ΠΑΝΕΠΙΣΤΗΜΙΟ ΣΤΕΡΕΑΣ ΕΛΛΑΔΑΣ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΙΑΤΡΙΚΗ

Εφαρμογή Αλγορίθμων βασισμένων σε χαρακτηριστικά του Ανοσοποιητικού Συστήματος για την κατηγοριοποίηση βιοϊατρικών δεδομένων.

Πτυχιακή Εργασία

Κυράννα Τζαφέρη

Επιβλέπων:
Κων/νος Δελήμπασης

Λαμία, 2010

ΠΑΝΕΠΙΣΤΗΜΙΟ ΣΤΕΡΕΑΣ ΕΛΛΑΔΑΣ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΙΑΤΡΙΚΗ

Εφαρμογή Αλγορίθμων βασισμένων σε χαρακτηριστικά του Ανοσοποιητικού Συστήματος για την κατηγοριοποίηση βιοϊατρικών δεδομένων

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Επιβλέπων :

Κων/νος Δελήμπασης

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή
Λαμία,

.....	.....	.....
Κων/νος Δελήμπασης,	Βασίλειος Πλαγιανάκος,	Αντώνιος Αλετράς

Λαμία, 2010

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον επιβλέπων καθηγητή μου, Δελήμπαση Κωνσταντίνο, που με τις επιστημονικές του γνώσεις με βοήθησε να ολοκληρώσω την διπλωματική μου εργασία, για την οποία εργάστηκα από τον Δεκέμβριο του 2008 έως και τον Ιούλιο του 2010.

Ακόμη, θα ήθελα να ευχαριστήσω την οικογένεια μου, αλλά και τους φίλους μου για την κατανόηση και την ψυχολογική τους υποστήριξη η οποία ήταν απαραίτητη για την πραγματοποίηση της εργασίας αυτής.

Τέλος, θέλω να εκφράσω τις ευχαριστίες μου σε φίλους συμφοιτητές και πλέον συναδέλφους για την βοήθεια τους, την οποία μου έδωσαν απλόχερα όταν τη χρειάστηκα.

Περίληψη

Σκοπός της παρούσας πτυχιακής εργασίας είναι η διερεύνηση της χρησιμότητας ενός σχετικά νέου αλγόριθμου κατηγοριοποίησης με μάθηση, βασισμένου στις αρχές λειτουργίας του ανοσοποιητικού συστήματος. Καθώς ο αλγόριθμος δεν έχει ακόμη καθιερωθεί μεταξύ των πολύ διαδεδομένων μεθόδων κατηγοριοποίησης, στόχος μας είναι η κατανόηση της λειτουργίας του, η διερεύνηση της εξάρτησης του από τις κυριότερες παραμέτρους του και η δυνατότητα του να αντιμετωπίσει ένα δύσκολο κλινικό πρόβλημα κατηγοριοποίησης δεδομένων βιοψίας λεπτής βελόνας θυρεοειδούς. Όλα τα παραπάνω γίνονται σε σύγκριση με δύο διαδεδομένες μεθόδους εποπτευόμενης κατηγοριοποίησης δεδομένων.

Για την ολοκλήρωση της πτυχιακής αυτής χρησιμοποιήθηκαν α) συνθετικά δεδομένα κατασκευασμένα από εμάς, β) δεδομένα διαθέσιμα από βάσεις δεδομένων και γ) πραγματικά κλινικά δεδομένα, τα οποία είναι δείγματα από δειγματοληψία λεπτής βελόνας για να ελεγθεί εάν υπάρχει καρκίνος στον θυρεοειδή αδένα. Τα τελευταία δεδομένα, προέρχονται από το Α’ τμήμα παθολογίας της ιατρικής σχολής Αθηνών, κατά το χρονικό διάστημα 2000-2004.

Αρχικά περιγράφεται το πρόβλημα του καρκίνου του θυρεοειδούς και οι διαστάσεις του. Περιγράφεται επίσης η έννοια της κατηγοριοποίησης, καθώς και η έννοια των ευριστικών αλγορίθμων.

Έπειτα περιγράφονται τα δεδομένα και δύο υπό σύγκριση μέθοδοι κατηγοριοποίησης, περισσότερο διαδεδομένοι από την προτεινόμενη.

Γίνεται περιγραφή του βιολογικού ανοσοποιητικού συστήματος και παρουσιάζονται οι ομοιότητες και οι αναλογίες του αλγορίθμου τεχνητού ανοσοποιητικού συστήματος με αυτό.

Τελικά, παρουσιάζονται τα αποτελέσματα από την χρήση του κάθε αλγόριθμου σε όλα τα διαθέσιμα δεδομένα και γίνονται οι απαραίτητοι σχολιασμοί. Ιδιαίτερη σημασία δίνεται στα αποτελέσματα των μεθόδων στα κλινικά δεδομένα, τα οποία για να μελετηθούν διεξοδικά χωρίζονται σε μη ύποπτα και ύποπτα, βάσει των αποτελεσμάτων της βιοψίας λεπτής βελόνας. Παρουσιάζονται γενικά συμπεράσματα σχετικά με τον υπό έρευνα αλγόριθμο, αλλά και ειδικά σε σχέση με το συγκεκριμένο κλινικό πρόβλημα, το οποίο μπορεί να περιγραφεί ως βελτίωση της διάγνωσης του καρκίνου του θυρεοειδούς, βάσει των αποτελεσμάτων της βιοψίας λεπτής βελόνας.

Λέξεις Κλειδιά: καρκίνος του θυρεοειδούς, βιοψία λεπτής βελόνας, κατηγοριοποίηση δεδομένων, τεχνητό ανοσοποιητικό σύστημα.

Abstract

The aim of this dissertation is the assessment of the usefulness of a relatively new classification algorithm with supervision that is based on the principles of the mammalian immune system, called AIRS. As AIRS may not be considered yet a well established classification algorithm, our goals include: a) the understanding of its operation, b) the investigation of the relation between the achieved classification accuracy and the parameters of the algorithmic settings and c) is usefulness in tackling a difficult classification problem, with clinical data coming from thyroid Fine Needle Aspiration. The aforementioned goals are pursued in comparison with two other well established classification methods.

In this study we used a) synthetic data generated by us, b) available data from databases for testing feature selection and classification algorithms and c) clinical data from a number of subjects taken from Fine Needle Aspiration. These data were originally acquired by A' Department of pathology of medical University of Athens.

Firstly we describe the problem of thyroid cancer as well as the concept of classification and the concept of heuristic and evolutionary algorithms. Then follows the description of the data and the two wide spread supervised classification methods in comparison. Finally, the biological immune system is briefly described, the artificial immune algorithm is presented in details and the analogies between them are outlined.

Finally, we present the results from the application of each algorithm on the available data and comments are underlined. Special attention is drown to the results of all the classification method considering the clinical data, which were separated into non-suspicious and suspicious data, based on the results of the FNA examination. General conclusions are drawn considering the AIRS algorithm, as well as specific conclusions considering the applicability of classification techniques to the clinical problem of improving thyroid cancer diagnosis based on the results of the FNA.

Finally, the datasets are tested with a number of classification techniques and the accuracy of the subject classification is measured. All the results are presented to lead us to a point.

Key Words: thyroid cancer, fine needle aspiration, data classification, artificial immune system

Περιεχόμενα

1 Περιγραφή κλινικού προβλήματος

Εισαγωγή

Στο κεφάλαιο αυτό γίνεται μια εισαγωγή στο κλινικό πρόβλημα που θα μελετήσουμε και στον τρόπο με τον οποίο γίνεται διάγνωση για τον καρκίνο του θυρεοειδούς. Γίνεται επίσης και μια εισαγωγή τόσο στην έννοια της ταξινόμησης ή κατηγοριοποίησης, όσο και στους μιμητικούς αλγόριθμους πάνω στους οποίους βασίζεται και ο αλγόριθμος που μελετάμε.

1.1 Ο καρκίνος του θυρεοειδούς

1.1.1 Ο θυρεοειδής αδένας και οι συνέπειες της δυσλειτουργίας του

Ο θυρεοειδής είναι ένας αδένας σε σχήμα πεταλούδας και βρίσκεται στη βάση του λαιμού, μπροστά από την τραχεία και αποτελεί τον «κύριο ρυθμιστή» του μεταβολισμού, βλέπε Εικόνα 1. Εάν ο θυρεοειδής υπολειτουργεί, παράγει πολύ μικρή ποσότητα θυρεοειδικής ορμόνης, προκαλώντας μια κατάσταση που ονομάζεται υποθυρεοειδισμός. Οι άνθρωποι με υποθυρεοειδισμό χρησιμοποιούν την ενέργεια με αργούς ρυθμούς και επιβραδύνεται έτσι ο μεταβολισμός τους. Αντίθετα, εάν ο θυρεοειδής υπερλειτουργεί, ο αδένας απελευθερώνει πολύ μεγάλη ποσότητα θυρεοειδικής ορμόνης στην κυκλοφορία του αίματος, προκαλώντας μία κατάσταση που ονομάζεται υπερθυρεοειδισμός η οποία επιταχύνει τον μεταβολισμό.

Η διάγνωση για τις παραπάνω περιπτώσεις γίνεται εύκολα με μια εξέταση αίματος σε ένα ιατρικό κέντρο, όπου ελέγχονται τα επίπεδα των θυρεοειδικών ορμονών του ασθενούς. Και φυσικά, υπάρχουν αποτελεσματικές θεραπείες που υποκαθιστούν ή μειώνουν τα επίπεδα των θυρεοειδικών ορμονών επιτρέποντας στους ασθενείς να ρυθμίσουν την κατάστασή τους και να ζήσουν φυσιολογική ζωή.

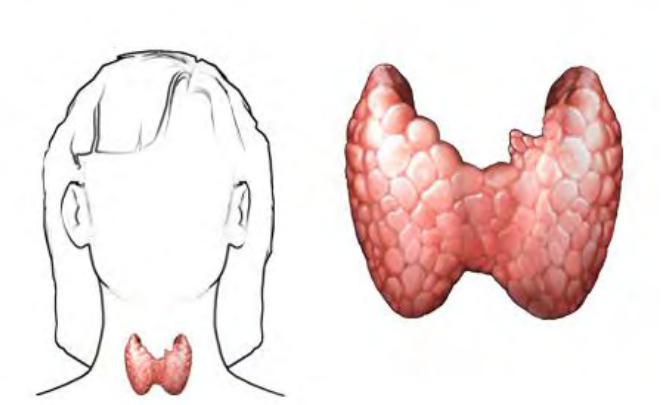
Πέραν όμως από αυτές τις δύο περιπτώσεις, όπου ο θυρεοειδής δεν λειτουργεί καλά, έχουμε και τις περιπτώσεις όπου μπορεί να εμφανιστεί κάποιος όγκος στην περιοχή αυτή. Αυτό αναφέρεται ως καρκίνος του θυρεοειδούς αδένα. Οι περισσότερες περιπτώσεις καρκίνου του θυρεοειδούς αδένα είναι ιάσιμες. Υπάρχουν όμως και περιπτώσεις όπου μπορεί να εξελιχτούν, γι' αυτό η έγκαιρη διάγνωση έχει μεγάλη σημασία. Ο καρκίνος του θυρεοειδούς μπορεί να εμφανιστεί σε οποιαδήποτε ηλικία και έχει προκύψει από έρευνες ότι οι γυναίκες έχουν δυο με τρείς φορές περισσότερες πιθανότητες να προσβληθούν σε σχέση με τους άντρες.

