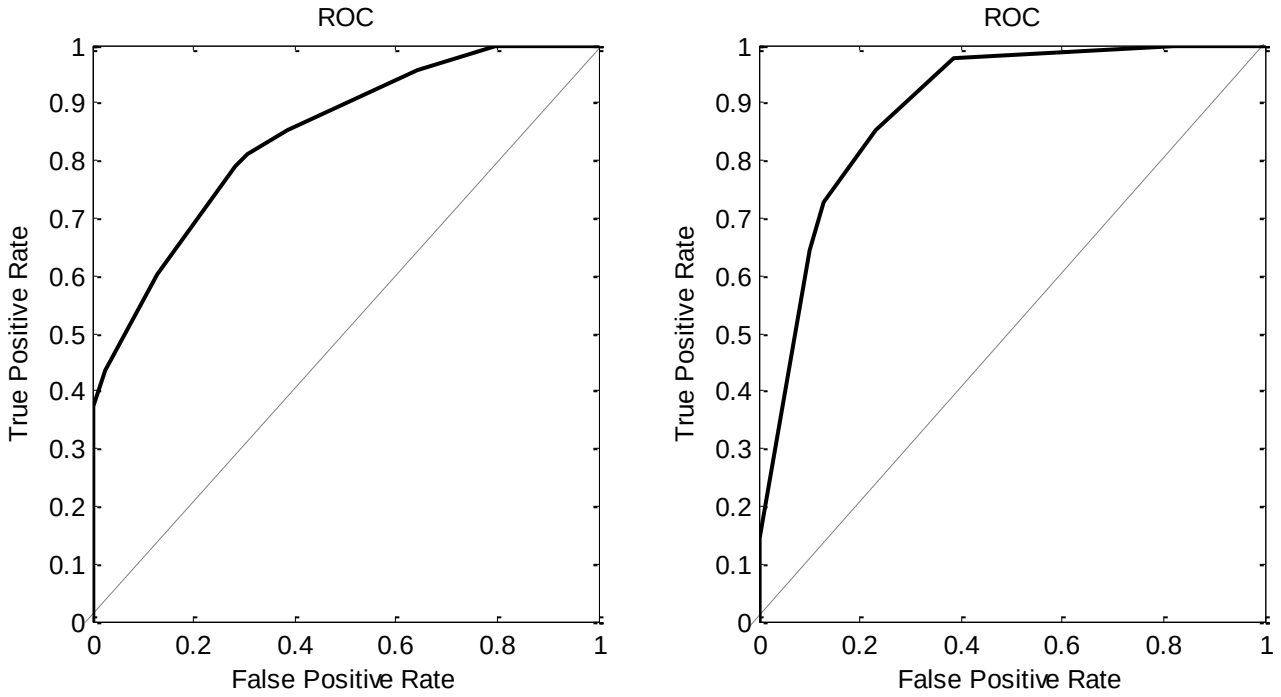
- Μετά την επιλογή χαρακτηριστικών:

  Πίνακας Σύγχυσης:

41	9
7	30

Ακρίβεια : 0.8161
Ευαισθησία : 0.8541
Προσδιοριστικότητα : 0.7692

Η κατηγοριοποίηση με τη χρήση της μεθόδου SOM είναι αρκετά επιτυχημένη με τους 3 δείκτες να κυμαίνονται σε παρόμοια επίπεδα. Όπως φαίνεται και στις καμπύλες ROC του σχήματος 20 η επιλογή χαρακτηριστικών επιφέρει αξιοσημείωτη βελτίωση με την προσδιοριστικότητα να κινείται σε λίγο χαμηλότερα επίπεδα από τους άλλους δείκτες.



Σχήμα 20: Καμπύλη ROC του κατηγοριοποιητή SOM για τη διάγνωση Καρδιακής Νόσου αριστερά με χρήση όλων των χαρακτηριστικών και δεξιά μετά την επιλογή χαρακτηριστικών μετά την επιλογή χαρακτηριστικών.

5.3.5. Κατηγοριοποιητής FCM

– Με όλα τα χαρακτηριστικά:

Πίνακας Σύγχυσης:

31	24
17	15

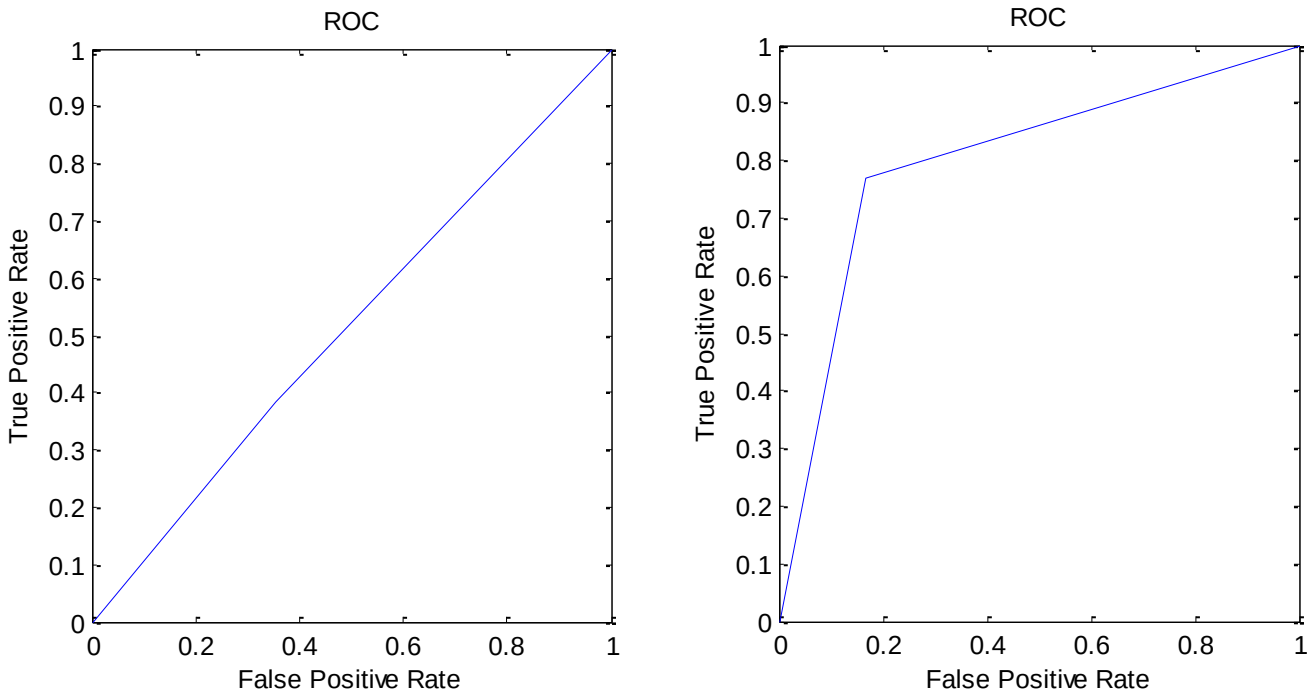
Ακρίβεια : 0.5287
Ευαισθησία : 0.6458
Προσδιοριστικότητα : 0.3846

– Μετά την επιλογή χαρακτηριστικών:

Πίνακας Σύγχυσης:

40	9
8	30

Ακρίβεια : 0.8046
Ευαισθησία : 0.8333
Προσδιοριστικότητα : 0.7692



Σχήμα 21: Καμπύλη ROC του κατηγοριοποιητή FCM για τη διάγνωση Καρδιακής Νόσου αριστερά με χρήση όλων των χαρακτηριστικών και δεξιά μετά την επιλογή χαρακτηριστικών μετά την επιλογή χαρακτηριστικών.

Όπως είναι εμφανές από τις καμπύλες του σχήματος 21, η κατηγοριοποίηση του συγκεκριμένου συνόλου δεδομένων με τη χρήση του αλγορίθμου FCM είναι ενδεικτική της μεγάλης βελτίωσης στην απόδοση που μπορεί να επιφέρει η επιλογή χαρακτηριστικών. Παρ' όλο που η χρήση όλων των χαρακτηριστικών έχει ως αποτέλεσμα μια κακή κατηγοριοποίηση, μετά την επιλογή χαρακτηριστικών προκύπτει κατηγοριοποίηση με πολύ ικανοποιητικές επιδόσεις.

5.3.6. Αποτίμηση Κατηγοριοποιητών

Το συγκεκριμένο σύνολο δεδομένων κατηγοριοποιείται πολύ ικανοποιητικά από όλους τους κατηγοριοποιητές. Τα στατιστικά μέτρα που αξιολογούνται είναι για όλους του κατηγοριοποιητές πολύ παρεμφερή, από πλευράς ευαισθησίας όμως ξεχωρίζει ο αλγόριθμος SVM, ενώ ακολουθεί ο αλγόριθμος BK. Από τους κατηγοριοποιητές της οικογένειας kNN ο 7NN λειτουργεί καλύτερα καθώς οι 3NN και 5NN παρουσιάζουν αστάθειες. Στο ίδιο επίπεδο με τον 7NN βρίσκονται και οι SOM οι FCM.

5.4. Διάγνωση Νόσου Parkinson

Το διαθέσιμο σύνολο δεδομένων περιλαμβάνει 195 διανύσματα 22 χαρακτηριστικών τα οποία μειώνονται σε 15 με τη χρήση της επιλογής χαρακτηριστικών.

5.4.1. Κατηγοριοποιητής BK

– Με όλα τα χαρακτηριστικά:
Πίνακας Σύγχυσης:

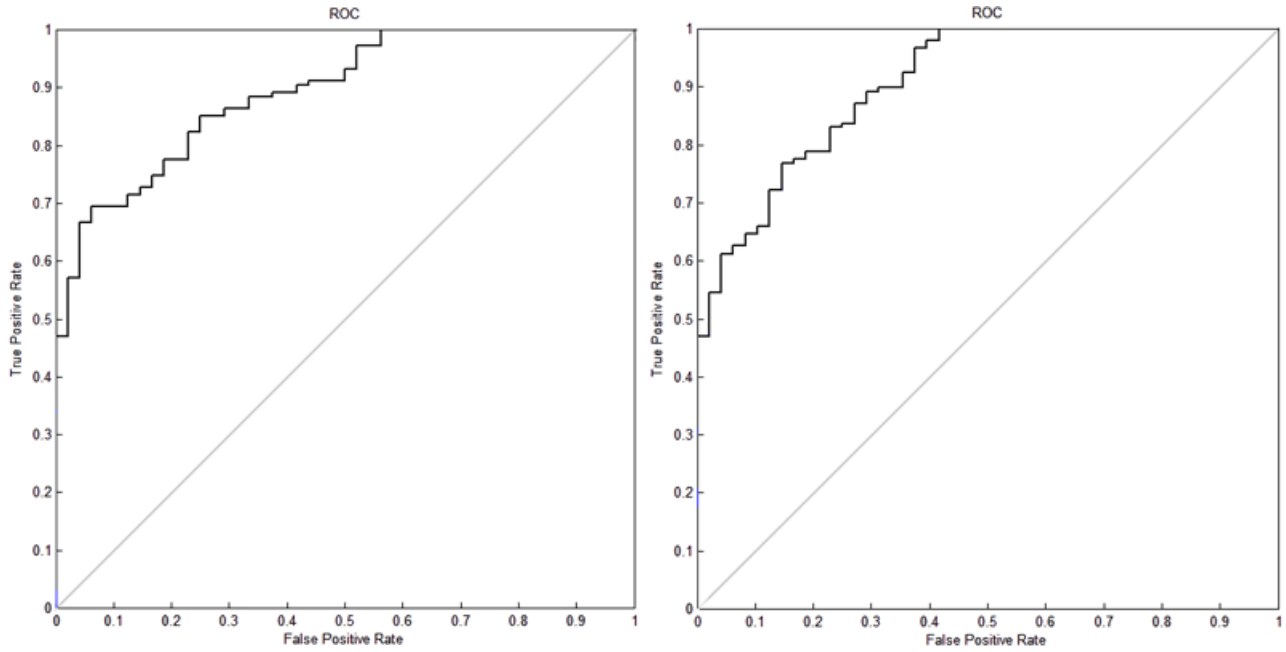
23	6
25	141

Ακρίβεια : 0.8718 ± 0.0162
Ευαισθησία : 0.5710 ± 0.0235
Προσδιοριστικότητα : 0.9590 ± 0.0217

– Μετά την επιλογή χαρακτηριστικών:
Πίνακας Σύγχυσης:

28	9
20	147

Ακρίβεια : 0.8790 ± 0.0104
Ευαισθησία : 0.6082 ± 0.0176
Προσδιοριστικότητα : 0.9670 ± 0.0076



Σχήμα 22: Καμπύλη ROC του κατηγοριοποιητή ΒΚ για τη διάγνωση της Νόσου Parkinson αριστερά με χρήση όλων των χαρακτηριστικών και δεξιά μετά την επιλογή χαρακτηριστικών

Η χρήση του κατηγοριοποιητή ΒΚ έχει πολύ καλά αποτελέσματα, όσον αφορά στην ακρίβεια και ειδικά την προσδιοριστικότητα η οποία εμφανίζει πολύ καλή τιμή. Στον αντίποδα, η ευαισθησία έχει πολύ χαμηλή τιμή. Μετά την επιλογή χαρακτηριστικών, η απόδοση του κατηγοριοποιητή αυξάνεται ελαφρά με όλους τους δείκτες να αυξάνονται η γενική εικόνα της κατηγοριοποίησης όμως παραμένει ίδια.

5.4.2. Κατηγοριοποιητής SVM

Στην κατηγοριοποίηση χρησιμοποιήθηκε πολυωνυμική συνάρτηση πυρήνα.

– Με όλα τα χαρακτηριστικά:

Πίνακας Σύγχυσης:

18	9
6	64

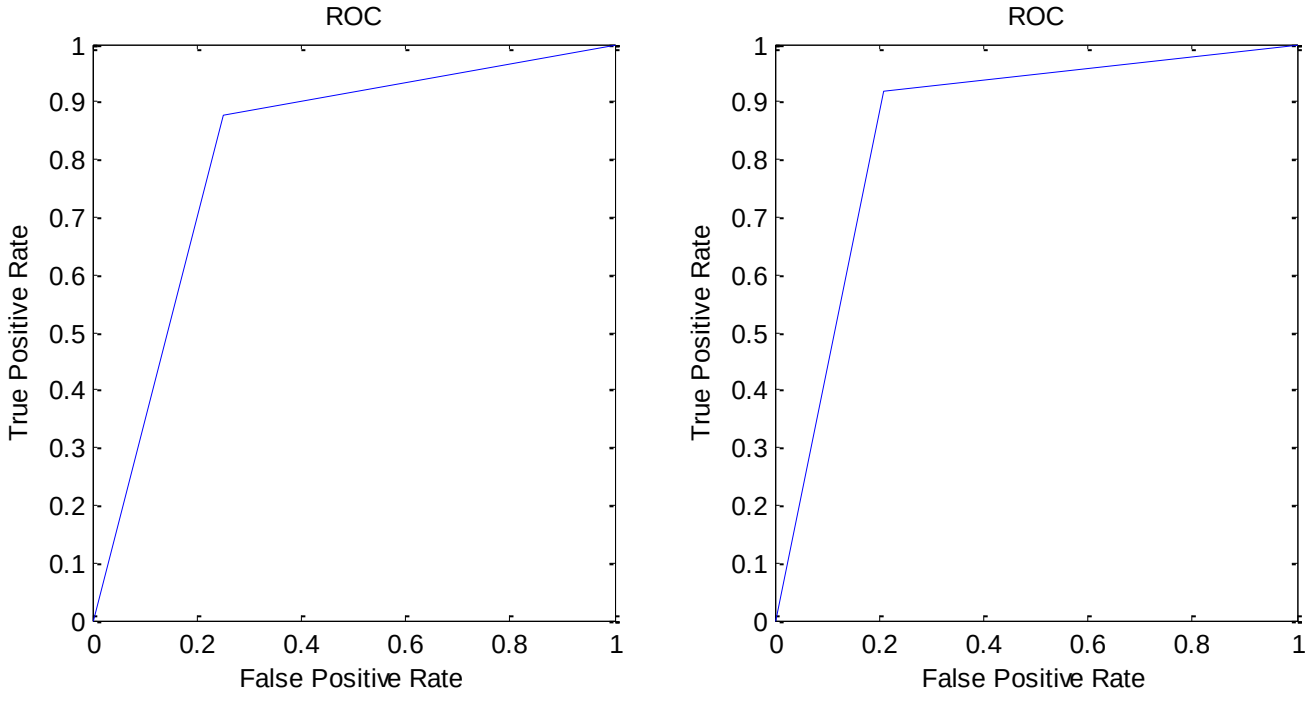
Ακρίβεια : 0.8464 ± 0.0399
Ευαισθησία : 0.7292 ± 0.1098
Προσδιοριστικότητα : 0.8849 ± 0.0429

– Μετά την επιλογή χαρακτηριστικών:

Πίνακας Σύγχυσης:

18	6
6	67

Ακρίβεια : 0.8784 ± 0.0369
Ευαισθησία : 0.7750 ± 0.0925
Προσδιοριστικότητα : 0.9123 ± 0.0297



Σχήμα 23: Καμπύλη ROC του κατηγοριοποιητή SVM για τη διάγνωση της Νόσου Parkinson αριστερά με χρήση όλων των χαρακτηριστικών και δεξιά μετά την επιλογή χαρακτηριστικών

Η απόδοση του κατηγοριοποιητή SVM είναι πολύ καλή και όπως καταδεικνύουν και οι καμπύλες ROC του σχήματος 23 βελτιώνεται μετά την επιλογή χαρακτηριστικών. Το μόνο σημείο που χρειάζεται προσοχή είναι η κατηγοριοποίηση των διανυσμάτων της θετικής κλάσης, καθώς η ευαισθησία βρίσκεται σε χαμηλότερα επίπεδα και παρουσιάζει αρκετά μεγάλες διακυμάνσεις ανάλογα με το σύνολο εκπαίδευσης.

5.4.3. Κατηγοριοποιητής kNN

5.4.3.1. Κατηγοριοποιητής 3NN

- Με όλα τα χαρακτηριστικά:

Πίνακας Σύγχυσης:

16	7
17	95

Ακρίβεια : 0.8148 ± 0.0264
Ευαισθησία : 0.4545 ± 0.1000
Προσδιοριστικότητα : 0.9314 ± 0.0467

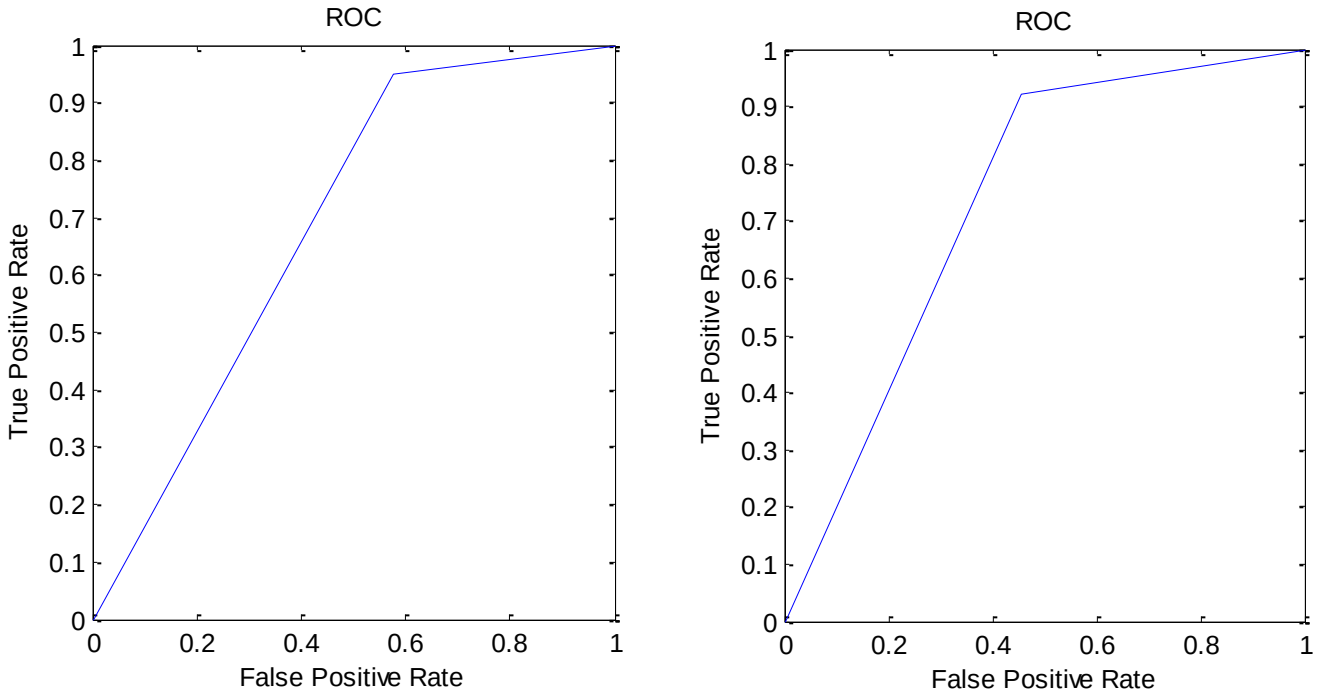
- Μετά την επιλογή χαρακτηριστικών:

Πίνακας Σύγχυσης:

18	8
15	94

Ακρίβεια : 0.8289 ± 0.0278
Ευαισθησία : 0.5152 ± 0.0799
Προσδιοριστικότητα : 0.9304 ± 0.0327

Στην περίπτωση αυτή η ακρίβεια είναι αρκετά ικανοποιητική, οφείλεται όμως σε μεγαλύτερο ποσοστό στη μεγάλη προσδιοριστικότητα και το μεγάλο πλήθος των προτύπων της κλάσης «υγιής», αφού η ευαισθησία είναι πολύ χαμηλή και ασταθής. Όπως φαίνεται και στο σχήμα 24, η επιλογή χαρακτηριστικών διατηρεί αυτή την εικόνα, με τις μέσες τιμές να παρουσιάζουν μια μικρη αύξηση.



Σχήμα 24: Καμπύλη ROC του κατηγοριοποιητή 3NN για τη διάγνωση της Νόσου Parkinson αριστερά με χρήση όλων των χαρακτηριστικών και δεξιά μετά την επιλογή χαρακτηριστικών

5.4.3.2. Κατηγοριοποιητής 5NN

– Με όλα τα χαρακτηριστικά:
Πίνακας Σύγχυσης:

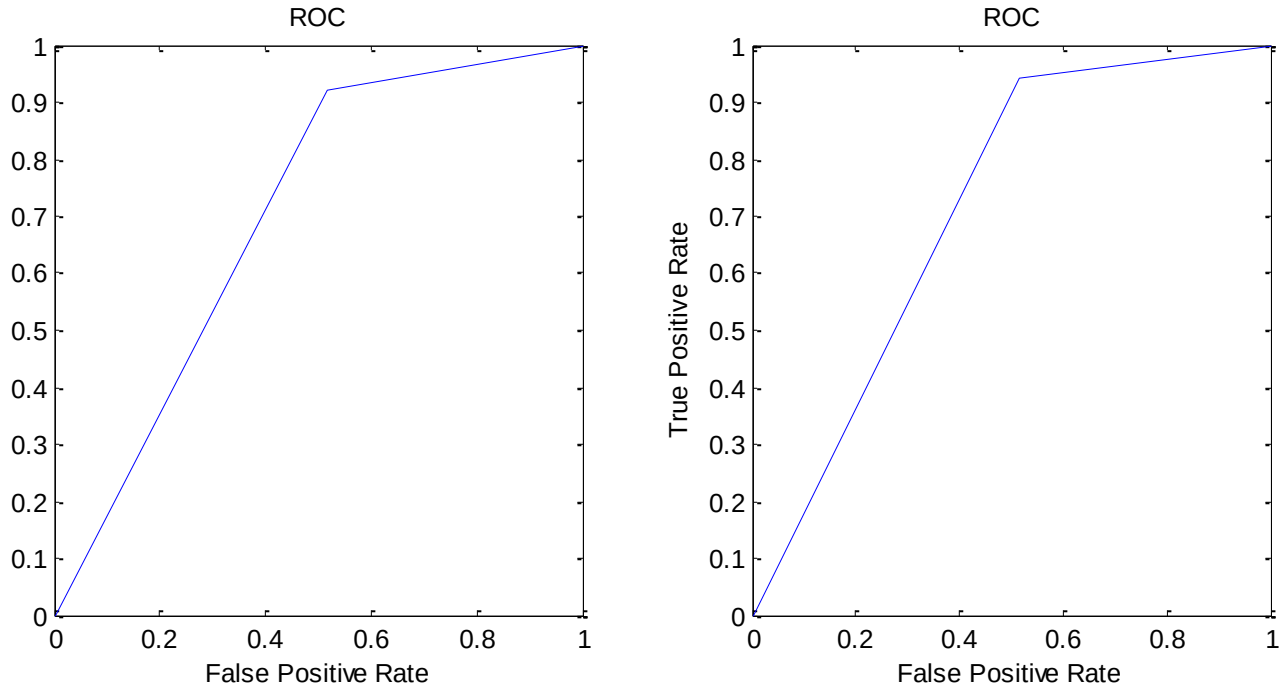
15	6
18	96

Ακρίβεια : 0.8133 ± 0.0191
Ευαισθησία : 0.4303 ± 0.0806
Προσδιοριστικότητα : 0.9373 ± 0.0232

– Μετά την επιλογή χαρακτηριστικών:
Πίνακας Σύγχυσης:

16	6
17	96

Ακρίβεια : 0.8233 ± 0.0288
Ευαισθησία : 0.4500 ± 0.1114
Προσδιοριστικότητα : 0.9441 ± 0.0415



Σχήμα 25: Καμπύλη ROC του κατηγοριοποιητή 5NN για τη διάγνωση της Νόσου Parkinson αριστερά με χρήση όλων των χαρακτηριστικών και δεξιά μετά την επιλογή χαρακτηριστικών

Η εικόνα της κατηγοριοποίησης με τη χρήση της μεθόδου 5NN είναι παραπλήσια με αυτή των 3NN. Η ακρίβεια κινείται σε καλά επίπεδα ενώ η ευαισθησία είναι ασταθής με μικρή μέση τιμή και η προσδιοριστικότητα είναι πάρα πολύ καλή. Όπως βλέπουμε και στο σχήμα 25 η επιλογή χαρακτηριστικών επιφέρει μια μικρή βελτίωση, διατηρώντας την υψηλή προσδιοριστικότητα αλλά και την χαμηλή ευαισθησία.