Είναι ενθαρρυντικό το γεγονός ότι οι περισσότερες μάζες που εμφανίζονται στην περιοχή αυτή, σε ποσοστό 90% είναι καλοήθεις. Επιπλέον οι μάζες που αποδεικνύονται ότι είναι καρκινικοί όγκοι έχουν πολύ καλή πρόγνωση. Κάθε μάζα όμως που θα εμφανιστεί στην περιοχή, θα πρέπει να ελεγχθεί πλήρως καθώς η μη

αφαίρεση ενός όγκου μπορεί να προκαλέσει προβλήματα στην κατάποσή και την αναπνοή και ακόμα και βραχνάδα.

Υπάρχουν τέσσερα είδη καρκίνου του θυρεοειδούς αδένα. Ο πρώτος τύπος είναι ο Θηλώδης τύπος, ο οποίος αποτελεί το 80% των περιπτώσεων, εμφανίζεται στα κύτταρα όπου παράγουν τη θυροξίνη και συχνά υποτροπιάζει. Επόμενος είναι ο Θυλακιώδης τύπος, ο οποίος εμφανίζεται στα ίδια κύτταρα και έχει ποσοστό ίασης της τάξης του 75% σε διάστημα πέντε χρόνων από την διάγνωση. Ο συγκεκριμένος τύπος μπορεί να διηθήσει στους λεμφαδένες που βρίσκονται κοντά στον θυρεοειδή και να κάνει μεταστάσεις. Σειρά έχει ο Μυελοειδής τύπος, ο οποίος αποτελεί το 5-10% των περιπτώσεων, μπορεί να κάνει και αυτός μεταστάσεις και εκκρίνει μια ορμόνη, την καλσιτονίνη. Αυτός ο τύπος μπορεί να συνυπάρχει και με άλλους καρκίνους του ενδοκρινολογικού συστήματος και το ποσοστό επιβίωσης του, στα πέντε χρόνια από τη διάγνωση, ανέρχεται στο 70%. Τέταρτος και τελευταίος τύπος είναι ο Αναπλαστικός τύπος, ο οποίος είναι η σπανιότερη μορφή και η πλέον επικίνδυνη καθώς εξελίσσεται πολύ γρήγορα και είναι ιδιαίτερα επιθετικός. Μπορεί να διηθήσει γρήγορα στους τοπικούς ιστούς, στις δομές του λαιμού και να προκαλέσει αναπνευστικά προβλήματα.

Η πρώτη θεραπευτική ενέργεια που έγινε για τον καρκίνο του θυρεοειδούς, ήταν η χειρουργική αφαίρεση του όγκου. Στη συνέχεια, οι περισσότεροι ασθενείς, υποβάλλονταν σε θεραπεία με ραδιενεργό ιώδιο, το οποίο απορροφάται από τον θυρεοειδή και καταστρέφει τα υπολειπόμενα καρκινικά κύτταρα. Άλλοι θεραπευτικοί τρόποι που μπορούν να χρησιμοποιηθούν για τον καρκίνο του θυρεοειδούς, είναι η ακτινοθεραπεία και η ορμονοθεραπεία.



Εικόνα 1 Θυρεοειδής αδένας

1.1.2 Ιστορικά στοιχεία

Μετά τον 2ο παγκόσμιο πόλεμο, άρθρα από την ιατρική βιβλιογραφία έδειξαν ότι η εμφάνιση καρκίνου σε κονδύλους του θυρεοειδούς, που μπορούν να αφαιρεθούν χειρουργικά ανέρχονταν σε ποσοστό 20 με 30%. Ορισμένοι πίστευαν ότι οι περιπτώσεις καρκίνου του θυρεοειδούς που οδήγησαν σε θάνατο [1] ήταν ελάχιστες. Έτσι δημιουργήθηκε μια διαμάχη σχετικά με τον αληθινό κίνδυνο των κονδυλωμάτων του θυρεοειδούς. Μέχρι την δεκαετία του ’50 ήταν κοινώς αποδεκτό ότι η νοσηρότητα και η θνησιμότητα του καρκίνου του θυρεοειδούς δεν δικαιολογούσε την χειρουργική αφαίρεση όλων των κονδυλωμάτων [2]. Σήμερα η διαμάχη δεν επικεντρώνεται πια στο αν θα πρέπει να αφαιρεθούν κάποια κονδυλώματα ή όχι αλλά, στα μέσα με τα οποία θα επιλεχθούν τα ασθενή κονδυλώματα για χειρουργική βιοψία.

Κατά τη δεκαετία του ’60 και ΄70 έγινε εμφανές στους παθολόγους στις Ηνωμένες Πολιτείες ότι εικόνες ραδιονουκλεϊδίου και υπέρηχοι του θυρεοειδούς θα ήταν επιτυχής μέθοδοι για την διαπίστωση της αναγκαιότητας της αφαίρεσης όχι περισσότερων από 10 με 20% των κονδυλωμάτων. Και γύρω στο 1950 στις σκανδιναβικές χώρες η προσοχή των ερευνητών εστιάστηκε στην χρήση της λεπτής βελόνας για την λήψη κυτταρολογικού δείγματος από τα κονδυλώματα του θυρεοειδούς [3]με σκοπό τον καθορισμό της πιθανής διάγνωσης.

Αναφορές 25 χρόνων για την συγκεκριμένη μέθοδο προκάλεσαν ελάχιστο ενδιαφέρον στους επιστήμονες άλλων χωρών. Οι λόγοι ήταν θεωρητικοί και περιελάμβαναν την δυσαρέσκεια για την ασταθή αναφερόμενη ευαισθησία των ευρωπαϊκών μελετών, την αποτυχία των συντακτών να δώσουν κατευθύνσεις για την διαδικασία της βιοψίας προς αποφυγή της χειρουργικής επέμβασης και τέλος η γενική γνώση ότι στην κυτταρολογία δεν μπορεί να στηριχθεί διάγνωση με ένα μόνο τμήμα ιστού αλλά χρειάζεται πολλά τμήματα. Μέχρι το 1940 οι καναδοί είχαν αναφέρει εμπειρίες με βιοψία λεπτής βελόνας[3] και ομάδες επιστημόνων είχαν αξιολογήσει την large-needle biopsy.[4]

1.1.3 Σημερινοί προβληματισμοί

Σήμερα θεωρείται βάσει της νοσηρότητας και της θνησιμότητας του, ότι ο καρκίνος του θυρεοειδούς δεν κατατάσσεται ως σημαντικό δημόσιο πρόβλημα υγείας. Λόγω της πληθώρας διαγνωστικών δοκιμών και χειρουργικών λοβοτομών, που έγιναν για να επιβεβαιώσουν ή να αποκλείσουν την παρουσία του, ο καρκίνος του θυρεοειδούς θεωρείται κυρίως μια ασθένεια οικονομικής σπουδαιότητας. Ζώντας σε κοινωνίες που ενδιαφέρονται για τη συγκράτηση των ιατρικών δαπανών, εμφανίστηκε η ανάγκη να επιλεχθούν με προσοχή πιο οικονομικά αποδοτικές διαγνωστικές μέθοδοι.

Από εμπειρίες ερευνητών διαπιστώνεται ότι η βιοψία με χρήση βελόνων είναι πολύ ακριβής για την επιλογή των ασθενών με κονδυλώματα, για τη διαγνωστική

λοβοτομή και είναι φτηνότερη από οποιοδήποτε συνδυασμό άλλων διαγνωστικών μέσων. Η χρήση της έχει μειώσει κατά 50% τις διαδικασίες που ορίζονται και έχει διπλασιάσει τον αριθμό καρκινωμάτων που προσδιορίστηκαν σε 100 χειρουργικές επεμβάσεις.[5] Η μείωση στο μισό των χειρουργικών λογαριασμών αλλά και των λογαριασμών των νοσοκομείων για τη διαχείριση των κονδυλωμάτων είναι ένα σημαντικό επίτευγμα. Επίσης τώρα πια μπορεί να προσδιοριστεί εάν τα κονδυλώματα που εντοπίζονται μπορούν να χαρακτηριστούν ως καλοήθη ή ακόμα να κάνουμε διάγνωση του καρκίνου σε πιο αρχικό στάδιο. Βέβαια για να καθοριστεί εάν αυτό είναι επίσης συμφέρον, και αν μπορεί να επηρεάσει θετικά τη νοσηρότητα και τη θνησιμότητα του καρκίνου του θυρεοειδούς απαιτούνται πολλά έτη μελέτης ακόμα.

Εν κατακλείδι, τώρα πια οι περισσότεροι παθολόγοι συμφωνούν ότι ούτε η αφαίρεση όλων των κονδυλωμάτων του θυρεοειδούς, αλλά ούτε και η αφαίρεση κανενός κονδυλώματος είναι μια λογική προσέγγιση. Επομένως, υιοθετούν κάποια διαδικασία επιλογής για τον ορισμό της χειρουργικής λοβοτομής. Η πιο οικονομικά αποδοτική μέθοδος επιλογής είναι βιοψία βελόνων.

1.1.4 Βιοψία λεπτής βελόνας (Fine Needle Aspiration -FNA)

Η βιοψία λεπτής βελόνας είναι μια μέθοδος κατά την οποία μια λεπτή βελόνα χρησιμοποιείται για να αφαιρέσει ένα δείγμα κυττάρων από την ύποπτη περιοχή για διαγνωστικούς σκοπούς. Η δειγματοληψία με αυτή τη μέθοδο φαίνεται στην Εικόνα 2. Το υλικό που λαμβάνεται μετατρέπεται σε κυτταρολογικό δείγμα κατάλληλο για επεξεργασία με μικροσκόπιο. Σαν μέθοδος θεωρείται ελάχιστα επεμβατική και οικονομικά αποδοτική με διαγνωστική ακρίβεια της τάξης το 90-99%. Παρά την επιτυχία της, η μέθοδος αυτή δεν ήταν διαδεδομένη μέχρι τη δεκαετία του 1980. Οι λόγοι ήταν οι ακόλουθοι: δεν ήταν σίγουροι οι επιστήμονες για την αποδοτικότητα και αποτελεσματικότητα της διαδικασίας, επίσης υπήρχε ο φόβος εμφύτευσης του όγκου στην διαδρομή που δημιουργείται από την βελόνα και τέλος οι χειρούργοι δεν ήταν πρόθυμοι να σταματήσουν τη χρήση των τυπικών ιστολογικών τεχνικών [6]. Σήμερα ο σκεπτικισμός ορισμένων ιστοπαθολόγων για την τεχνική αυτή έχει μειωθεί και μαζί του και ο φόβος ότι μπορεί η μέθοδος αυτή να αντικαταστήσει την διάγνωση ιστού.