5.4.3.3. Κατηγοριοποιητής 7NN

– Με όλα τα χαρακτηριστικά:

Πίνακας Σύγχυσης:

12	8
21	94

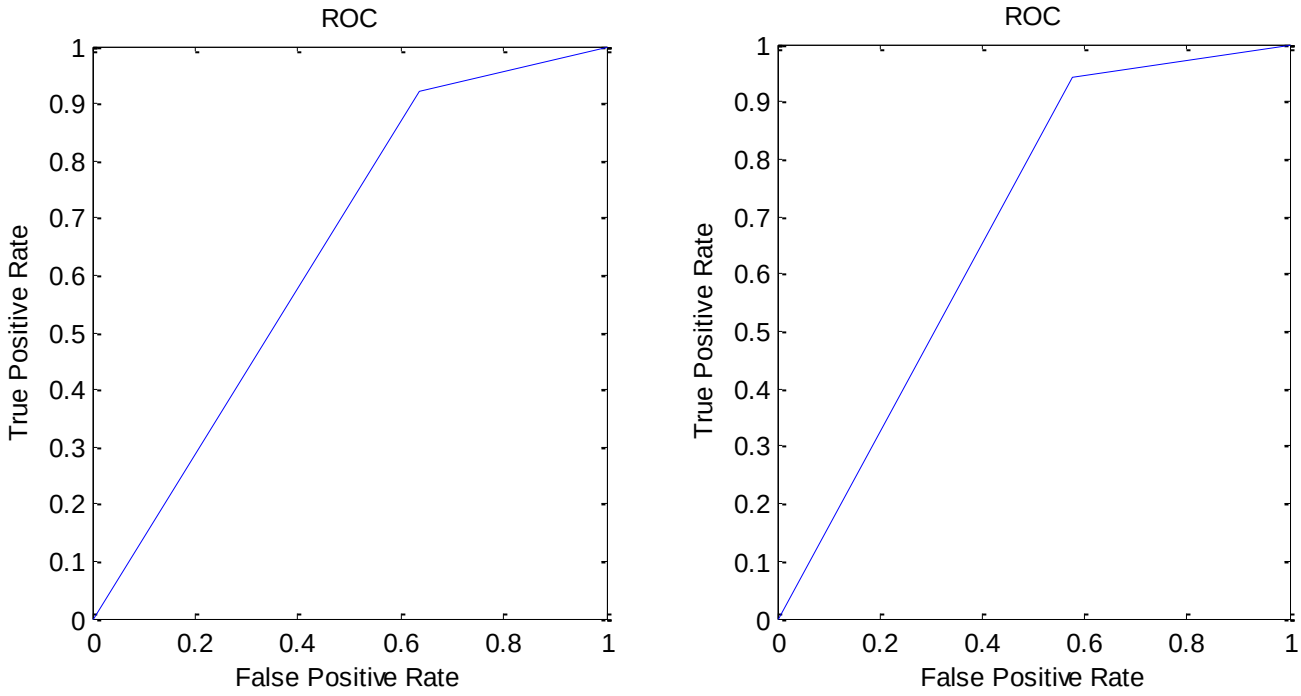
Ακρίβεια : 0.7822 ± 0.0389
Ευαισθησία : 0.3121 ± 0.0881
Προσδιοριστικότητα : 0.9343 ± 0.0395

– Μετά την επιλογή χαρακτηριστικών:

Πίνακας Σύγχυσης:

14	6
19	96

Ακρίβεια : 0.8115 ± 0.0218
Ευαισθησία : 0.4061 ± 0.0524
Προσδιοριστικότητα : 0.9426 ± 0.0339



Σχήμα 26: Καμπύλη ROC του κατηγοριοποιητή 7NN για τη διάγνωση της Νόσου Parkinson αριστερά με χρήση όλων των χαρακτηριστικών και δεξιά μετά την επιλογή χαρακτηριστικών

Η κατηγοριοποίηση με τη χρήση των 7NN ακολουθεί τις επιδόσεις των δύο άλλων κατηγοριοποιητών της οικογένειας 3NN και 5NN, καθώς οι τιμές των μέτρων απόδοσης είναι παρεμφερείς και η γενικότερη συμπεριφορά ίδια, με τη διαφορά ότι οι τιμές των μέτρων είναι λίγο χαμηλότερες από τις προηγούμενες όταν χρησιμοποιούνται όλα τα χαρακτηριστικά (βλ. σχήμα 26).

5.4.4. Κατηγοριοποιητής SOM

– <u>Με όλα τα χαρακτηριστικά</u>**:**
Πίνακας Σύγχυσης:

27	9
21	138

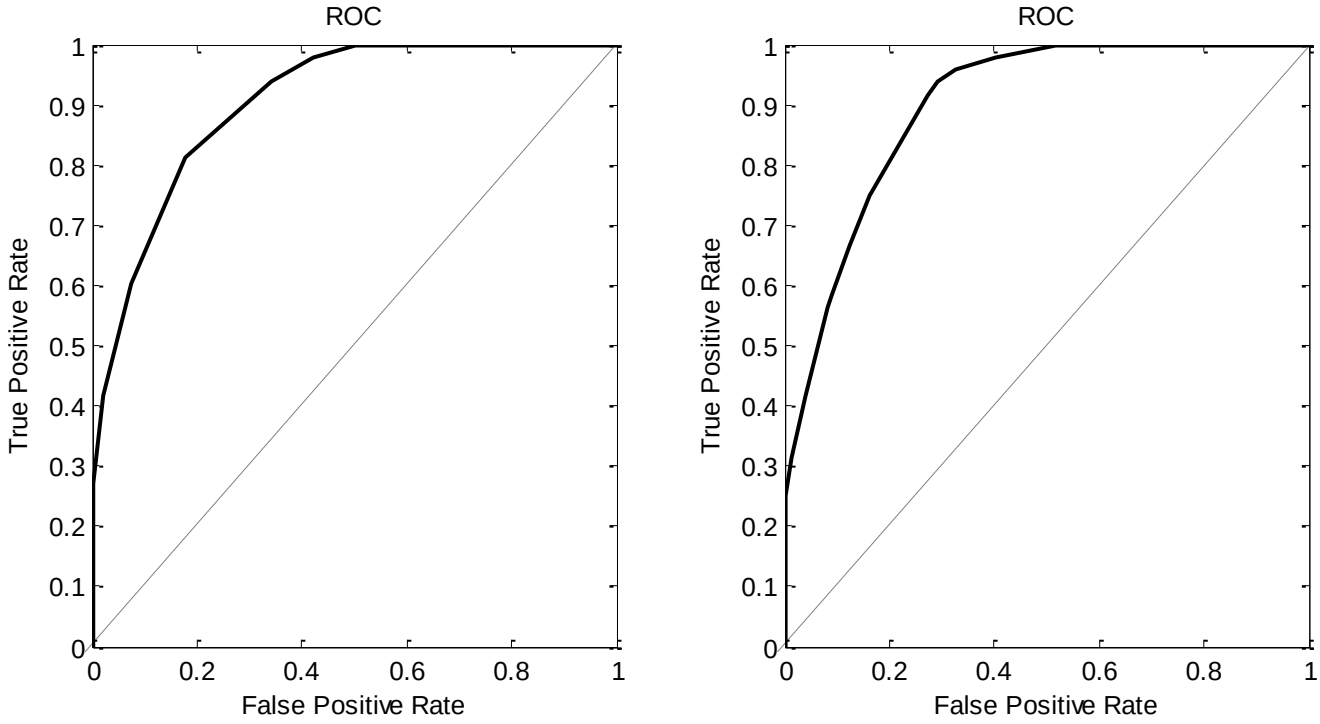
Ακρίβεια : 0.8462
Ευαισθησία : 0.5625
Προσδιοριστικότητα : 0.9388

– Μετά την επιλογή χαρακτηριστικών:
Πίνακας Σύγχυσης:

27	12
21	135

Ακρίβεια : 0.8308
Ευαισθησία : 0.5625
Προσδιοριστικότητα : 0.9184

Η κατηγοριοποίηση με χρήση SOM έχει αρκετά καλή απόδοση με έμφαση στην πολύ καλή προσδιοριστικότητα και την χαμηλή ευαισθησία. Εδώ παρουσιάζεται και μία περίπτωση όπου η επιλογή χαρακτηριστικών χειροτερεύει ελαφρά τα αποτελέσματα του κατηγοριοποιητή καθώς μειώνεται η προσδιοριστικότητα (βλ. σχήμα 27).



Σχήμα 27: Καμπύλη ROC του κατηγοριοποιητή SOM για τη διάγνωση της Νόσου Parkinson αριστερά με χρήση όλων των χαρακτηριστικών και δεξιά μετά την επιλογή χαρακτηριστικών

5.4.5. Κατηγοριοποιητής FCM

– Με όλα τα χαρακτηριστικά:

Πίνακας Σύγχυσης:

32	39
16	108

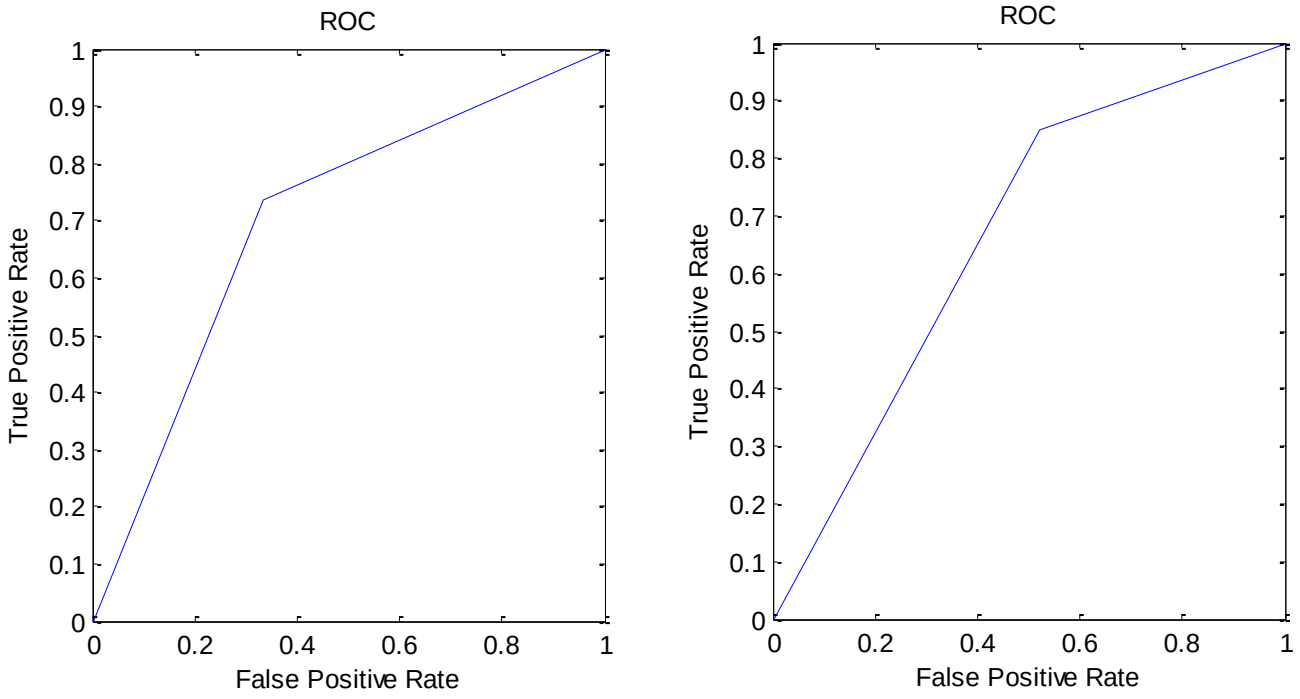
Ακρίβεια : 0.7179
Ευαισθησία : 0.6667
Προσδιοριστικότητα : 0.7347

– Μετά την επιλογή χαρακτηριστικών:

Πίνακας Σύγχυσης:

23	22
25	125

Ακρίβεια : 0.7590
Ευαισθησία : 0.4792
Προσδιοριστικότητα : 0.8503



Σχήμα 28: Καμπύλη ROC του κατηγοριοποιητή FCM για τη διάγνωση της Νόσου Parkinson αριστερά με χρήση όλων των χαρακτηριστικών και δεξιά μετά την επιλογή χαρακτηριστικών

Χρησιμοποιώντας τον κατηγοριοποιητή FCM προκύπτουν μέτριες επιδόσεις. Μετά την επιλογή χαρακτηριστικών οι δείκτες βελτιώνονται, με εξαίρεση την ευαισθησία η οποία μειώνεται αρκετά (βλ. σχήμα 28).

5.4.6. Αποτίμηση Κατηγοριοποιητών

Η κατηγοριοποίηση της συγκεκριμένης περίπτωσης συνόλου δεδομένων παρουσιάζει ικανοποιητική ακρίβεια, η ακρίβεια όμως αυτή οφείλεται στην πολύ καλή προσδιοριστικότητα σε συνδυασμό με το πολύ μεγαλύτερο πλήθος προτύπων στην κλάση των υγιών. Από την κατηγοριοποίηση με τον αλγόριθμο BK προκύπτει η καλύτερη προσδιοριστικότητα με χαμηλότερη όμως ευαισθησία. Ο καλύτερος συνδυασμός και των τριών δεικτών προκύπτει από τη μέθοδο SVM. Λίγο υποδεέστερη είναι η κατηγοριοποίηση που εκτελείται από τον αλγόριθμο SOM ενώ οι τρεις κατηγοριοποιητές της οικογένειας kNN παρουσιάζουν παρόμοια απόδοση οπότε αν απαιτείται γρήγορη κατηγοριοποίηση επιλέγεται ο 3NN ενώ αν αυτό είναι αδιάφορο επιλέγεται ο 5NN. Τη χειρότερη απόδοση παρουσιάζει ο αλγόριθμος FCM.

5.5. Διάγνωση από Τομογραφία SPECT

Το διαθέσιμο σύνολο δεδομένων περιλαμβάνει 267 διανύσματα με 21 χαρακτηριστικά από τα οποία -σύμφωνα με το αποτέλεσμα της επιλογής χαρακτηριστικών- είναι αρκετά μόνο 2.

5.5.1. Κατηγοριοποιητής BK

- Με όλα τα χαρακτηριστικά:

  Πίνακας Σύγχυσης:

135	53
22	57

  Ακρίβεια : 0.7190 ± 0.0340
  Ευαισθησία : 0.8138 ± 0.0444
  Προσδιοριστικότητα : 0.5834 ± 0.0650

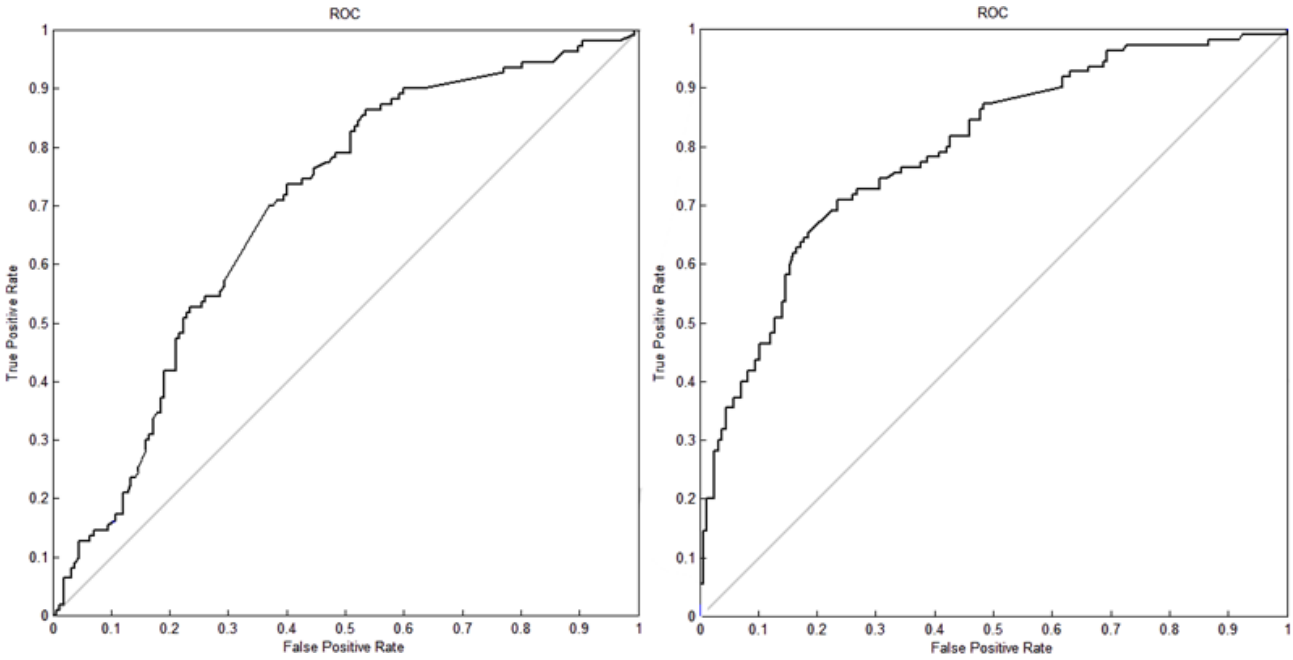
- Μετά την επιλογή χαρακτηριστικών:

  Πίνακας Σύγχυσης:

139	52
18	58

  Ακρίβεια : 0.7334 ± 0.0049
  Ευαισθησία : 0.8240 ± 0.0049

Προσδιοριστικότητα : 0.6036 ± 0.0137



Σχήμα 29: Καμπύλη ROC του κατηγοριοποιητή ΒΚ για τη διάγνωση από τομογραφία SPECT,αριστερά με χρήση όλων των χαρακτηριστικών και δεξιά μετά την επιλογή χαρακτηριστικών

Συμπεριλαμβάνοντας όλα τα χαρακτηριστικά η κατηγοριοποίηση γίνεται ικανοποιητικά παρουσιάζοντας υψηλή τιμή ευαισθησίας αλλά χαμηλή τιμή προσδιοριστικότητας. Τα χαρακτηριστικά που αφαιρέθηκαν κατά την επιλογή χαρακτηριστικών φαίνεται να αυξάνουν ελαφρά την ικανότητα κατηγοριοποίησης του ΒΚ ως προς τα θετικά πρότυπα αφού η προσδιοριστικότητα και η ευαισθησία παρουσιάζουν μία μικρή αύξηση (βλ. σχήμα 29).

5.5.2. Κατηγοριοποιητής SVM

Η συνάρτηση πυρήνα που χρησιμοποιείται είναι πολυωνυμική.

– Με όλα τα χαρακτηριστικά:

Πίνακας Σύγχυσης:

57	29
21	26

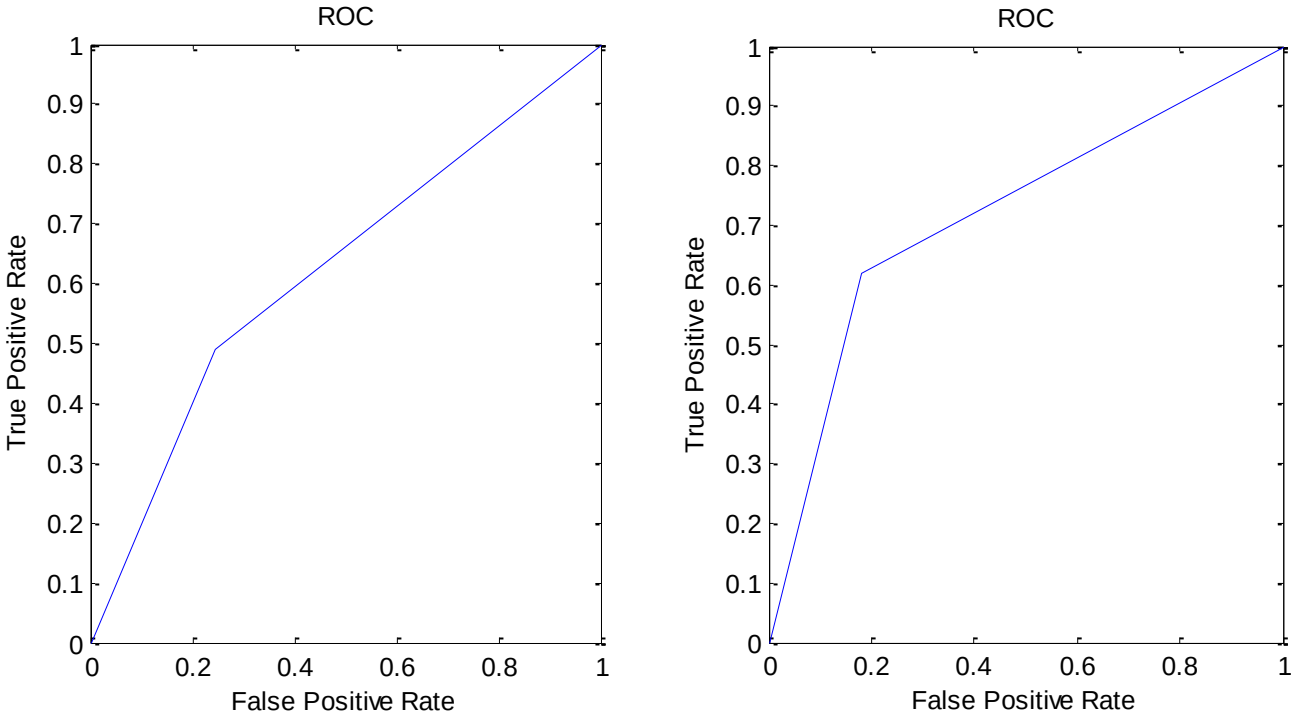
Ακρίβεια : 0.6308 ± 0.0514
Ευαισθησία : 0.7231 ± 0.0660
Προσδιοριστικότητα : 0.5000 ± 0.0832

– Μετά την επιλογή χαρακτηριστικών:

Πίνακας Σύγχυσης:

65	22
13	33

Ακρίβεια : 0.7376 ± 0.0306
Ευαισθησία : 0.8231 ± 0.0169
Προσδιοριστικότητα : 0.6164 ± 0.0655



Σχήμα 30: Καμπύλη ROC του κατηγοριοποιητή SVM για τη διάγνωση από τομογραφία SPECT, αριστερά με χρήση όλων των χαρακτηριστικών και δεξιά μετά την επιλογή χαρακτηριστικών

Στην περίπτωση αυτή με τη χρήση όλων των χαρακτηριστικών έχουμε μη ικανοποιητικές αποδόσεις στην κατηγοριοποίηση με την ευαισθησία να ξεχωρίζει από τα άλλα μέτρα. Μετά την επιλογή χαρακτηριστικών τα αποτελέσματα βελτιώνονται σημαντικά, φτάνοντας σε ικανοποιητικά επίπεδα, δεδομένης της υψηλής ευαισθησίας που παρουσιάζεται. Η προσδιοριστικότητα αυξάνεται μεν, αλλά παραμένει σε μέτρια επίπεδα (βλ. σχήμα 30).

5.5.3. Κατηγοριοποιητής kNN

5.5.3.1. Κατηγοριοποιητής 3NN

– Με όλα τα χαρακτηριστικά:
Πίνακας Σύγχυσης:

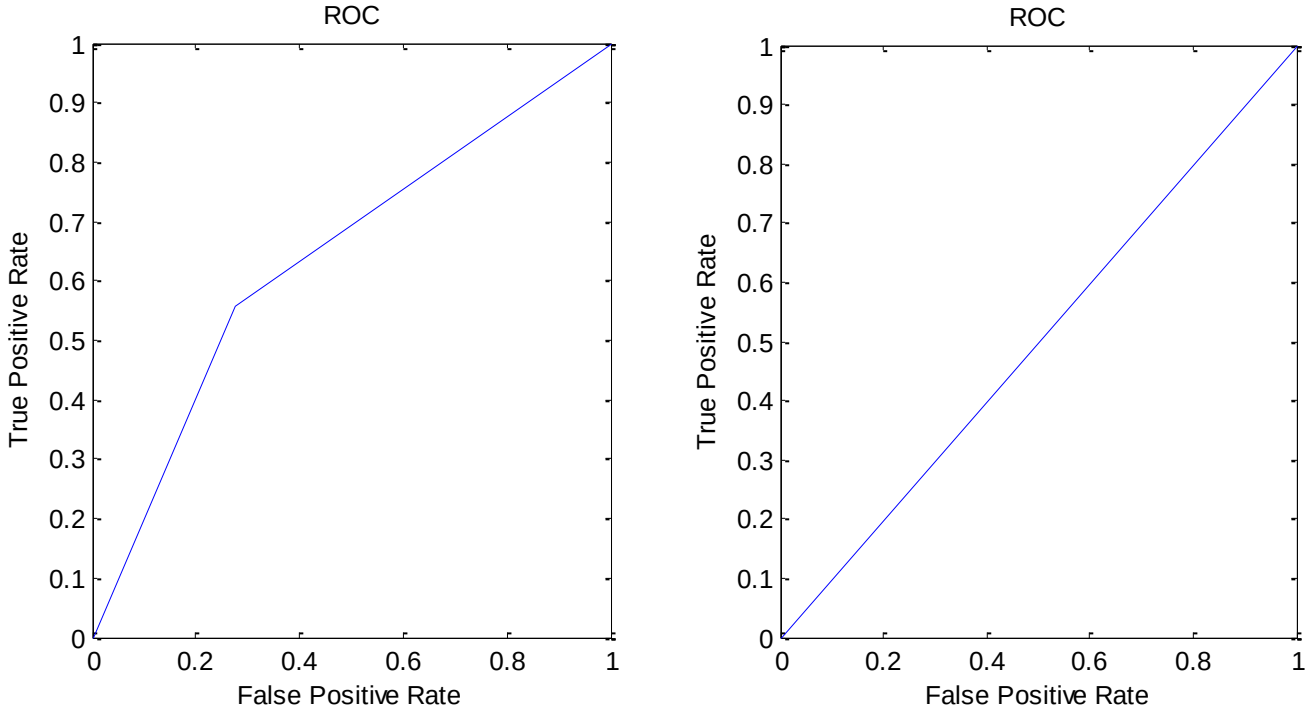
79	34
30	43

Ακρίβεια : 0.6634 ± 0.0387
Ευαισθησία : 0.7367 ± 0.0766
Προσδιοριστικότητα : 0.5597 ± 0.1104

– Μετά την επιλογή χαρακτηριστικών:
Πίνακας Σύγχυσης:

82	36
27	41

Ακρίβεια : 0.6202 ± 0.1241
Ευαισθησία : 0.7670 ± 0.2502
Προσδιοριστικότητα : 0.4123 ± 0.3191



Σχήμα 31: Καμπύλη ROC του κατηγοριοποιητή 3NN για τη διάγνωση από τομογραφία SPECT, αριστερά με χρήση όλων των χαρακτηριστικών και δεξιά μετά την επιλογή χαρακτηριστικών

Όπως είναι εμφανές στις καμπύλες του σχήματος 31 η κατηγοριοποίηση με χρήση της μεθόδου 3NN χαρακτηρίζεται μη ικανοποιητική καθώς τόσο πριν την επιλογή χαρακτηριστικών όσο και μετά οι δείκτες απόδοσης κυμαίνονται σε χαμηλές τιμές. Η

επιλογή χαρακτηριστικών μάλιστα επιφέρει μείωση της ακρίβειας παρά την αύξηση της ευαισθησίας καθώς η προσδιοριστικότητα μειώνεται σε σχέση με την αρχική.