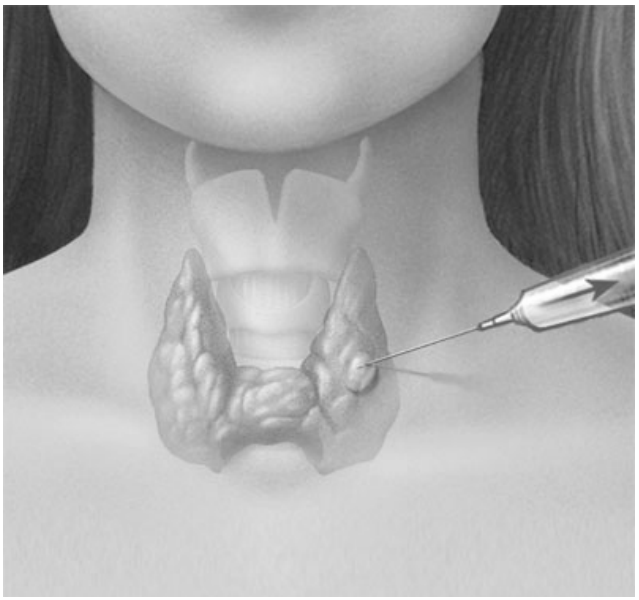
Όπως βέβαια όλες οι μέθοδοι και τεχνικές έχουν πλεονεκτήματα και μειονεκτήματα, έτσι και η βιοψία λεπτής βελόνας έχει τα δικά της. Όσον αφορά τα πλεονεκτήματα της, η βιοψία λεπτής βελόνας είναι ασφαλής μέθοδος, δίνει μια σύντομη αναφορά του προβλήματος, είναι ευαίσθητη και ακριβής για τη διάγνωση κακοήθειας των όγκων, επίσης απαιτεί περιορισμένο εξοπλισμό και προκαλεί ελάχιστη δυσαρέσκεια στον ασθενή. Είναι μια διαδικασία που μπορεί να πραγματοποιηθεί και σε εξωτερικούς ασθενείς νοσοκομείων, που μειώνει την διεξαγωγή των διερευνητικών διαδικασιών, και επιτρέπει την οριστική διάγνωση σε μη επιδεχόμενους χειρουργική επέμβαση ασθενείς. Τέλος η μέθοδος αυτή δεν

χρειάζεται ίαση των πληγών, είναι εύκολα επαναλαμβανόμενη και οικονομικά αποδεκτή [6].

Στα μειονεκτήματα της μεθόδου έχουμε να αναφέρουμε ότι η βιοψία λεπτής βελόνας σαν διαδικασία απαιτεί εμπειρία και ικανότητα. Ένα συγκεκριμένο ποσοστό ασθενών έχει δηλώσει δυσαρέσκεια για τον τρόπο πραγματοποίησης της διαδικασίας. Εμπειρία απαιτείται επίσης και για την ερμηνεία των αποτελεσμάτων [6]. Επιπλέον, η διαγνωστική ακρίβεια της μεθόδου, παρά το ότι είναι ιδιαίτερα υψηλή, υπολείπεται κατά της ιστολογικής διάγνωσης, καθώς ένα μικρό αλλά όχι αμελητέο ποσοστό των περιπτώσεων δεν μπορεί να χαρακτηριστεί με βεβαιότητα βάσει των αποτελεσμάτων της μεθόδου και κατά συνέπεια χαρακτηρίζεται ως «ύποπτο - suspicious» και αντιμετωπίζεται κλινικά σαν κακοήθες καρκίνωμα.

Η διαγνωστική ακρίβεια της βιοψίας λεπτής βελόνας εξαρτάται από πολλούς παράγοντες. Ορισμένοι από αυτούς είναι ο τύπος και η περιοχή από όπου θα ληφθεί το δείγμα, η εμπειρία του ιατρού που πραγματοποιεί τη διαδικασία λήψης, η ποιότητα της προετοιμασίας του δείγματος και οι διαγνωστικές δεξιότητες των ειδικών παθολόγων που θα κάνουν την διάγνωση [7]. Μάλιστα μελέτες έχουν δείξει ότι η απόδοση είναι μεγαλύτερη όταν το ίδιο άτομο που πραγματοποιεί τη δειγματοληψία, κάνει την προετοιμασία και ερμηνεύει τα αποτελέσματα.

Στην περίπτωση ύποπτων συμπτωμάτων στον θυρεοειδή αδένα, η βιοψία λεπτής βελόνας είναι η πλέον αποδοτική και οικονομική μέθοδος για την αξιολόγηση του, είτε πραγματοποιηθεί σε εξωτερικούς ασθενείς, είτε σε ασθενείς που νοσηλεύονται.



Εικόνα 2 Δειγματοληψία με χρήση λεπτής βελόνας

1.1.5 Σκοπός της παρούσας πτυχιακής

Όπως αναφέρθηκε και στα μειονεκτήματα της μεθόδου FNA ένα ποσοστό των περιπτώσεων δεν είναι δυνατό να χαρακτηριστεί με βεβαιότητα και κατά συνέπεια θεωρείται ως «ύποπτο». Όταν εφαρμόστηκε η βιοψία λεπτής βελόνας στα δεδομένα που αναλύουμε στις επόμενες παραγράφους, κατηγοριοποίησε τα δεδομένα σε τρείς κλάσεις. Χαρακτήρισε τα δείγματα σε καλοήθη με ποσοστό 69%, σε κακοήθη, με ποσοστό 4%, αλλά και σε ύποπτα (suspicious), με ποσοστό 10%. Στους ασθενείς των οποίων τα δείγματα χαρακτηρίστηκαν ύποπτα, πραγματοποιήθηκε επέμβαση και ιστολογική εξέταση - βιοψία και διαπιστώθηκε ότι μόνο το 25% των δειγμάτων αυτών είναι πραγματικά κακοήθη. Η κλινική αυτή πρακτική υπαγορεύεται από την προφανή ανάγκη να ελαχιστοποιούνται οι ψευδώς αρνητικές διαγνώσεις, εις βάρος των πιθανών ψευδώς θετικών.

Σκοπός της παρούσας πτυχιακής εργασίας είναι να διερευνηθεί η δυνατότητα εφαρμογής ενός αλγόριθμου κατηγοριοποίησης, ο οποίος θα λαμβάνει ως είσοδο μία σειρά χαρακτηριστικών που προκύπτουν από την βιοψία λεπτής βελόνας και θα είναι σε θέση να μειώσει τα ύποπτα (suspicious) δείγματα. Έτσι θα μπορούν τα αποτελέσματα της εξέτασης FNA να είναι περισσότερο αξιόπιστα και να μην αναγκάζουν τον ασθενή να υποβληθεί σε περαιτέρω εξετάσεις, οι οποίες είναι σωματικά και ψυχικά επώδυνες.

Στο πλαίσιο της εργασίας θα εφαρμοστούν κάποιοι γνωστοί αλγόριθμοι κατηγοριοποίησης στα διαθέσιμα δεδομένα βιοψίας θυρεοειδούς, καθώς και ένας σχετικά νέος αλγόριθμος κατηγοριοποίησης, ο οποίος βασίζεται σε ένα υπολογιστικό ανάλογο του ανθρώπινου ανοσοποιητικού συστήματος. Ο αλγόριθμος αυτός είναι σχετικά νέος, κατά συνέπεια θα μελετηθεί διεξοδικά και θα εξεταστεί η συμπεριφορά του σε σχέση με τις παραμέτρους του.

1.2 Κατηγοριοποίηση

Μέχρι τώρα είδαμε τον τρόπο με τον οποίο γίνεται η λήψη του δείγματος από το ανθρώπινο σώμα, για να γίνει έπειτα επεξεργασία και διάγνωση. Ο αλγόριθμος που μελετάται στην πτυχιακή αυτή και με την χρήση του επεξεργαζόμαστε τα δεδομένα από βιοψίες λεπτής βελόνας, βασίζεται στην έννοια της κατηγοριοποίησης.

Κατηγοριοποίηση ή ταξινόμηση είναι ο προσδιορισμός της κατηγορίας στην οποία ανήκει ένα αντικείμενο ή φαινόμενο ή πρότυπο ή μέτρηση κλπ. Ο όρος ταξινόμηση είναι πιο γενικός από τον όρο κατηγοριοποίηση, περιλαμβάνοντας πολλές φορές και την έννοια της διάταξης των κατηγοριών.

Τα προβλήματα κατηγοριοποίησης έχουν ως είσοδο ένα σύνολο από αντικείμενα, κάθε ένα από τα οποία περιγράφεται από ένα διάνυσμα χαρακτηριστικών (feature vector) και παράγουν ως έξοδο την κατηγορία ή κλάση

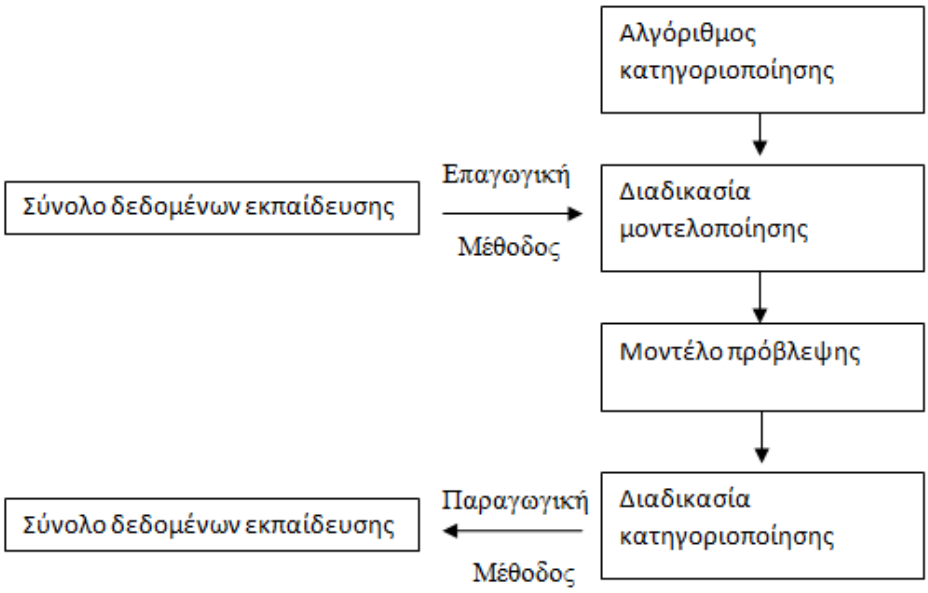
στην οποία ανήκουν τα αντικείμενα. Η επιλογή της κατηγορίας στην οποία κατατάσσεται ένα αντικείμενο γίνεται από ένα προκαθορισμένο σύνολο κατηγοριών.

Ανάλογα με το πρόβλημα και την περίπτωση που έχουμε να μελετήσουμε, οι σχέσεις της κατηγοριοποίησης προκύπτουν με δυο τρόπους. Είτε από μαθηματικές σχέσεις και αλγορίθμους που γενικεύουν τα παρατηρούμενα δεδομένα εκπαίδευσης δημιουργώντας μοντέλα, όπως συμβαίνει στην περίπτωση της μηχανικής μάθησης, είτε βάσει ευριστικών κανόνων οι οποίοι συνήθως προκύπτουν από την εμπειρική γνώση ενός ανθρώπου-ειδικού. Στην πτυχιακή αυτή θα χρησιμοποιήσουμε τον πρώτο τρόπο κατηγοριοποίησης, με την χρήση εποπτευόμενων αλγορίθμων [8].

Πιο αναλυτικά η διαδικασία της εποπτευόμενης κατηγοριοποίησης πραγματοποιείται ως εξής. Λαμβάνεται ως είσοδος ένα σύνολο δεδομένων εκ των προτέρων κατηγοριοποιημένο, το οποίο θα ονομάζουμε υποσύνολο εκπαίδευσης. Στη συνέχεια εκτελείται ο αλγόριθμος κατηγοριοποίησης ο οποίος εκπαιδεύεται στην κατασκευή του μοντέλου κατηγοριοποίησης. Πιο συγκεκριμένα, ο αλγόριθμος ρυθμίζει τις παραμέτρους του έτσι ώστε να μεγιστοποιεί την ακρίβεια κατηγοριοποίησης του υποσυνόλου εκπαίδευσης.