5.5.3.2. Κατηγοριοποιητής 5NN

– Με όλα τα χαρακτηριστικά:

Πίνακας Σύγχυσης:

82	35
27	42

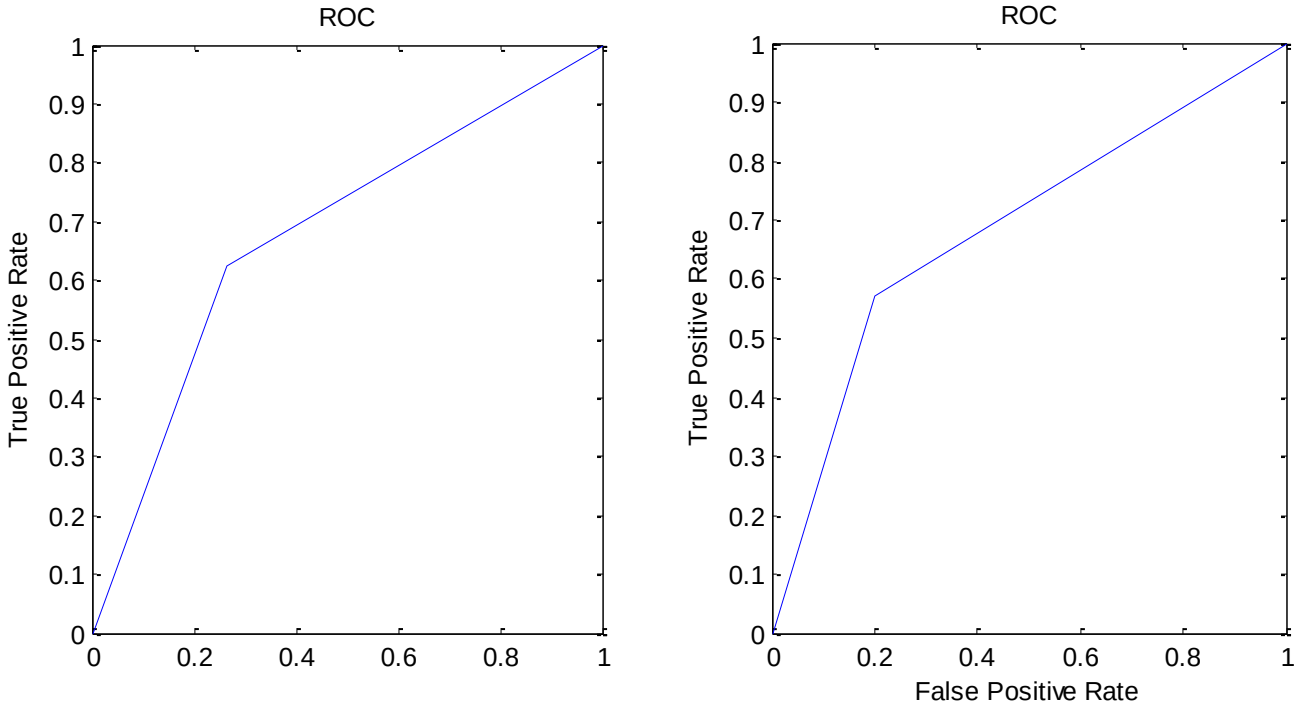
Ακρίβεια : 0.6785 ± 0.0320
Ευαισθησία : 0.7417 ± 0.0597
Προσδιοριστικότητα : 0.5890 ± 0.0735

– Μετά την επιλογή χαρακτηριστικών:

Πίνακας Σύγχυσης:

84	29
25	48

Ακρίβεια : 0.6457 ± 0.0960
Ευαισθησία : 0.7784 ± 0.2480
Προσδιοριστικότητα : 0.4578 ± 0.3382



Σχήμα 32: Καμπύλη ROC του κατηγοριοποιητή 5NN για τη διάγνωση από τομογραφία SPECT, αριστερά με χρήση όλων των χαρακτηριστικών και δεξιά μετά την επιλογή χαρακτηριστικών

Η εικόνα της κατηγοριοποίησης με χρήση της μεθόδου 5NN είναι παραπλήσια με αυτή των 3NN. Οι στατιστικοί δείκτες παραμένουν σε χαμηλά επίπεδα ενώ η επιλογή χαρακτηριστικών μειώνει τις επιδόσεις. (βλ. σχήμα 32)

5.5.3.3. Κατηγοριοποιητής 7NN

– Με όλα τα χαρακτηριστικά:

Πίνακας Σύγχυσης:

82	30
27	47

Ακρίβεια : 0.6769 ± 0.0323
Ευαισθησία : 0.7546 ± 0.0725
Προσδιοριστικότητα : 0.5669 ± 0.0784

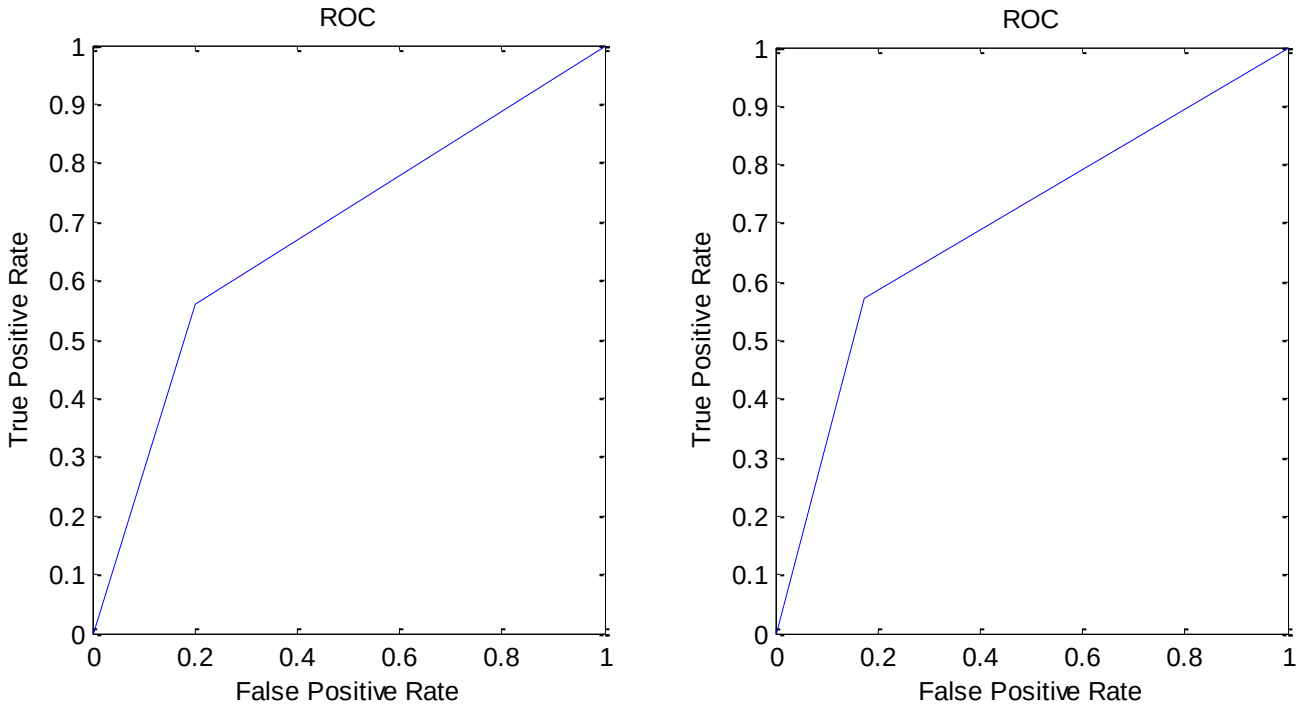
– Μετά την επιλογή χαρακτηριστικών:

Πίνακας Σύγχυσης:

85	29
24	48

Ακρίβεια : 0.7073 ± 0.0616
Ευαισθησία : 0.8092 ± 0.1427
Προσδιοριστικότητα : 0.5630 ± 0.2060

Στην περίπτωση αυτή η κατηγοριοποίηση πριν την επιλογή χαρακτηριστικών δεν μπορεί να κριθεί ικανοποιητική. Μετά την επιλογή χαρακτηριστικών όμως η ευαισθησία αυξάνεται σε πολύ καλά επίπεδα αυξάνοντας έτσι τη συνολική ακρίβεια. Αξίζει να σημειωθεί η αστάθεια που παρουσιάζεται στην προσδιοριστικότητα η οποία παρουσιάζει και πολύ χαμηλή μέση τιμή. (βλ. σχήμα 33)



Σχήμα 33: Καμπύλη ROC του κατηγοριοποιητή 7NN για τη διάγνωση από τομογραφία SPECT, αριστερά με χρήση όλων των χαρακτηριστικών και δεξιά μετά την επιλογή χαρακτηριστικών

5.5.4. Κατηγοριοποιητής SOM

– Με όλα τα χαρακτηριστικά:

Πίνακας Σύγχυσης:

131	45
26	65

Ακρίβεια : 0.7341
Ευαισθησία : 0.8344
Προσδιοριστικότητα : 0.5909

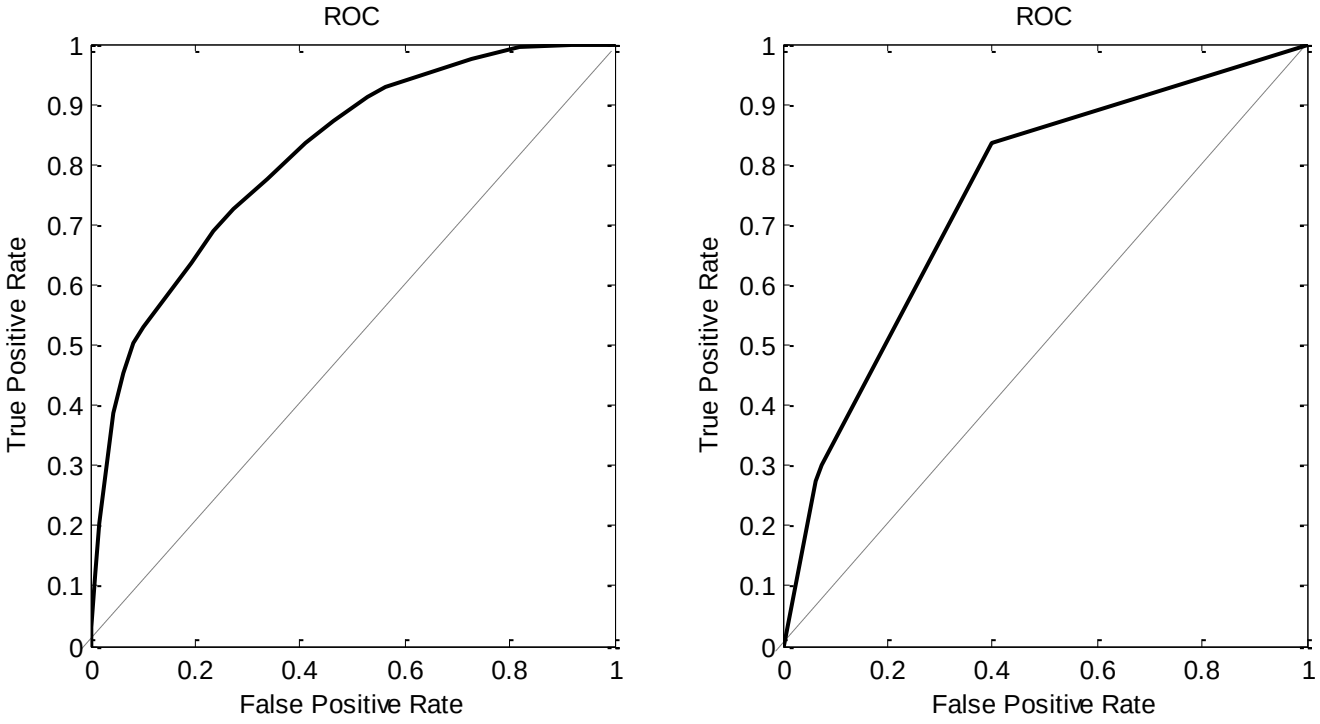
– Μετά την επιλογή χαρακτηριστικών:

Πίνακας Σύγχυσης:

131	44
26	66

Ακρίβεια : 0.7378
Ευαισθησία : 0.8344

Προσδιοριστικότητα : 0.6000



Σχήμα 34: Καμπύλη ROC του κατηγοριοποιητή SOM για τη διάγνωση από τομογραφία SPECT, αριστερά με χρήση όλων των χαρακτηριστικών και δεξιά μετά την επιλογή χαρακτηριστικών

Στην κατηγοριοποίηση με τη χρήση του αλγορίθμου SOM παρατηρείται η ίδια συμπεριφορά είτε με χρήση όλων των χαρακτηριστικών είτε μετά την επιλογή χαρακτηριστικών (σχήμα 34). Η τιμή της ευαισθησίας που προκύπτει είναι αρκετά υψηλή, η προσδιοριστικότητα όμως διατηρείται σε χαμηλά επίπεδα κι έτσι συνολικά η ακρίβεια είναι μετρίως ικανοποιητική.

5.5.5. Κατηγοριοποιητής FCM

– Με όλα τα χαρακτηριστικά:
Πίνακας Σύγχυσης:

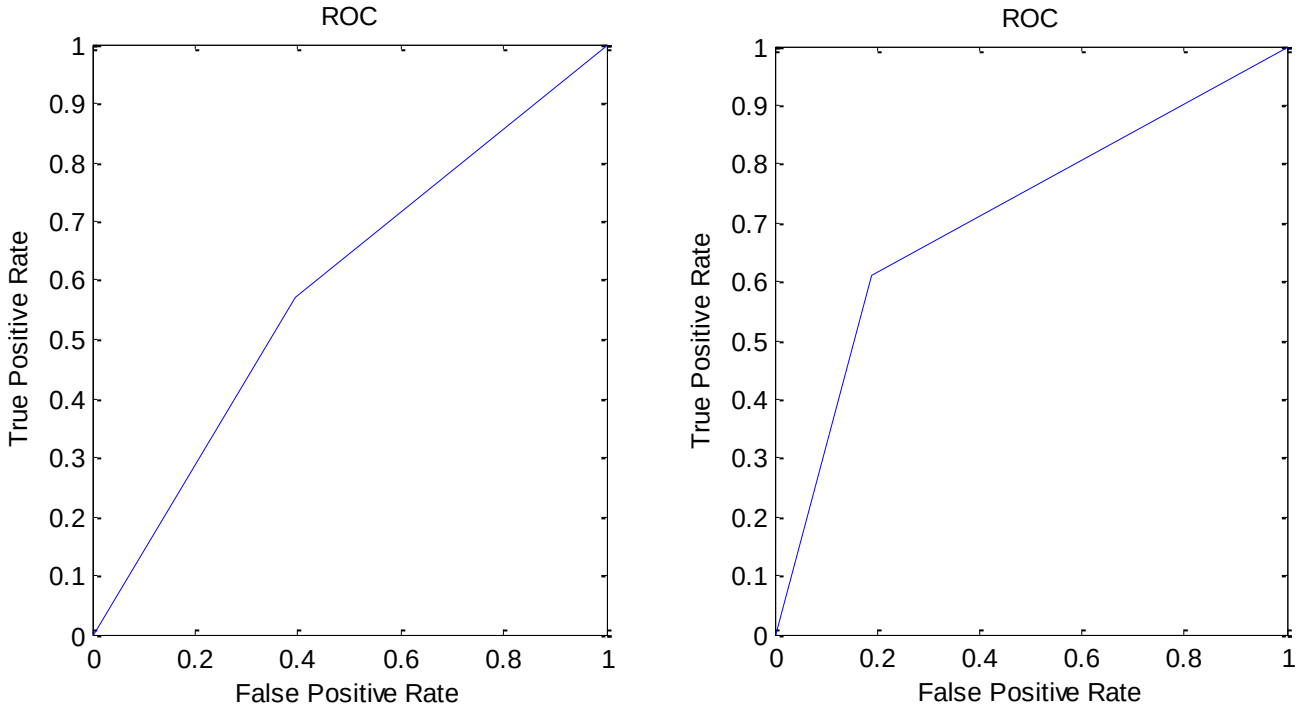
95	47
62	63

Ακρίβεια : 0.5918
Ευαισθησία : 0.6051
Προσδιοριστικότητα : 0.5727

– Μετά την επιλογή χαρακτηριστικών:
Πίνακας Σύγχυσης:

127	43
30	67

Ακρίβεια : 0.7266
Ευαισθησία : 0.8089
Προσδιοριστικότητα : 0.6091



Σχήμα 35: Καμπύλη ROC του κατηγοριοποιητή FCM για τη διάγνωση από τομογραφία SPECT, αριστερά με χρήση όλων των χαρακτηριστικών και δεξιά μετά την επιλογή χαρακτηριστικών

Στην περίπτωση αυτή η κατηγοριοποίηση με χρήση όλων των χαρακτηριστικών είναι κρίνεται ανεπαρκής καθώς οι τιμές των στατιστικών δεικτών είναι πολύ χαμηλές. Μετά την επιλογή χαρακτηριστικών επέρχεται σημαντική αύξηση της ακρίβειας του κατηγοριοποιητή, με την ευαισθησία να ανέρχεται σε αρκετά καλό επίπεδο, όμως η προσδιοριστικότητα παραμένει σε χαμηλά επίπεδα (βλ. σχήμα 35).

5.5.6. Αποτίμηση Κατηγοριοποιητών

Το υπό εξέταση σύνολο δεδομένων δεν κατηγοριοποιείται πολύ ικανοποιητικά από κανένα κατηγοριοποιητή. Μεγάλες αποκλίσεις στην ακρίβεια δεν υπάρχουν, ενώ η προσδιοριστικότητα δεν είναι καλή, οπότε μεγαλύτερο ενδιαφέρον έχει να ληφθεί η μεγαλύτερη δυνατή ευαισθησία. Σ' αυτό το πλαίσιο, ξεχωρίζουν οι αλγόριθμοι BK και SOM με τη μεγαλύτερη ευαισθησία και γενικότερα παραπλήσιους δείκτες, ενώ με μικρή διαφορά ακολουθεί ο SVM. Ελαφρά μικρότερες επιδόσεις παρουσιάζουν οι αλγόριθμοι FCM και 7NN ο οποίος επιλέγεται ως ο καλύτερος από τους αλγορίθμους kNN.

5.6. Διάγνωση Θυρεοειδούς

Το διαθέσιμο σύνολο δεδομένων περιλαμβάνει 841 διανύσματα με 21 χαρακτηριστικά, τα οποία μετά τη επιλογή χαρακτηριστικών μειώνονται σε 15.

5.6.1. Κατηγοριοποιητής ΒΚ

- Με όλα τα χαρακτηριστικά:

  Πίνακας Σύγχυσης:

699	76
14	52

Ακρίβεια : 0.9032 ± 0.0225
Ευαισθησία : 0.9852 ± 0.0074
Προσδιοριστικότητα : 0.4080 ± 0.1188

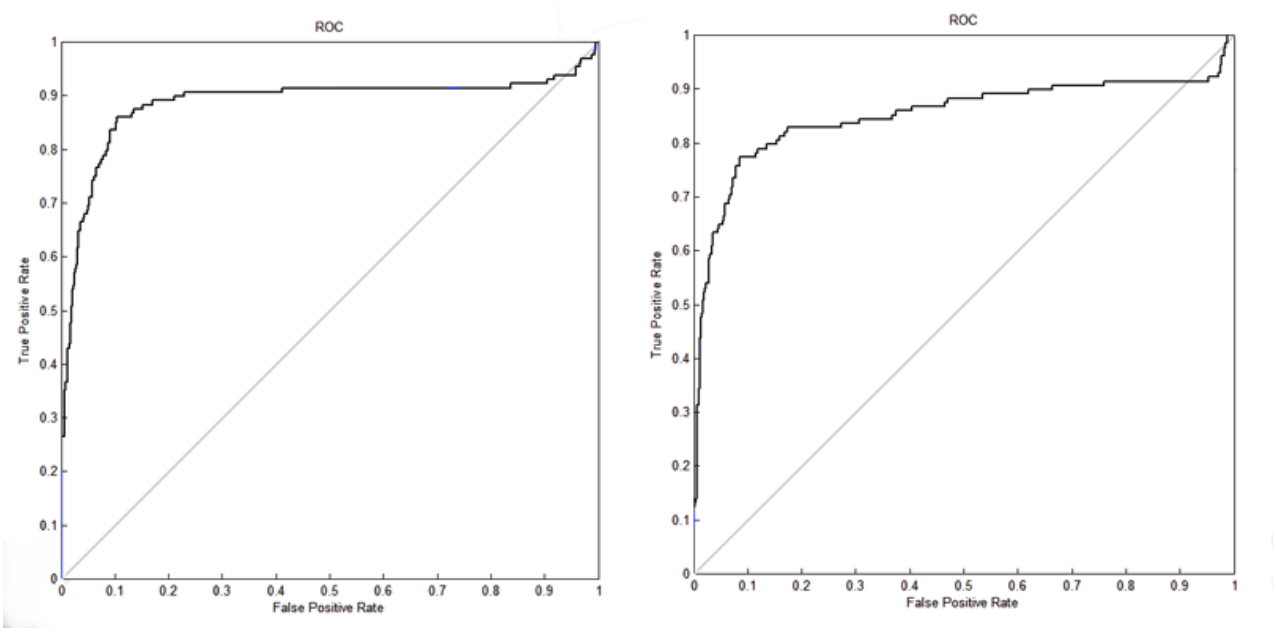
- Μετά την επιλογή χαρακτηριστικών:

  Πίνακας Σύγχυσης:

641	67
72	61

Ακρίβεια : 0.8983 ± 0.0216
Ευαισθησία : 0.9784 ± 0.0103
Προσδιοριστικότητα : 0.4518 ± 0.1118

Στην περίπτωση αυτή αξιοσημείωτη είναι η πολύ υψηλή ευαισθησία, σε αντίθεση με τη χαμηλή προσδιοριστικότητα η οποία έχει παρουσιάζει και μεγάλη τυπική απόκλιση. Παρά τη χαμηλή προσδιοριστικότητα η ακρίβεια είναι πολύ μεγάλη λόγω του γεγονότος ότι τα πρότυπα της κλάσης των ασθενών είναι πολλαπλάσια σε σχέση με εκείνα των υγιών. Η επιλογή χαρακτηριστικών αφήνει ανεπηρέαστους σχεδόν τους δείκτες, με εξαίρεση μια μικρή αύξηση της μέσης τιμής της προσδιοριστικότητας η οποία όμως δεν προσθέτει στην ικανότητα διάκρισης της κλάσης των «υγιών» (σχήμα 36).



Σχήμα 36: Καμπύλη ROC του κατηγοριοποιητή ΒΚ για τη διάγνωση του θυρεοειδούς αριστερά με χρήση όλων των χαρακτηριστικών και δεξιά μετά την επιλογή χαρακτηριστικών

5.6.2. Κατηγοριοποιητής SVM

Η κατηγοριοποίηση γίνεται με χρήση γραμμικού πυρήνα.

- Με όλα τα χαρακτηριστικά:

  Πίνακας Σύγχυσης:

341	24
15	40

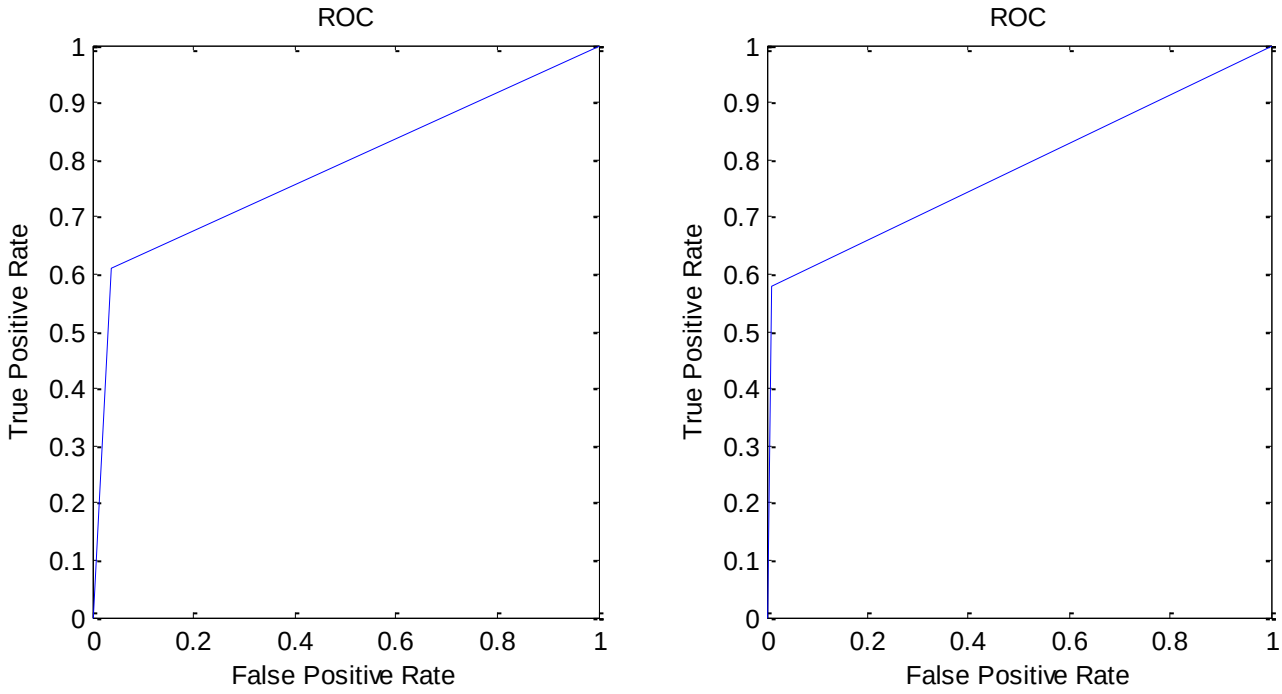
Ακρίβεια : 0.9169 ± 0.0126
Ευαισθησία : 0.9711 ± 0.0139
Προσδιοριστικότητα : 0.6156 ± 0.0590

- Μετά την επιλογή χαρακτηριστικών:

  Πίνακας Σύγχυσης:

351	28
5	36

Ακρίβεια : 0.9207 ± 0.0081
Ευαισθησία : 0.9834 ± 0.0073
Προσδιοριστικότητα : 0.5719 ± 0.0511



Σχήμα 37: Καμπύλη ROC του κατηγοριοποιητή SVM για τη διάγνωση του θυρεοειδούς, αριστερά με χρήση όλων των χαρακτηριστικών και δεξιά μετά την επιλογή χαρακτηριστικών

Η εφαρμογή της κατηγοριοποίησης SVM έχει ως αποτέλεσμα πολύ καλή ακρίβεια στις προβλέψεις του κατηγοριοποιητή με αιχμή τη πολύ μεγάλη τιμή ευαισθησίας αφού η προσδιοριστικότητα είναι πολύ χαμηλή. Η επιλογή χαρακτηριστικών επιφέρει μια μικρή περεταίρω αύξηση στην ευαισθησία και κατά συνέπεια και στην ακρίβεια, η οποία όμως συνοδεύεται από μικρή πτώση της προσδιοριστικότητας σε ακόμη χαμηλότερο επίπεδο. (σχήμα 37)

5.6.3. Κατηγοριοποιητής kNN

5.6.3.1. Κατηγοριοποιητής 3NN

– <u>Με όλα τα χαρακτηριστικά</u>:
Πίνακας Σύγχυσης:

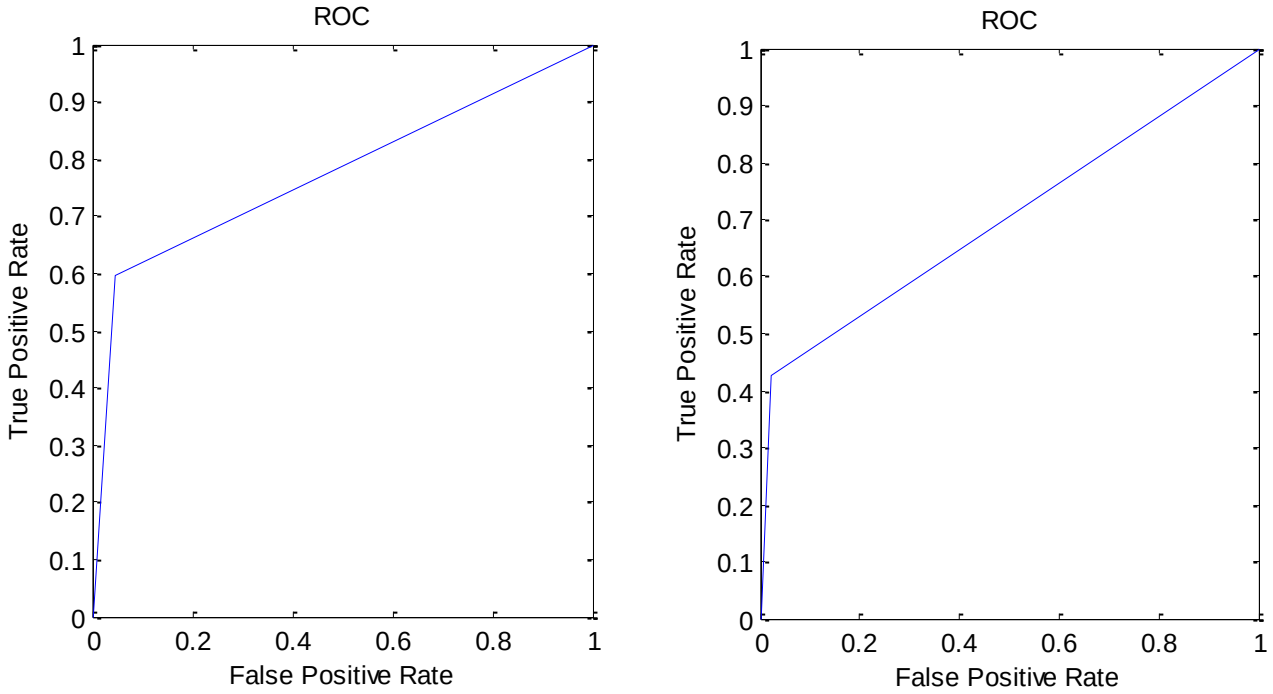
477	36
22	53

Ακρίβεια : 0.9071 ± 0.0082
Ευαισθησία : 0.9629 ± 0.0086
Προσδιοριστικότητα : 0.5938 ± 0.0346