Για να μπορέσει ο αλγόριθμος να πετύχει τον στόχο του, να ανακαλύψει δηλαδή τις συσχετίσεις μεταξύ των τιμών των γνωρισμάτων πρόβλεψης και των κλάσεων, αναλύει το σύνολο εκπαίδευσης και χρησιμοποιώντας κατάλληλες τεχνικές, δημιουργεί ένα σύνολο κανόνων πρόβλεψης. Οι κανόνες αποτελούνται από δυο μέρη: (α) από τις συνθήκες πρόβλεψης και (β) από την κλάση πρόβλεψης. Το παραγόμενο σύνολο κανόνων απαρτίζει το μοντέλο πρόβλεψης, το οποίο στη συνέχεια ενσωματώνεται σε ένα σύστημα κατηγοριοποίησης. Η τυπική δομή ενός συστήματος κατηγοριοποίησης φαίνεται στην Εικόνα 3.

Το σύνολο των κανόνων πρόβλεψης πρέπει να προσαρμόζεται στα δεδομένα του συνόλου εκπαίδευσης αλλά και να είναι ικανό να προβλέπει δεδομένα που δεν έχει συναντήσει στο παρελθόν. Συμπεραίνουμε λοιπόν ότι μια βασική ιδιότητα ενός αποτελεσματικού αλγορίθμου κατηγοριοποίησης είναι η δημιουργία κανόνων πρόβλεψης με μεγάλη ικανότητα γενίκευσης. [9]



Εικόνα 3 Τυπική δομή ενός συστήματος κατηγοριοποίησης

1.3 Εξελικτικοί και Μιμητικοί αλγόριθμοι

Ένας εξελικτικός αλγόριθμος, ορίζεται σαν μια αλγοριθμική διαδικασία που διατηρεί έναν πληθυσμό ατόμων, τον οποίο εξελίσσει σύμφωνα με κάποιους κανόνες επιλογής και κάποιους τελεστές, όπως ο ανασυνδυασμός και η μετάλλαξη. Συνήθως κάθε άτομο του πληθυσμού κωδικοποιεί μία λύση του προβλήματος που αντιμετωπίζει ο αλγόριθμος. Αν το πρόβλημα είναι τύπου βελτιστοποίησης, τότε η λύση αποτελεί ένα σημείο του χώρου στον οποίο ορίζεται η συνάρτηση κόστους που βελτιστοποιείται. Αν το πρόβλημα είναι τύπου κατηγοριοποίησης, τότε η λύση μπορεί να αποτελείται από ένα σύνολο κανόνων βάσει των οποίων προκύπτει η κλάση ενός αντικειμένου. Σε κάθε εκτέλεση λοιπόν, έχουμε ένα σύνολο λύσεων σε αντίθεση με άλλους ευριστικούς αλγορίθμους που συνήθως σε κάθε εκτέλεση εξελίσσουν μια ενιαία λύση ή ένα μέρος της τελικής λύσης που επιδιώκεται. Οι εξελικτικοί και μιμητικοί αλγόριθμοι είναι ευριστικοί αλγόριθμοι εμπνευσμένοι από τη βιολογία και τους φυσικούς ή κοινωνικούς μηχανισμούς εξέλιξης. Ο κυριότερος εκπρόσωπος των εξελικτικών αλγορίθμων είναι οι Γενετικοί Αλγόριθμοι [39].

Πολλές μελέτες έχουν γίνει μέχρι σήμερα για την βελτίωση της ποιότητας των αποτελεσμάτων που προκύπτουν από τους εξελικτικούς αλγορίθμους (ΕΑ). Μια τέτοια προσέγγιση συνιστά και η δημιουργία ενός μοντέλου στο οποίο διάφοροι αλγόριθμοι, καλούνται να συλλειτουργήσουν προκειμένου να υπάρχει όφελος κατά τη διερεύνηση του χώρου λύσεων από τα καλύτερα χαρακτηριστικά από τις μεθόδους αυτές [10]. Το μοντέλο αυτό καλείται υβριδικό. Το πλέον διαδεδομένο υβριδικό μοντέλο είναι ο συνδυασμός ενός ΕΑ με ένα αλγόριθμο τοπικής βελτιστοποίησης.

2 Υλικά και μέθοδοι

2.1 Δεδομένα που χρησιμοποιήθηκαν

2.1.1 Συνθετικά δεδομένα

Για τον έλεγχο της μεθόδου που προτείνουμε στην πτυχιακή αυτή χρησιμοποιήσαμε δεδομένα που είτε τα δημιουργήσαμε, είτε τα πήραμε έτοιμα από βάσεις δεδομένων.

Τα δεδομένα που δημιουργήσαμε εμείς τα ονομάσαμε «συνθετικά δεδομένα» και αποτελούνται από δυο κλάσεις. Επιλέξαμε να χρησιμοποιήσουμε δεδομένα με δύο (2) χαρακτηριστικά, έτσι ώστε να είναι εύκολη η δημιουργία αυθαίρετα πολύπλοκων δεδομένων, καθώς και η οπτικοποίηση των αποτελεσμάτων της κατηγοριοποίησης. Πιο συγκεκριμένα, απεικονίζουμε τα συνθετικά δεδομένα στο καρτεσιανό επίπεδο χρησιμοποιώντας ως συντεταγμένες τα δύο χαρακτηριστικά, ενώ η κλάση τους απεικονίζεται με χρήση κατάλληλου χρώματος. Η κατασκευή των συνθετικών δεδομένων έγινε με κατάλληλο λογισμικό επεξεργασίας εικόνας. Πριν την εκτέλεση της μεθόδου χωρίζουμε τα δεδομένα μας σε δεδομένα εκπαίδευσης και δεδομένα ελέγχου, δηλαδή ένα τμήμα των δεδομένων χρησιμοποιείται για την εκπαίδευση του συστήματος και τα υπόλοιπα για την επαλήθευση της εγκυρότητας. Στόχος μας είναι να μπορέσει ο αλγόριθμος να ταξινομήσει σωστά τα δεδομένα.

Με την τεχνική που περιγράφτηκε προηγουμένως είναι δυνατό να δημιουργηθούν με γραφικό ή αλγοριθμικό τρόπο συνθετικά δεδομένα. Δημιουργήσαμε διαφορετικές εικόνες με διαφορετικό διαχωρισμό των δυο κλάσεων ώστε να μπορούμε να δούμε τα διαφορετικά αποτελέσματα που θα έχει ο αλγόριθμος. Ο διαχωρισμός για τα συνθετικά δεδομένα (α) αποτελούνται από το υποσύνολο της κλάσης A και B, D_A, D_B αντίστοιχα, τα οποία παράγονται τυχαία ως εξής:

$$D_A = \{(x_i, y_i)\}, i = 1,..., N_A, x_i = \Re([1,64]), y_i = \Re([1,128])$$
$$D_B = \{(x_i, y_i)\}, i = 1,..., N_A, x_i = \Re((64,128]), y_i = R([1,128])$$

όπου $\Re([a,b])$ τυχαίος πραγματικός αριθμός που ακολουθεί επίπεδη κατανομή στο διάστημα $[a,b]$.

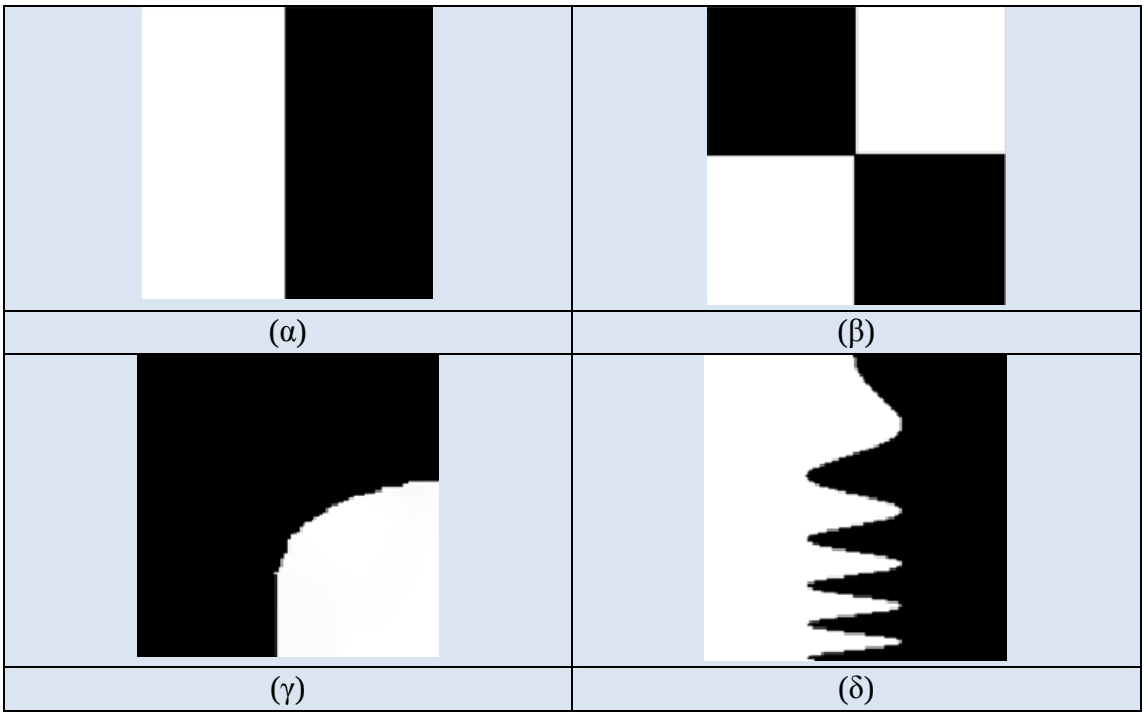
Ο διαχωρισμός για τα συνθετικά δεδομένα (β) αποτελούνται από το υποσύνολο της κλάσης A και B, D_A, D_B αντίστοιχα, τα οποία παράγονται τυχαία ως εξής:

$$D_A = \{(x_i, y_i)\}, i = 1,..., N_A, (x_i, y_i) \in [1,64] \times [1,64] \cup (64,128] \times (64,128]$$
$$D_B = \{(x_i, y_i)\}, i = 1,..., N_A, (x_i, y_i) \in [1,64] \times (64,128] \cup (64,128] \times [1,64]$$

Για τα συνθετικά δεδομένα (γ), η καμπύλη διαχωρισμού έγινε με τυχαίο τρόπο από τον χρήστη. Ενώ για τα συνθετικά δεδομένα (δ) χρησιμοποιήθηκε η παρακάτω εξίσωση καμπύλης διαχωρισμού

$$y = 64 + 32 * \sin\left(2\pi\left(\frac{x}{32}\right)\frac{x}{128}\right)$$

Η δυσκολία των συνθετικών δεδομένων (δ) οφείλεται σε δυο προάγοντες. Ο πρώτος είναι ότι η συνάρτηση είναι ημιτονοειδής και έτσι η μια κλάση καλύπτεται από την άλλη σε ορισμένα σημεία και ο δεύτερος είναι ότι η καμπύλη αυτή έχει μονοτομικά αυξανόμενη συχνότητα, όπως παρατηρούμε και στην παρακάτω εικόνα.

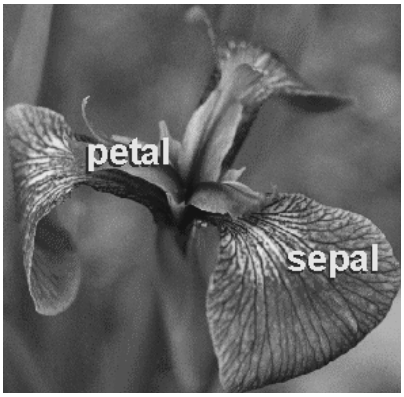


Εικόνα 4 Συνθετικά δεδομένα (α), έχουμε δύο κλάσεις οι οποίες χωρίζονται γραμμικά, συνθετικά δεδομένα (β), δύο κλάσεις και πάλι γραμμικά διαχωρίσιμες, με μεγαλύτερο βαθμό δυσκολίας στην ταξινόμηση, συνθετικά δεδομένα (γ), δυο κλάσεις για τις οποίες ο διαχωρισμός έγινε με τυχαία καμπύλη από τον χρήστη και στα συνθετικά δεδομένα (δ), έχουμε δυο κλάσεις διαχωρισμένες με ημιτονοειδή καμπύλη μη σταθερής συχνότητας.

2.1.2 Iris Δεδομένα

Επόμενα δεδομένα που χρησιμοποιήσαμε, είναι τα δεδομένα Iris τα οποία τα είναι διαθέσιμα για ανάπτυξη και έλεγχο αλγορίθμων κατηγοριοποίησης και επιλογής χαρακτηριστικών [11][12]. Τα δεδομένα αυτά αποτελούνται από πενήντα δείγματα από διαφορετικά είδη της οικογένειας του φυτού iris. Κάθε είδος του iris φυτού μπορεί να ταξινομηθεί ανάλογα με το μήκος και πλάτος των σέπαλων του και των πετάλων του, βλέπε Εικόνα 5. Έχουμε λοιπόν δυο κλάσεις κάθε μία εκ των οποίων περιέχει πενήντα δείγματα και τέσσερα χαρακτηριστικά.

Οι μετρήσεις αυτές των δεδομένων, μας δίνουν ένα μέσο όρο και μία τυπική απόκλιση γύρω από την οποία κινείται κάθε είδος της οικογένειας. Ο στόχος μας είναι και σε αυτήν την περίπτωση να μπορέσουμε με την μέθοδο που προτείνουμε να ταξινομήσουμε σωστά τα χαρακτηριστικά της κάθε οικογένειας. Για την εκπαίδευση όπως και παραπάνω θα χρησιμοποιηθεί ένα τμήμα μόνο των δεδομένων και τα υπόλοιπα θα χρησιμοποιηθούν για να εξεταστεί η εγκυρότητα της μεθόδου [13].



Εικόνα 5: Το φυτό Ίρις, από τα πέταλα και τα σέπαλα του οποίου προκύπτουν τα δεδομένα που αναλύουμε στην παράγραφο αυτή. Τα χαρακτηριστικά των κλάσεων που δημιουργούνται αναφέρονται στο μήκος και το πλάτος των πετάλων και των σέπαλων.

2.1.3 Πραγματικά δεδομένα

Τελευταία και σημαντικότερα δεδομένα είναι τα αποτελέσματα από πραγματικές βιοψίες κυττάρων του θυρεοειδούς για διάγνωση κακοήθειας. Τα δεδομένα προέρχονται από την εξέταση 9162 ασθενών, του Α’ τμήματος παθολογίας στην ιατρική σχολή Αθηνών, κατά το χρονικό διάστημα 2000-2004 [14]. Τα δείγματα αυτά εξετάστηκαν με την μέθοδο Βιοψίας Λεπτής Βελόνας (FNA) και από αυτόν τον συνολικό αριθμό δειγμάτων τα 723 θεωρήθηκαν ανεπαρκή για διάγνωση. Τα υπόλοιπα δείγματα από τη βιοψία FNA χαρακτηρίστηκαν ως κακοήθη, καλοήθη και

ύποπτα. Τελικά από τα 2016 δείγματα που χρησιμοποιήθηκαν στην μελέτη τα 1866 χαρακτηρίστηκαν ως καλοήθη. Από αυτά τα 140 προέκυψαν βάσει των ιστολογικών δεδομένων ή συμπληρωματικών κλινικών και εργαστηριακών αποτελεσμάτων και αξιολογήσεων απεικονίσεων και τα 1726 προέκυψαν από περιπτώσεις όπου η ιστολογική εξέταση δεν ήταν εφικτή. Στην παρούσα εργασία, τα δείγματα που χαρακτηρίστηκαν ως καλοήθη αποτελούν την πρώτη κλάση και η δεύτερη κλάση αποτελείται από 150 δείγματα που χαρακτηρίστηκαν ως κακοήθη και προέρχονται μόνο από ιστολογικά δείγματα. Σύμφωνα με την εξέταση Βιοψίας Λεπτής Βελόνας, τα 1848 δείγματα της πρώτης κλάσης χαρακτηρίστηκαν ως καλοήθη, τα 2 ως κακοήθη και τα 16 ως ύποπτα. Αντίστοιχα από τη δεύτερη κλάση τα 97 χαρακτηρίστηκαν κακοήθη, τα 16 καλοήθη και τα 37 ως ύποπτα. Τα παραπάνω συνοψίζονται στον Πίνακα 1.

Πλήθος δεδομένων	**Αποτελέσματα ιστολογικής εξέτασης**	**Αποτελέσματα κατηγοριοποίησης FNA**
2016	1866 Καλοήθη	1848 Καλοήθη
		2 Κακοήθη
		16 Ύποπτα
	150 Κακοήθη	97 Κακοήθη
		16 Καλοήθη
		37 Ύποπτα

Πίνακας 1 Δεδομένα από την βιοψία λεπτής βελόνας και η ταξινόμησή τους. Αποτελέσματα από μελέτη που πραγματοποιήθηκε σε εξέταση ασθενών, του Α' τμήματος παθολογίας στην ιατρική σχολή Αθηνών, κατά το χρονικό διάστημα 2000-2004.

Κάθε ένα από τα δείγματα που χρησιμοποιήθηκαν έχει 67 κυτταρολογικά χαρακτηριστικά γνωρίσματα, τα οποία έχουν δύο δυνατές τιμές, το 0 και το 1 που υποδεικνύουν την απουσία ή την ύπαρξη στο δείγμα αντίστοιχα. Τα χαρακτηριστικά αυτά αναφέρονται στον πίνακα 2.

Όπως αναφέραμε και παραπάνω, η βιοψία λεπτής βελόνας χώρισε τα δεδομένα σε 2 ομάδες, τα ύποπτα (suspicious) και τα μη ύποπτα (non-suspicious). Οι δοκιμές του αλγορίθμου ταξινόμησης έγιναν αρχικά με τα μη-ύποπτα δεδομένα, καθώς ήταν πιο εύκολο να ταξινομηθούν και έπειτα με τα ύποπτα.

1	ΚΟΛΛΟΕΙΔΕΣ ΠΟΛΥ
2	ΚΟΛΟΕΙΔΕΣ ΛΙΓΟ
3	ΥΠΑΡΞΗ ΚΟΛΟΕΙΔΟΥΣ
4	ΑΙΜΑ
5	ΑΜΥΛΟΕΙΔΕΣ
6	ΜΙΚΡΗ ΚΥΤΤΑΡΟΒΡΙΘΕΙΑ
7	ΜΕΓΑΛΗ ΚΥΤΤΑΡΟΒΡΙΘΕΙΑ
8	ΥΠΑΡΞΗ ΚΥΤΤΑΡΟΒΡΙΘΕΙΑΣ

9	ΛΕΜΦΟΚΥΤΤΑΡΑ ΠΟΛΛΑ
10	ΛΕΜΦΟΚΥΤΤΑΡΑ ΛΙΓΑ
11	ΥΠΑΡΞΗ ΛΕΜΦΟΚΥΤΤΑΡΩΝ
12	ΝΕΚΡΩΤΙΚΟ
13	ΑΣΒΕΣΤΙΟ
14	ΛΙΠΟΕΙΔΗ
15	ΦΑΓΟΚΥΤΤΑΡΑ ΠΟΛΛΑ
16	ΦΑΓΟΚΥΤΤΑΡΑ ΛΙΓΑ
17	ΥΠΑΡΞΗ ΦΑΓΟΚΥΤΤΑΡΩΝ
18	ΠΟΛΥΜΟΡΦΟΠΥΡΗΝΑ
19	ΓΙΓΑΝΤΟΚΥΤΤΑΡΑ
20	ΆΛΛΟ
21	DETRITUS
22	ΣΥΝΔΕΤΙΚΟ ΥΠΟΣΤΡΩΜΑ
23	ΨΑΜΜΩΔΗ
24	ΙΝΟΒΛΑΣΤΕΣ
25	ΙΣΤΙΟΚΥΤΤΑΡΑ
26	ΚΑΛΗ ΣΥΝΟΧΗ
27	ΧΑΛΑΡΗ ΣΥΝΟΧΗ
28	ΜΟΝΟΣΤΟΙΒΑΔΩΣΗ
29	ΜΙΚΡΟΘΥΛΑΚΙΑ
30	ΘΥΛΑΚΙΩΔΕΙΣ ΣΧΗΜΑΤΙΣΜΟΙ ΠΟΛΛΟΙ
31	ΘΥΛΑΚΙΩΔΕΙΣ ΣΧΗΜΑΤΙΣΜΟΙ ΛΙΓΟΙ
32	ΥΠΑΡΞΗ ΘΥΛΑΚΙΩΔΕΙΣ ΣΧΗΜΑΤΙΣΜΟΙ
33	ΣΥΜΠΑΓΕΙΣ ΣΧΗΜΑΤΙΣΜΟΙ
34	ΘΗΛΩΔΕΙΣ ΣΧΗΜΑΤΙΣΜΟΙ ΠΟΛΛΟΙ
35	ΘΗΛΩΔΕΙΣ ΣΧΗΜΑΤΙΣΜΟΙ ΛΙΓΟΙ
36	ΥΠΑΡΞΗ ΘΗΛΩΔΕΙΣ ΣΧΗΜΑΤΙΣΜΟΙ
37	ΕΛΕΥΘΕΡΑ
38	ΜΟΝΟΠΥΡΗΝΑ ΚΥΤΤΑΡΑ
39	ΔΙΠΥΡΗΝΑ ΚΥΤΤΑΡΑ
40	ΠΟΛΥΠΥΡΗΝΑ ΚΥΤΤΑΡΑ
41	ΘΥΛΑΚΙΚΑ ΚΥΤΤΑΡΑ
42	ΑΤΥΠΑ ΚΥΤΤΑΡΑ ΠΟΛΛΑ
43	ΑΤΥΠΑ ΚΥΤΤΑΡΑ ΛΙΓΑ
44	ΥΠΑΡΞΗ ΑΤΥΠΑ ΚΥΤΤΑΡΑ
45	ΕΠΙΔΕΡΜΟΕΙΔΗ ΚΥΤΤΑΡΑ
46	ΟΞΥΦΙΛΑ
47	ΑΝΙΣΟΚΥΤΤΑΡΩΣΗ
48	ΟΜΑΛΗ
49	ΠΟΛΥΜΟΡΦΙΑ ΚΥΤΤΑΡΩΝ
50	ΣΑΦΗ ΟΡΙΑ
51	ΜΗ ΣΑΦΗ ΟΡΙΑ
52	ΑΤΡΑΚΤΟΕΙΔΗ
53	ΚΛΑΣΣΙΚΟ
54	ΠΛΑΣΜΟΚΥΤΟΕΙΔΕΣ
55	ΔΙΑΥΓΕΣ

56	ΠΟΛΥΓΩΝΙΚΑ
57	ΚΟΚΚΙΩΣΗ
58	ΚΑΙΝΟΤΟΠΙΩΣΗ
59	ΕΚΚΕΝΤΡΟΣ
60	ΚΕΝΤΡΙΚΟΣ
61	ΑΝΙΣΟΚΑΡΥΩΣΗ
62	ΕΓΚΛΕΙΣΤΑ
63	ΠΥΚΝΩΤΙΚΟΣ ΠΥΡΗΝΑΣ
64	ΙΣΟΠΥΡΗΝΩΣΗ
65	ΠΥΡΗΝΙΚΟΣ ΣΥΝΩΣΤΙΣΜΟΣ
66	ΠΥΡΗΝΙΑ
67	ΠΟΛΥΜΟΡΦΙΑ ΠΥΡΗΝΑ

Πίνακας 2 Ιστολογικά Χαρακτηριστικά των Δεδομένων που έχουν ληφθεί με βιοψία λεπτής βελόνας. Στα δεδομένα δίνουν την τιμή 0 ή 1 ανάλογα με το αν υπάρχουν ή όχι.