– <u>Μετά την επιλογή χαρακτηριστικών</u>:
Πίνακας Σύγχυσης:

483	48
16	41

Ακρίβεια : 0.8929 ± 0.0126
Ευαισθησία : 0.9748 ± 0.0151
Προσδιοριστικότητα : 0.4337 ± 0.0504



Σχήμα 38: Καμπύλη ROC του κατηγοριοποιητή 3NN για τη διάγνωση του θυρεοειδούς, αριστερά με χρήση όλων των χαρακτηριστικών και δεξιά μετά την επιλογή χαρακτηριστικών

Ο κατηγοριοποιητής 3NN εξάγει πολύ καλής ακρίβειας αποτελέσματα, γεγονός που οφείλεται στην πολύ μεγάλη τιμή ευαισθησίας, παρά τη χαμηλή προσδιοριστικότητα. Μετά την επιλογή χαρακτηριστικών η ευαισθησία παρουσιάζει μικρή αύξηση, όμως η προσδιοριστικότητα μειώνεται σημαντικά όπως φαίνεται στη δεύτερη καμπύλη ROC του σχήματος 38.

5.6.3.2. Κατηγοριοποιητής 5NN

– Με όλα τα χαρακτηριστικά:
Πίνακας Σύγχυσης:

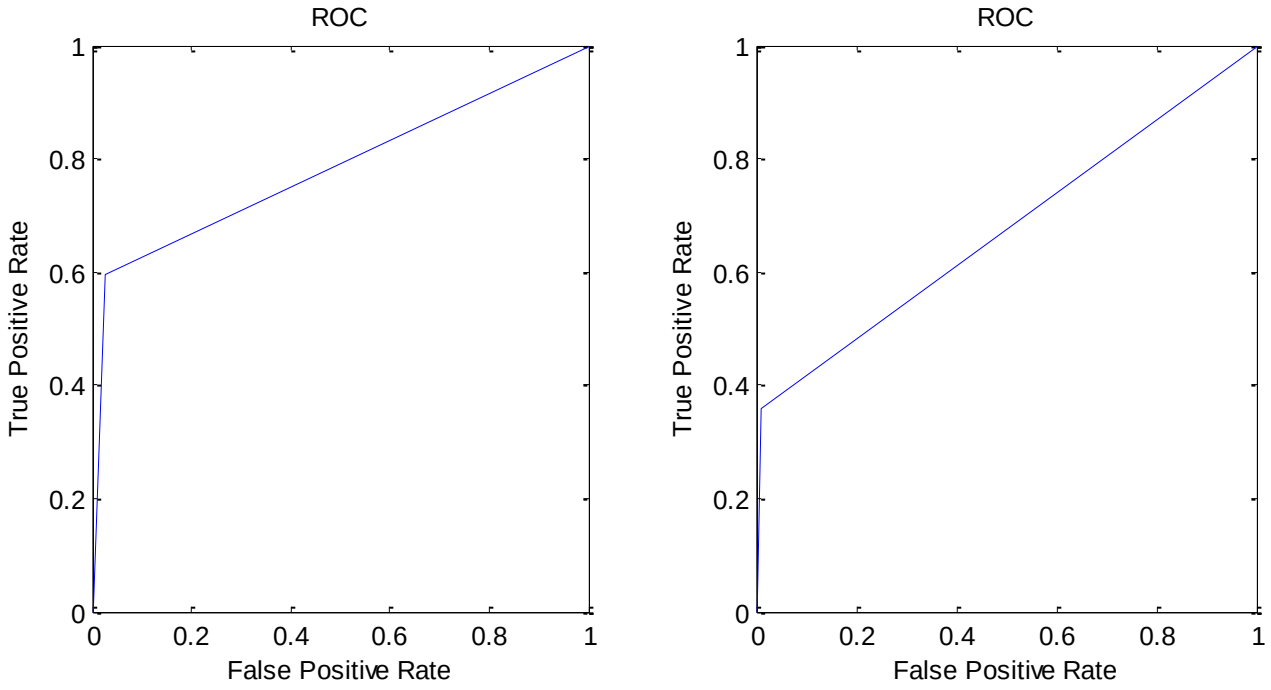
485	37
14	52

Ακρίβεια : 0.9133 ± 0.0067
Ευαισθησία : 0.9707 ± 0.0109
Προσδιοριστικότητα : 0.5910 ± 0.0555

– Μετά την επιλογή χαρακτηριστικών:
Πίνακας Σύγχυσης:

492	56
7	33

Ακρίβεια : 0.8937 ± 0.0079
Ευαισθησία : 0.9850 ± 0.0096
Προσδιοριστικότητα : 0.3820 ± 0.0408



Σχήμα 39: Καμπύλη ROC του κατηγοριοποιητή 5NN για τη διάγνωση του θυρεοειδούς, αριστερά με χρήση όλων των χαρακτηριστικών και δεξιά μετά την επιλογή χαρακτηριστικών

Η εικόνα της κατηγοριοποίησης είναι παραπλήσια με αυτή της κατηγοριοποίησης 3NN, με πολύ υψηλή ευαισθησία αλλά κακή προσδιοριστικότητα, η οποία μετά την επιλογή χαρακτηριστικών φθάνει σε ακόμα χαμηλότερα επίπεδα απ' ό,τι στην προηγούμενη περίπτωση συμπαρασύροντας τη συνολική ακρίβεια (βλ. σχήμα 39).

5.6.3.3. Κατηγοριοποιητής 7NN

– Με όλα τα χαρακτηριστικά:
Πίνακας Σύγχυσης:

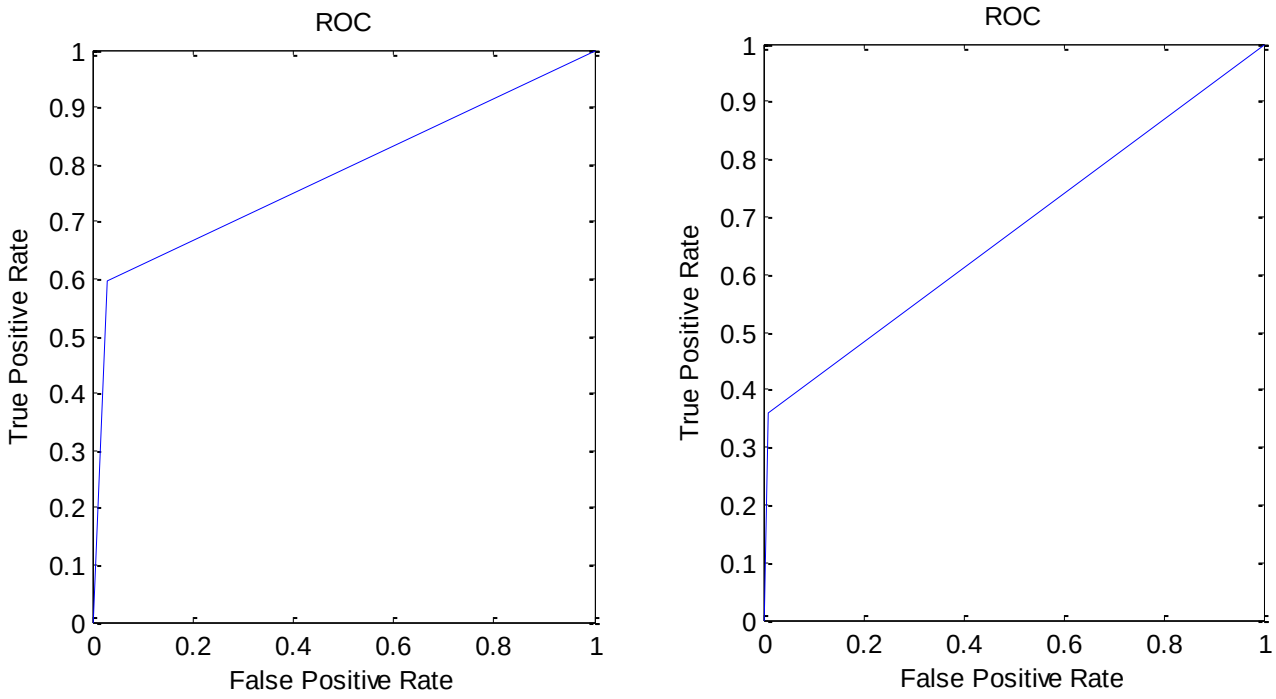
486	41
13	48

Ακρίβεια : 0.9084 ± 0.0065
Ευαισθησία : 0.9714 ± 0.0075
Προσδιοριστικότητα : 0.5551 ± 0.0450

– Μετά την επιλογή χαρακτηριστικών:
Πίνακας Σύγχυσης:

495	60
4	29

Ακρίβεια : 0.8943 ± 0.0094
Ευαισθησία : 0.9917 ± 0.0051
Προσδιοριστικότητα : 0.3483 ± 0.0537



Σχήμα 40: Καμπύλη ROC του κατηγοριοποιητή 7NN για τη διάγνωση του θυρεοειδούς, αριστερά με χρήση όλων των χαρακτηριστικών και δεξιά μετά την επιλογή χαρακτηριστικών

Η εικόνα της κατηγοριοποίησης 7NN είναι ίδια με αυτές που προέκυψαν από την κατηγοριοποίηση 3NN και 5NN, δηλαδή λαμβάνεται πολύ υψηλή ευαισθησία και κακή προσδιοριστικότητα, η οποία μετά την επιλογή χαρακτηριστικών φθάνει σε ακόμα χαμηλότερα επίπεδα με συνέπεια τη μείωση της συνολικής ακρίβειας (βλ. σχήμα 40).

5.6.4. Κατηγοριοποιητής SOM

– Με όλα τα χαρακτηριστικά:

Πίνακας Σύγχυσης:

688	47
25	81

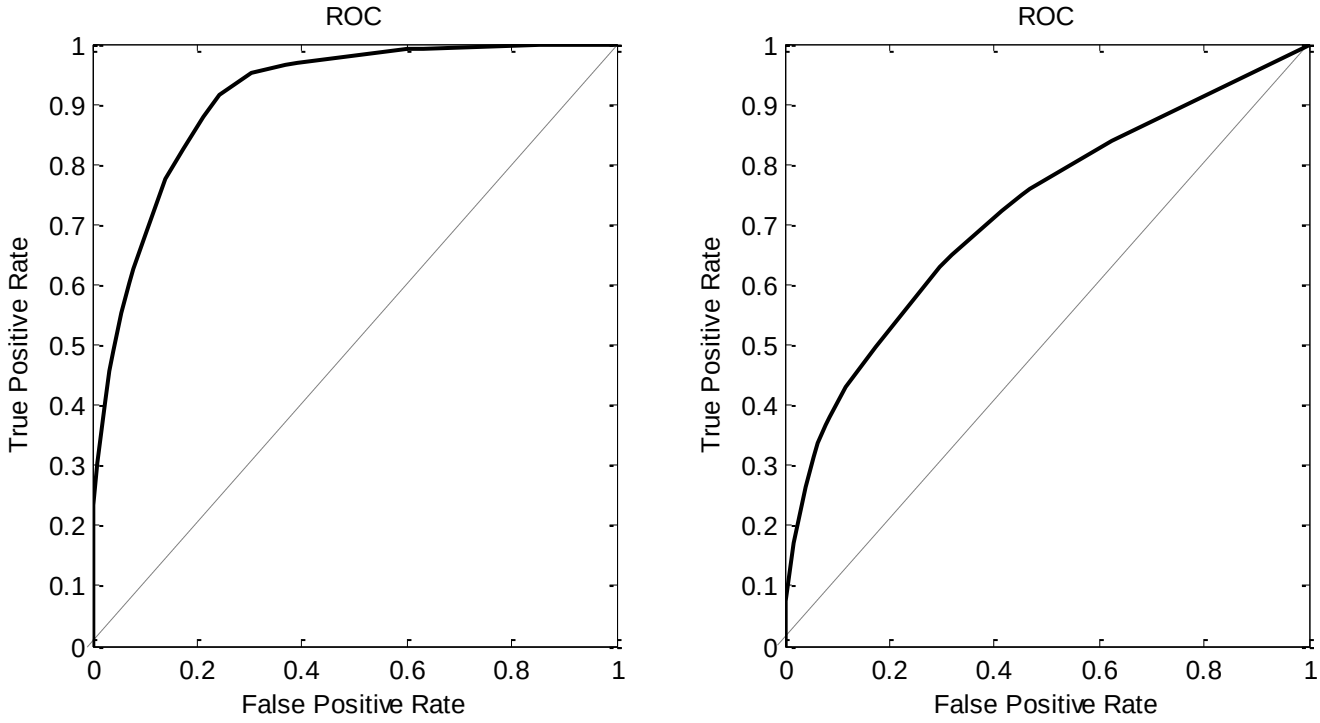
Ακρίβεια : 0.9144
Ευαισθησία : 0.9649
Προσδιοριστικότητα : 0.6328

– Μετά την επιλογή χαρακτηριστικών:

Πίνακας Σύγχυσης:

713	128
0	0

Ακρίβεια : 0.8478
Ευαισθησία : 1
Προσδιοριστικότητα : 0



Σχήμα 41: Καμπύλη ROC του κατηγοριοποιητή SOM για τη διάγνωση του θυρεοειδούς, αριστερά με χρήση όλων των χαρακτηριστικών και δεξιά μετά την επιλογή χαρακτηριστικών

Η εφαρμογή της μεθόδου SOM αποδίδει πολύ καλή ευαισθησία και κατά συνέπεια πολύ καλή ακρίβεια παρά την όχι τόσο ικανοποιητική προσδιοριστικότητα αφού τα πρότυπα της κλάσης των ασθενών είναι πολλαπλάσια των υγιών. Μετά την επιλογή χαρακτηριστικών όμως όλα τα πρότυπα κατηγοριοποιούνται στην κλάση των ασθενών με αποτέλεσμα η ευαισθησία να είναι απόλυτη και η προσδιοριστικότητα μηδενική. Κατά συνέπεια, όπως φαίνεται στο σχήμα 41, και η ακρίβεια της κατηγοριοποίησης μειώνεται αντιστοίχως.