2.2 *Διαχωρισμός των δεδομένων σε υποσύνολο εκπαίδευσης και ελέγχου*

Κάθε φορά που θέλαμε να εφαρμόσουμε μια μέθοδο κατηγοριοποίησης, διαχωρίζαμε τα δεδομένα σε δύο ξένα υποσύνολα: ελέγχου D_{test} και εκπαίδευσης D_{train} του αλγορίθμου. Ο διαχωρισμός αυτός σε κάθε περίπτωση, γίνεται με τον παρακάτω τρόπο:

Για κάθε ένα στοιχείο από το σύνολο των N διαθέσιμων δεδομένων, ορίζουμε μια τυχαία μεταβλητή r η οποία υπολογίζεται με τυχαίο τρόπο στο διάστημα [0,1] ακολουθώντας επίπεδη κατανομή. Εάν η μεταβλητή αυτή είναι μικρότερη από το ποσοστό διαχωρισμού που θέλουμε να ορίσουμε το στοιχείο κατατάσσεται στο σύνολο ελέγχου, αλλιώς κατατάσσεται στο σύνολο εκπαίδευσης. Ακολουθεί ο ψευδοκώδικας της συνάρτησης διαχωρισμού των δεδομένων.

$D_{train} = \varnothing$
$D_{test} = \varnothing$
ΓΙΑ ΚΑΘΕ στοιχείο $d_i, i = 1,...,N$
$r = \Re([0,1])$
IF r<0.6
$D_{train} = D_{train} \cup d_i$
Else
$D_{test} = D_{train} \cup d_i$

Στα συνθετικά δεδομένα και στα δεδομένα Iris οι δυο κλάσεις έχουν ίδιο αριθμό στοιχείων και έτσι μπορεί να θεωρηθεί ότι τα σύνολα ελέγχου και εκπαίδευσης, μπορούν να συμπληρωθούν με την ίδια πιθανότητα και από τις δυο κλάσεις. Στα πραγματικά δεδομένα όμως όπου οι δυο κλάσεις, καλοήθη και κακοήθη δεδομένα, διαφέρουν σημαντικά ως προς το πλήθος τους, θεωρήσαμε πως θα ήταν καλύτερο το ποσοστό διαχωρισμού να μην είναι της τάξης 90-10 για test και train αλλά της τάξης 60-40. Με τον τρόπο αυτό αυξάνουμε το επιτρεπόμενο πλήθος στα δεδομένα εκπαίδευσης αυξάνοντας την πιθανότητα να ληφθούν και δεδομένα από την λιγότερο πολυπληθή κλάση.

Την παραπάνω διαδικασία την εφαρμόσαμε στα non-suspicious δεδομένα, ενώ για τα suspicious δεδομένα εργαστήκαμε ως εξής. Κάναμε αρχικά διαχωρισμό σε δεδομένα ελέγχου και δεδομένα εκπαίδευσης στα non-suspicious με ποσοστό 60-40, το οποίο μας δίνει τα καλύτερα αποτελέσματα όπως φαίνεται σε παρακάτω παραγράφους. Στη συνέχεια ορίσαμε ως δεδομένα ελέγχου τα suspicious για να ελέγξουμε τα αποτελέσματα τους.

2.3 Συγκριτικές μέθοδοι κατηγοριοποίησης

Πέραν της μεθόδου κατηγοριοποίησης δεδομένων με χρήση τεχνητού ανοσοποιητικού συστήματος που μελετάμε στη συγκεκριμένη εργασία χρησιμοποιήσαμε και κάποιες άλλες μεθόδους κατηγοριοποίησης για να μπορέσουμε να συγκρίνουμε τα αποτελέσματα και να δούμε την απόδοση της μεθόδου. Οι μέθοδοι που επιλέξαμε ήταν τα τεχνητά νευρωνικά δίκτυα(ΤΝΔ) και ο αλγόριθμος k κοντινότερων γειτόνων (kNN).

2.3.1 Κατηγοριοποιητής τεχνητών νευρωνικών δικτύων (ΤΝΔ)

Η πρώτη μέθοδος κατηγοριοποίησης που θα περιγράψουμε είναι τα νευρωνικά δίκτυα. Τα νευρωνικά δίκτυα είναι μια ιδιαίτερη προσέγγιση στη δημιουργία συστημάτων με νοημοσύνη καθώς αποφεύγουν να αναπαραστήσουν ρητά τη γνώση και να υιοθετήσουν ειδικά σχεδιασμένους αλγόριθμους αναζήτησης. Αντίθετα βασίζονται σε βιολογικά πρότυπα καθώς χρησιμοποιούν δομές και διαδικασίες που μιμούνται τις αντίστοιχες του ανθρώπινου εγκεφάλου.

Τα τεχνητά νευρωνικά δίκτυα είναι συστήματα επεξεργασίας δεδομένων που αποτελούνται από πλήθος τεχνητών νευρώνων οργανωμένων σε δομές παρόμοιες με αυτές του ανθρωπίνου εγκεφάλου. Συνήθως οι τεχνητοί νευρώνες είναι οργανωμένοι σε μία σειρά από στρώματα ή επίπεδα (layers). Το πρώτο από αυτά ονομάζεται επίπεδο εισόδου (input layer) και χρησιμοποιείται για την εισαγωγή δεδομένων. Στη συνέχεια μπορεί να ακολουθούν προαιρετικά, ένα ή περισσότερα ενδιάμεσα ή κρυφά

επίπεδα (hidden layers), ενώ στο τέλος υπάρχει το επίπεδο εξόδου (output layer). Μια αναπαράσταση ενός τέτοιου δικτύου φαίνεται στο σχήμα 2.

Όταν δεν υπάρχουν συνδέσεις μεταξύ νευρώνων ενός επιπέδου και νευρώνων προηγούμενου επιπέδου τα τεχνητά νευρωνικά δίκτυα χαρακτηρίζονται ως δίκτυα με πρόσθια τροφοδότηση (feedforward). Στην αντίθετη περίπτωση, καθώς και στην περίπτωση σύνδεσης μεταξύ νευρώνων ίδιου επιπέδου, τα τεχνητά νευρωνικά δίκτυα χαρακτηρίζονται ως δίκτυα με ανατροφοδότηση (feedback ή recurrent) [15]. Τα πρόσθιας τροφοδότησης ΤΝΔ είναι τα πρώτα και τα απλούστερα νευρωνικά δίκτυα που επινοήθηκαν. Σε αυτά τα δίκτυα οι πληροφορίες κινούνται μόνο σε μία κατεύθυνση από τους κόμβους εισαγωγής, μέσω κρυμμένων κόμβων, στους κόμβους εξόδου, χωρίς να υπάρχει κύκλος ή βρόχος στο δίκτυο

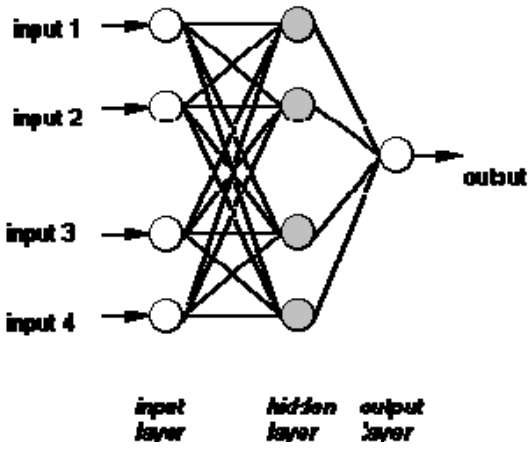
Το θεμελιώδες στοιχείο σε ένα ΤΝΔ είναι ο τεχνητός νευρώνας. Ο κάθε νευρώνας του ΤΝΔ είναι μια σύναψη των εισόδων του. Η λειτουργία του περιλαμβάνει τον αθροιστή, ο οποίος προσθέτει τα επηρεασμένα από τα βάρη σήματα εισόδου και τη συνάρτηση ενεργοποίησης (activation function), ένα είδος φίλτρου το οποίο διαμορφώνει την τελική τιμή του σήματος εξόδου y, σε συνάρτηση με την ποσότητα S και την τιμή κατωφλίου της συνάρτησης ενεργοποίησης. Οι είσοδοι μπορεί να είναι, είτε κάποια εξωτερικά σήματα, ή οι έξοδοι από άλλους νευρώνες. Υπάρχει επίσης και ένα διαφορετικό στοιχείο εισόδου, που είναι γνωστό ως bias (βάρος) και χρησιμοποιείται προκειμένου να ληφθούν υπόψη επιδράσεις παραγόντων ανεξάρτητων από τις εισόδους. Έτσι για τον νευρώνα j ο οποίος συνδέεται με N νευρώνες του προηγούμενου επιπέδου, με τιμές βαρύτητας w_{ji} και με bias b_j, η έξοδος θα είναι της μορφής:

$$y_j = f\left(\sum_{i=1}^{N-1} w_{ji}x_i + b_j\right)$$

όπου f η συνάρτηση ενεργοποίησης η οποία συνήθως είναι της μορφής:

$$f(x) = \frac{1}{1+e^{-x}}$$

Όταν η τιμή της συνάρτησης ενεργοποίησης ξεπεράσει ένα κατώφλι, τότε ο νευρώνας παράγει ένα σήμα εξόδου το οποίο θα διαδοθεί στους νευρώνες του επόμενου στρώματος του νευρωνικού δικτύου.



Εικόνα 6 Αναπαράσταση ενός δικτύου πρόσθιας τροφοδότησης

Ο κατηγοριοποιητής νευρωνικού δικτύου ανήκει στην κατηγορία των κατηγοριοποιητών με επίβλεψη (supervised classifiers). Αυτό σημαίνει ότι υπάρχει μια φάση εκπαίδευσης του δικτύου, κατά τη διάρκεια της οποίας τα βάρη του μεταβάλλονται, με βάση κάποιο κανόνα εκμάθησης, μέχρι η παραγόμενη έξοδος του ΤΝΔ να συγκλίνει με την επιθυμητή. Όταν το ΤΝΔ χρησιμοποιείται σαν κατηγοριοποιητής, τότε το επίπεδο εξόδου έχει τόσους νευρώνες όσες και οι κλάσεις των δεδομένων. Η κατηγοριοποίηση γίνεται βάσει του νευρώνα του επιπέδου εξόδου που ενεργοποιείται. Τα δεδομένα του υποσυνόλου εκμάθησης παρουσιάζονται στο ΤΝΔ και η διαφορά της εξόδου του από την επιθυμητή έξοδο χρησιμοποιείται για την μεταβολή των βαρών του. Τα δεδομένα του συνόλου εκμάθησης παρουσιάζονται συνεχώς στην είσοδο του ΤΝΔ μέχρι να ελαχιστοποιηθεί το μέσο τετραγωνικό σφάλμα μεταξύ στην επιθυμητή και την πραγματική απόκριση του δικτύου. Η ελαχιστοποίηση του σφάλματος επιτυγχάνεται με την εφαρμογή κάποιας μεθόδου βελτιστοποίησης. Μια από τις πιο γνωστές μεθόδους βελτιστοποίησης που εφαρμόζεται για την εκπαίδευση ενός νευρωνικού δικτύου είναι η μέθοδος της όπισθεν διάδοσης (backpropagation). Σύμφωνα με τη μέθοδο αυτή, η διαφορά της εξόδου από την επιθυμητή έξοδο διαδίδεται προς τα πίσω και χρησιμοποιείται για την ανανέωση των βαρών κάθε νευρώνα. Αυτή η διαδικασία επαναλαμβάνεται για κάθε διάνυσμα του συνόλου εκπαίδευσης. Όταν παρουσιαστούν όλα τα διανύσματα τότε έχει συμπληρωθεί μια εποχή (epoch). Μετά το τέλος κάθε επιτυχημένης εποχής το συνολικό σφάλμα, για όλα τα διανύσματα του συνόλου εκπαίδευσης, έχει ελαττωθεί. Η διαδικασία συνεχίζεται μέχρι τα βάρη κάθε νευρώνα να συγκλίνουν, το οποίο σημαίνει ότι η επιπλέον εκπαίδευση του δικτύου δεν αλλάζει σημαντικά τις τιμές τους. Αφού ολοκληρωθεί η εκπαίδευση του δικτύου μπορεί να χρησιμοποιηθεί για την κατηγοριοποίηση "άγνωστων" χαρακτηριστικών διανυσμάτων σε μια από τις προκαθορισμένες κατηγορίες.

Στην εργασία αυτή χρησιμοποιούμε μια εφαρμογή των τεχνητών νευρωνικών δικτύων, η οποία παρέχεται από το περιβάλλον ανάπτυξης Matlab[16]. Το πρόγραμμα αυτό δέχεται τα δεδομένα και περνώντας τα από ένα νευρωνικό δίκτυο τα

κατηγοριοποιεί και βγάζει το ποσοστό αυτών που τοποθετούνται στην σωστή κλάση. Τα βήματα της εφαρμογής του κατηγοριοποιητή νευρωνικών δικτύων είναι τα εξής:

Περιγραφή

Είσοδοι:

- Πίνακας δεδομένων

Έξοδοι:

- Η ακρίβεια κατηγοριοποίησης για κάθε κλάση

Βήματα της μεθόδου:

1. Διαχωρισμός δεδομένων σε 2 σύνολα: εκπαίδευσης και ελέγχου
2. Δημιουργία του νευρωνικού δικτύου
   i. Εισαγωγή των επιπέδων του νευρωνικού δικτύου
   ii. Εισαγωγή των συναρτήσεων μεταφοράς από το ένα επίπεδο του δικτύου στο άλλο
3. Εκπαίδευση του νευρωνικού δικτύου
4. Εισαγωγή των δεδομένων ελέγχου, εκπαίδευση και εξαγωγή της ακρίβεια για κάθε κλάση

Σύνταξη

[N1,N2,accurA,accurB]=NN

2.3.2 Κατηγοριοποιητής k- πλησιέστερων γειτόνων(kNN)

Ο αλγόριθμος *k* πλησιέστερων γειτόνων είναι ένας σειριακός αλγόριθμος, όπου τα στοιχεία συγχωνεύονται επαναληπτικά στις πλησιέστερες μεταξύ των κλάσεων που υπάρχουν, σε κάθε επανάληψη. Παίρνει τα δεδομένα σε δυο ομάδες σε αυτή του συνόλου εκπαίδευσης και αυτή του συνόλου ελέγχου. Ο k-NN εκπαιδεύεται, αποθηκεύοντας τα διανύσματα που αντιστοιχούν στα παραδείγματα εκπαίδευσης, και μαζί και τις εξόδους των στοιχείων αυτών. Αποθηκεύει δηλαδή τα σημεία ενός πολυδιάστατου χώρου, στον οποίο μπορούν να αναπαρασταθούν τα στοιχεία αυτά.

Κατά την φάση ταξινόμησης, δηλαδή κατά τη χρήση του εκπαιδευμένου k-NN, το σύστημα λαμβάνει τις νέες εισόδους, για τις οποίες δεν γνωρίζει την έξοδο και υπολογίζει για κάθε μία τη διανυσματική της αναπαράσταση, δηλαδή το αντίστοιχο σημείο στον πολυδιάστατο χώρο. Έπειτα, υπολογίζεται η απόσταση του σημείου του στοιχείου εισόδου από κάθε σημείο που αντιστοιχεί σε ένα αποθηκευμένο παράδειγμα εκπαίδευσης. Η απόσταση ορίζεται χρησιμοποιώντας Ευκλείδεια μετρική, ενώ στην περίπτωση μη ετερόκλητων χαρακτηριστικών, απαιτείται κανονικοποίηση των τιμών τους σε κοινή ακτίνα τιμών. Αφού

υπολογιστούν οι αποστάσεις αυτές, είναι εύκολο να βρεθούν τα k στοιχεία εκπαίδευσης με τη μικρότερη απόσταση από το σημείο της εισόδου και έτσι η είσοδος κατατάσσεται στην κατηγορία που είναι πιο συχνή μεταξύ των k κοντινότερων παραδειγμάτων εκπαίδευσης. Το k είναι ένας φυσικός αριθμός ο οποίος αποτελεί παράμετρο του αλγορίθμου. Συνήθως είναι ένας περιττός φυσικός αριθμός για να μπορεί να υπάρξει πλειοψηφία.

Ο αλγόριθμος απαιτεί περισσότερους υπολογισμούς κατά την κατάταξη νέων στοιχείων, όσο αυξάνει το πλήθος των παραδειγμάτων εκπαίδευσης, αφού υπολογίζεται κάθε φορά η απόσταση του από όλα τα παραδείγματα εκπαίδευσης. Έχει επίσης, μεγάλες απαιτήσεις μνήμης, αφού πρέπει να αποθηκεύονται όλα τα παραδείγματα εκπαίδευσης. Από την άλλη πλευρά, όμως, ο αλγόριθμος είναι εξαιρετικά απλός, είναι ταχύτατος κατά την εκπαίδευση, αφού απλά απομνημονεύει τα παραδείγματα εκπαίδευσης και μπορεί να μάθει υπερ-επιφάνειες διαχωρισμού οποιουδήποτε είδους, σε αντίθεση με γραμμικούς διαχωριστές [17][18].

Αφού γίνει η κατάταξη ο αλγόριθμος ελέγχει την απόδοση του με τρείς τρόπους. Αρχικά με τα δεδομένα ελέγχου, που λαμβάνει σαν παράμετρο μαζί με τα δεδομένα εκπαίδευσης. Έπειτα ο αλγόριθμός μας δίνει σαν έξοδο την μήτρα αληθείας (confusion matrix), η οποία μας δίνει το πλήθος των αληθών θετικών και αρνητικών σε σχέση με το σύνολο των δεδομένων, όπως και το σύνολο των ψευδών θετικών και αρνητικών. Και τέλος μπορούμε να ελέγξουμε την απόδοση του αλγορίθμου μέσω της καμπύλης OC (operating characteristic)[19].

Στη εργασία αυτή χρησιμοποιώντας τις εντολές της Matlab φτιάξαμε ένα πρόγραμμα το οποίο να δέχεται τα δεδομένα μας, και αφού υπολογίσει την απόσταση τους από k πλησιέστερους γείτονες, να κάνει την ταξινόμηση και να μας δώσει το ποσοστό των σωστά τοποθετημένων στοιχείων. Τα βήματα της εφαρμογής του αλγορίθμου αυτού είναι τα εξής:

Περιγραφή
Είσοδοι:

- Πίνακας δεδομένων

Έξοδοι:

- Μήτρα αληθείας [N1,N2,N3,N4]

Βήματα της μεθόδου:

1. Χωρισμός δεδομένων σε σύνολα εκπαίδευσης Train και ελέγχου Test, ξένα μεταξύ τους.
2. Για κάθε ένα από τα δεδομένα του συνόλου ελέγχου
   i. Εύρεση των κοντινότερων γειτόνων του από τα μέλη του συνόλου εκπαίδευσης.
   ii. Υπολογισμός της κλάσης της πλειοψηφίας των γειτόνων
3. Κατασκευή του πίνακα συνάφειας

Ο υπολογισμός της απόστασης και η κατηγοριοποίηση γίνεται με μια επιμέρους εφαρμογή, η οποία περιγράφεται ως εξής:

Είσοδοι:

- Δεδομένα ελέγχου, δεδομένα εκπαίδευσης, πλήθος γειτόνων και πλήθος κλάσεων

Έξοδοι:

- Μήτρα αληθείας [N1,N2,N3,N4] και οι κλάσεις των δεδομένων ελέγχου.

Σύνταξη

[class_out,Confusion]=newknn2(CTEST,CTRAIN,K,nclass)

3 Προτεινόμενη Μέθοδος

Στο κεφάλαιο αυτό αρχικά παρουσιάζεται ο τρόπος λειτουργίας του Βιολογικού Ανοσοποιητικού Συστήματος και στη συνέχεια γίνεται μια εισαγωγή στην έννοια του Τεχνητού Ανοσοποιητικού Συστήματος. Τέλος περιγράφεται ο τρόπος λειτουργίας ενός αλγορίθμου βασισμένου σε ένα τέτοιο σύστημα. Στόχος είναι να κατανοήσουμε την σύνδεση μεταξύ των δυο για να μπορέσουμε να περιγράψουμε τον αλγόριθμο που χρησιμοποιήσαμε.