5.6.5. Κατηγοριοποιητής FCM

- Με όλα τα χαρακτηριστικά:

Πίνακας Σύγχυσης:

406	21
307	107

Ακρίβεια : 0.6100
Ευαισθησία : 0.5694
Προσδιοριστικότητα : 0.8359

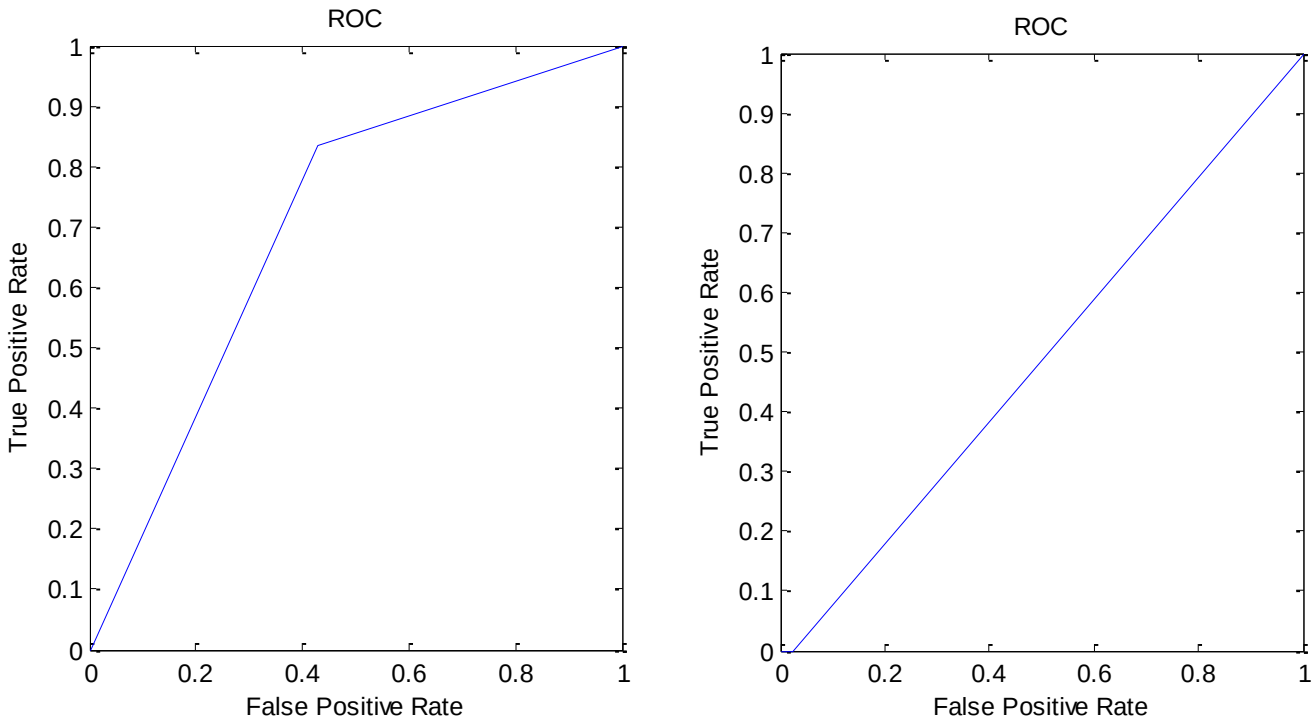
- Μετά την επιλογή χαρακτηριστικών:

Πίνακας Σύγχυσης:

696	128
17	0

Ακρίβεια : 0.8276
Ευαισθησία : 0.9762
Προσδιοριστικότητα : 0

Εδώ παρατηρείται μια ιδιαίτερη περίπτωση όπου η επιλογή χαρακτηριστικών αλλάζει άρδην την εικόνα της κατηγοριοποίησης. Από την κατηγοριοποίηση με χρήση όλων των χαρακτηριστικών προκύπτει αρκετά υψηλή προσδιοριστικότητα ενώ η χαμηλή ευαισθησία κρατάει την ακρίβεια σε χαμηλά επίπεδα επίσης. Μετά την επιλογή χαρακτηριστικών κανένα αρνητικό πρότυπο δεν ταξινομείται σωστά, προκύπτει δηλαδή μηδενική προσδιοριστικότητα ενώ η ευαισθησία είναι πολύ υψηλή σε πλήρη αντίθεση με την κατηγοριοποίηση με όλα τα χαρακτηριστικά.



Σχήμα 41: Καμπύλη ROC του κατηγοριοποιητή FCM για τη διάγνωση του θυρεοειδούς, αριστερά με χρήση όλων των χαρακτηριστικών και δεξιά μετά την επιλογή χαρακτηριστικών

5.6.6. Αποτίμηση Κατηγοριοποιητών

Η συγκεκριμένη περίπτωση είναι λίγο ιδιαίτερη καθώς παρά τη γενική τάση μεταξύ των κατηγοριοποιητών για μεγάλη ευαισθησία και μικρή προσδιοριστικότητα, η ταξινόμηση με τη χρήση του αλγορίθμου FCM παρουσιάζει μεγάλη προσδιοριστικότητα και μέτρια ευαισθησία. Μετά την επιλογή χαρακτηριστικών όμως ακολουθεί τη συμπεριφορά των άλλων κατηγοριοποιητών. Έτσι, θα μπορούσε να χρησιμοποιηθεί η ειδική περίπτωση του FCM σε συνδυασμό με τον κατηγοριοποιητή με την καλύτερη ευαισθησία από τους υπόλοιπους ώστε να προβλέπονται και οι δύο κλάσεις με πολύ καλή πιθανότητα. Αν ένα πρότυπο ταξινομηθεί από τον FCM ως ασθενής τότε με μεγάλη πιθανότητα αυτό θα ανήκει στην κλάση αυτή ανεξάρτητα από την έξοδο του έτερου κατηγοριοποιητή. Αντίστροφα, αν ένα πρότυπο ταξινομηθεί από τον έτερο κατηγοριοποιητή στην κλάση υγιής ανεξάρτητα από την έξοδο του FCM αυτό θα ανήκει με μεγάλη πιθανότητα στην κλάση αυτή. Όσον αφορά στον κατηγοριοποιητή με την καλύτερη ευαισθησία αυτός είναι ο SOM με τους υπόλοιπους κατηγοριοποιητές να ακολουθούν με κοντινές τιμές ευαισθησίας.

6. ΣΥΝΟΛΙΚΗ ΑΠΟΤΙΜΗΣΗ

Κάνοντας έναν συνολικό απολογισμό των κατηγοριοποιητών που χρησιμοποιήθηκαν, έγινε σαφές ότι δεν μπορεί να χαρακτηριστεί κάποιος κατηγοριοποιητής ως ο καλύτερος όλων. Κάθε σύνολο δεδομένων ανάλογα με το πώς είναι κατανεμημένα τα πρότυπα των δύο κλάσεων και αν υπάρχουν επικαλύψεις. Η κατανομή αυτή, εκτός από την απόδοση των κατηγοριοποιητών, καθορίζει και τα κριτήρια αξιολόγησης των κατηγοριοποιητών όπως για παράδειγμα στην περίπτωση που η ευαισθησία προκύπτει υψηλή και η προσδιοριστικότητα χαμηλη ή το αντίστροφο. Στην περίπτωση αυτή η μία από τις δύο κλάσεις είναι πιο καλά κατανεμημένη -υπό την έννοια ότι τα πρότυπα της δεν ξεφεύγουν από τα όρια της κλάσης ώστε να βρεθούν μεταξύ προτύπων της άλλης κλάσης- ενώ η άλλη κλάση παρουσιάζει συσχέτιση με την πρώτη. Υπό αυτές τις συνθήκες αντί της μεγάλης ακρίβειας μεγαλύτερη πρακτική χρησιμότητα έχει η μεγιστοποίηση της ευαισθησίας (ή της προσδιοριστικότητας), καθώς μπορεί να εξαχθεί ασφαλές συμπέρασμα για τη μία από τις δύο κλάσεις. Έτσι, αν ένα πρότυπο καταταχθεί στην κλάση «ασθενής» («υγιής») τότε η υψηλή ευαισθησία (προσδιοριστικότητα) εξασφαλίζει με μεγάλη πιθανότητα ότι το πρότυπο ανήκει όντως στην κλάση «ασθενής» («υγιής»).

Υπό αυτό το πρίσμα, τις καλύτερες επιδόσεις κατέγραψε ο αλγόριθμος Οπισθοδιάδοσης (BK) ο οποίος σε πολλές περιπτώσεις παρουσίασε την καλύτερη συμπεριφορά από όλους τους κατηγοριοποιητές ενώ και στις υπόλοιπες περιπτώσεις η απόδοσή του ήταν πολύ κοντά στον καλύτερο. Σαφή υστέρηση σε σχέση με τους υπόλοιπους κατηγοριοποιητές παρουσίασε στην πλειονότητα των περιπτώσεων ο αλγόριθμος Ασαφούς C-Μέσου (FCM). Η απόκλισή του από τον καλύτερο κατηγοριοποιητή ήταν σε μερικές περιπτώσεις μεγάλη ενώ σε άλλες πιο κοντά στην κυρίαρχη τάση. Οι υπόλοιποι κατηγοριοποιητές κυμάνθηκαν στα ίδια επίπεδα στο ενδιάμεσο διάστημα μεταξύ των δύο άκρων, με τη Μηχανή Διανυσμάτων Υποστήριξης (SVM) να προηγείται λίγο μεταξύ τους, και η μέθοδος Αυτο-Οργανούμενων Χαρτών (SOM) και την οικογένεια των αλγορίθμων k-Κοντινότερων Γειτόνων (kNN) να ακολουθούν. Όσον αφορά στην οικογένεια kNN, οι διαφορές μεταξύ των τριών εκδοχών στις περισσότερες περιπτώσεις ήταν πολύ μικρές, υπέρ της εκδοχής των 7-Κοντινότερων Γειτόνων συνήθως. Αν όμως το σύστημα της κατηγοριοποίησης επιβάλλει γρήγορους υπολογισμούς τότε η εκδοχή 3-Κοντινότερων Γειτόνων είναι πολύ ικανοποιητική δεδομένης της μικρής –αν υπάρχει- διαφοράς στην απόδοση.

Η επιλογή χαρακτηριστικών, πλην ελαχίστων εξαιρέσεων, βελτίωσε την απόδοση των κατηγοριοποιητών. Σε περιπτώσεις καλής κατηγοριοποίησης η βελτίωση αυτή δεν ήταν ιδιαίτερα θεαματική, ήταν όμως σταθερή και για κάθε δεδομένο σύνολο δεδομένων παρουσιάστηκε σε όλους τους κατηγοριοποιητές. Σημειώνεται ότι η μέθοδος SFFS που χρησιμοποιήθηκε δεν είναι παράγει το βέλτιστο υποσύνολο χαρακτηριστικών και δεδομένων των ιδιαιτεροτήτων των συνόλων δεδομένων η μικρή μόνο βελτίωση είναι αναμενόμενη.

Βιβλιογραφία

J. Bezdek, R. Ehrlich & W. Full. The Fuzzy *c*-Means Clustering Algorithm. *Computers & Geosciences Vol. 10*, 2-3:191-203, 1984.

C. Cortes & V. Vapnik. Support-vector network. *Machine Learning*, 20:273-297, 1995.

A. Frank & A. Asuncion. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, 2010

M. Goldstein. k_n-Nearest Neighbor Classification. *IEEE Transactions Theory, Vol. IT-18*, No. 5, 1972.

S. Haykin. Neural Networks and Machine Learning, Third Edition. Pearson, 2009

T. Kohonen. Self-Organizing Maps. Springer, Berlin, 1995.

M. Little, P. McSharry, S. Roberts, D. Costello, I Moroz. Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection. *BioMedical Engineering OnLine 2007*, 6:23, 2007

R. Lippmann. An introduction to computing with neural nets. *IEEE ASSP Magazine*, 4:4-22, 1987

H. Peng. Feature Selection Based on Mutual Information Criteria of Max-Dependency, Max- Relevance and Min-Redundancy, *IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27,* No. 8, 2005.

P. Pudil, J. Novivicova & J.Kittler. Floating search methods in feature slection. *Pattern Racognition Letters 15*, 1119-1125, 1994.

F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the Brain. *Psychological Review*, 65:386-408, 1958.

T. Theodoridis & K. Koutroumbas. Pattern Recognition, Third Edition, Elsevier, 2006.

V. Vapnik & A. Chervonenkis. The Theory of Pattern Recognition. Nauka, Moscow, 1974.

J. Vesanto & E.Alhoniemi. Clustering of the Self-Organising Map, *IEEE Transactions on Neural Networks, Vol. 11*, No. 3, 586-600, 2000.