3.1 Βιολογικό Ανοσοποιητικό Σύστημα

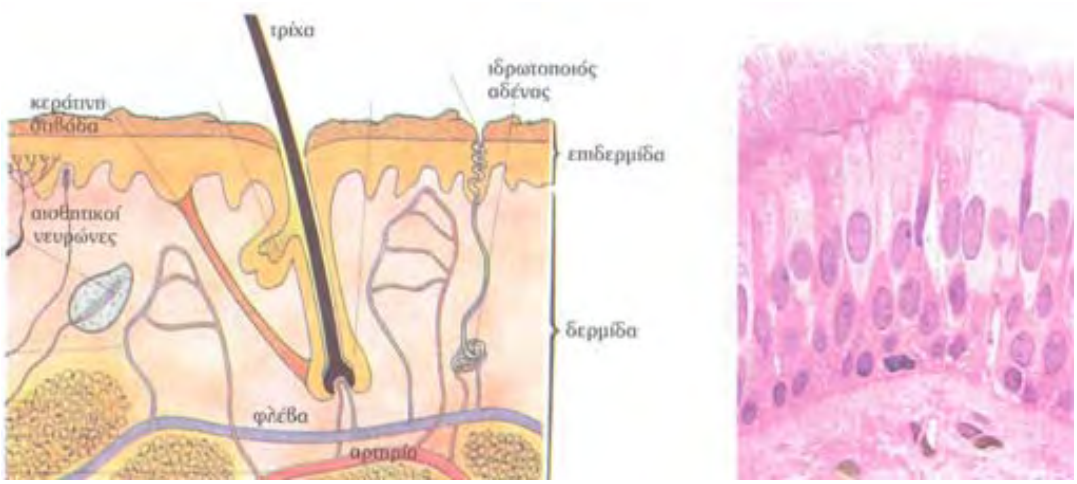
Το ανοσοποιητικό σύστημα είναι ένα σύστημα οργάνων υπεύθυνο για την άμυνα του οργανισμού. Αποτελείται από πολλά διαφορετικά όργανα και ιστούς. Τα σημαντικότερα από αυτά είναι ο μυελός των οστών και ο θύμος αδένας. Σε αυτά δημιουργούνται και αναπτύσσονται τα ειδικά κύτταρα του ανοσοποιητικού συστήματος. Δευτερεύοντα όργανα του ανοσοποιητικού συστήματος είναι οι αμυγδαλές, ο σπλήνας, τα λεμφογάγγλια και οι πλάκες Peyer.

Η άμυνα του οργανισμού εναντίον εξωτερικών παραγόντων επιτυγχάνεται με ένα σύνολο από μηχανισμούς, οι οποίοι μπορούν να διαχωριστούν τόσο με βάση τη θέση τους στο ανθρώπινο σώμα, όσο και με βάση την ιδιότητα τους να έχουν γενικευμένη δράση(μη ειδικοί αμυντικοί μηχανισμοί) ή εξειδικευμένη δράση(ειδικοί αμυντικοί μηχανισμοί).

Βασικό χαρακτηριστικό της μη ειδικής άμυνας είναι η δυνατότητα αντιμετώπισης οποιουδήποτε παθογόνου μικροοργανισμού και περιλαμβάνει μηχανισμούς που παρεμποδίζουν την είσοδο των μικροοργανισμών στον οργανισμό μας, αλλά και μηχανισμούς που αντιμετωπίζουν τους μικροοργανισμούς όταν έχουν πια εισέλθει στον οργανισμό μας.

Οι μηχανισμοί που εμποδίζουν τους παθογόνους οργανισμούς να εισέλθουν στο σώμα μας είναι το δέρμα και οι βλεννογόνοι των διάφορων οργάνων. Το δέρμα τους εμποδίζει λόγω της δομής του αλλά και λόγω των ουσιών που παράγονται από τους σμηγματογόνους και τους ιδρωτοποιούς αδένες. Οι βλεννογόνοι από την πλευρά τους, καλύπτουν κοιλότητες του οργανισμού και με τη βλέννα που εκκρίνουν παγιδεύουν τους μικροοργανισμούς και δεν τους επιτρέπουν να περάσουν στον οργανισμό, βλέπε Εικόνα 7.

Αν παρά τους παραπάνω φραγμούς, ένα μικρόβιο καταφέρει να περάσει στον οργανισμό, θα έρθει αντιμέτωπο με μια δεύτερη γραμμή αμυντικών μηχανισμών, στους οποίους ανήκει η φαγοκυττάρωση, η φλεγμονώδης αντίδραση, ο πυρετός και η δράση ορισμένων αντιμικροβιακών ουσιών.



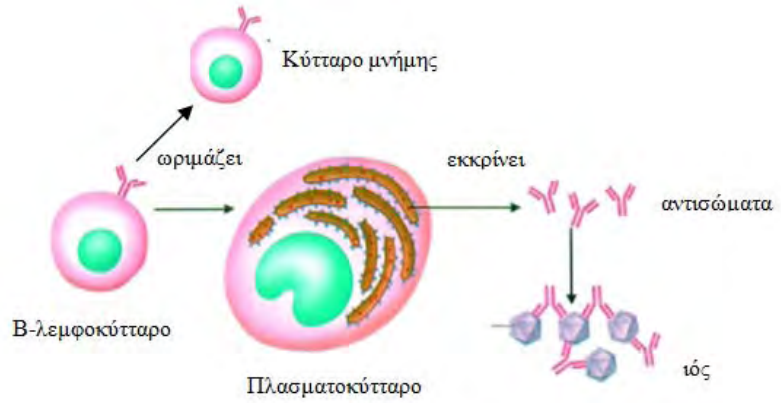
Εικόνα 7 Δέρμα και βλεννογόνοι αδένες στην άμυνα του οργανισμού[20]

Η προσοχή μας όμως εστιάζεται στην ειδική άμυνα του οργανισμού μας, την ανοσία, όπως ονομάζεται. Ο ανθρώπινος οργανισμός έχει την ικανότητα να αναγνωρίζει οποιαδήποτε ξένη προς αυτόν ουσία και να αντιδρά παράγοντας εξειδικευμένα κύτταρα και κυτταρικά προϊόντα, ώστε να την εξουδετερώσει. Η ικανότητα αυτή ονομάζεται ανοσία και αυτήν μιμείται και ο αλγόριθμος που εξετάζουμε, γι' αυτό έχει μεγάλη σημασία να κατανοήσουμε την διαδικασία. Μια ακόμα σημαντική έννοια, είναι αυτή της ξένης ουσίας που προκαλεί την ανοσοβιολογική απόκριση και ονομάζεται αντιγόνο. Ως αντιγόνο μπορεί να δράσει είτε ένας ολόκληρος μικροοργανισμός, όπως για παράδειγμα ένας ιός ή ένα βακτήριο, είτε ένα τμήμα αυτού, αλλά και οι τοξίνες που παράγονται από αυτόν.

Οι μηχανισμοί ειδικής άμυνας διαθέτουν δυο χαρακτηριστικά που τους κάνουν να ξεχωρίζουν από τους μη ειδικούς μηχανισμούς. Πρώτο χαρακτηριστικό είναι η εξειδίκευση, που σημαίνει ότι τα προϊόντα της ανοσοβιολογικής απόκρισης θα δράσουν μόνο εναντίον της ουσίας που προκάλεσε την παραγωγή τους και δεύτερο χαρακτηριστικό είναι η μνήμη, που είναι η ικανότητα του οργανισμού να «θυμάται» τα αντιγόνα με τα οποία έχει έρθει σε επαφή, έτσι ώστε μετά από μια πιθανή δεύτερη έκθεση του σε αυτά να αντιδρά πιο γρήγορα.

Τα κύτταρα που απαρτίζουν το ανοσοβιολογικό σύστημα είναι κυρίως τα λεμφοκύτταρα, τα οποία ανήκουν στα λευκά αιμοσφαίρια. Τα λεμφοκύτταρα είναι μικρά σφαιρικά κύτταρα και διακρίνονται σε δυο κατηγορίες, τα Τ-λεμφοκύτταρα και τα Β-λεμφοκύτταρα. Τα μεν Τ-λεμφοκύυταρα διαφοροποιούνται και ωριμάζουν στον θύμο αδένα και είναι απαραίτητα για την ολοκλήρωση της ανοσοβιολογικής απόκρισης και τα δε Β-λεμφοκύτταρα διαφοροποιούνται και ωριμάζουν στο μυελό των οστών. Τα δεύτερα συνθέτουν και παρουσιάζουν στην επιφάνεια τους ειδικές πρωτεΐνες που ονομάζονται ανοσοσφαιρίνες ή αντισώματα. Κάθε Β-λεμφοκύτταρο διαθέτει υποδοχείς – αντισώματα που αναγνωρίζουν ένα συγκεκριμένο αντιγόνο και

συνδέονται μαζί του. Εξαιτίας της σύνδεσης που πραγματοποιείται, το Β-λεμφοκύτταρο υφίσταται διαδοχικές διαιρέσεις, από τις οποίες παράγονται τα πλασματοκύτταρα και τα Β-λεμφοκύτταρα μνήμης, βλέπε Εικόνα 8. Τα πλασματοκύτταρα παράγουν και εκκρίνουν μεγάλες ποσότητες αντισωμάτων και τα Β-λεμφοκύτταρα μνήμης ενεργοποιούνται αμέσως μετά από επόμενη έκθεση του οργανισμού στο ίδιο αντιγόνο.



Εικόνα 8 Ωρίμανση Β-λεμφοκυττάρων

Η αντίδραση του οργανισμού μας στους παθογόνους οργανισμούς συνιστά την ανοσοβιολογική απόκριση και περιγράφεται ως εξής:

Όταν ένας παθογόνος μικροοργανισμός εμφανίζεται, ενεργοποιούνται τα μακροφάγα, τα οποία πέρα από τη δυνατότητα που έχουν να καταστρέφουν το μικρόβιο, μπορούν και να εκθέτουν στην επιφάνεια τους τμήματα του μικροβίου που έχουν καταστρέψει. Το τμήμα του μικροβίου που εκτίθεται συνδέεται με μια πρωτεΐνη της επιφάνειας των μακροφάγων. Η πρωτεΐνη αυτή είναι χαρακτηριστική για κάθε άτομο και ονομάζεται αντιγόνο ιστοσυμβατότητας.

Τα πρώτα κύτταρα που ενεργοποιούνται με την παρουσίαση του αντιγόνου είναι τα βοηθητικά Τα-λεμφοκύτταρα, τα οποία εκκρίνουν ουσίες για να ενεργοποιηθούν τα Β-λεμφοκύτταρα, προκειμένου αυτά να πολλαπλασιαστούν και τελικά να διαφοροποιηθούν σε πλασματοκύτταρα και Β-κύτταρα μνήμης, όπως αναφέραμε πιο πάνω. Τα πλασματοκύτταρα, εκκρίνουν και αυτά ποσότητες αντισωμάτων ειδικών για το συγκεκριμένο αντιγόνο, ενώ τα Β-λεμφοκύτταρα μνήμης θα ενεργοποιηθούν στην περίπτωση που ο οργανισμός εκτεθεί ξανά στο ίδιο αντιγόνο. [20]

3.2 Τεχνητό ανοσοποιητικό Σύστημα

Τα τελευταία χρόνια, το ενδιαφέρον για την διερεύνηση και την εκμετάλλευση των δυνατοτήτων ενός εποπτευόμενου υπολογιστικού συστήματος μάθησης, βασισμένο στο ανοσοποιητικό σύστημα των θηλαστικών γίνεται όλο και μεγαλύτερο, αποτελώντας έμπνευση για επιστήμονες και μηχανικούς, καθώς τα βιολογικά συστήματα όπως ο ανθρώπινος οργανισμός επιτελούν περίπλοκη επεξεργασία διαφόρων πληροφοριών, αποτελούμενα από απλούστερες επιμέρους μονάδες. Τέτοια συστήματα επεξεργασίας, εμπνευσμένα από βιολογικούς οργανισμούς μπορούν να διαχωριστούν σε τρεις κατηγορίες: στα νευρωνικά δίκτυα(*βλέπε2.3.1*), τους γενετικούς αλγορίθμους και στο τεχνητό ανοσοποιητικό σύστημα. Απ' αυτές τις κατηγορίες, τα νευρωνικά δίκτυα και οι γενετικοί αλγόριθμοι έχουν εφαρμοστεί ευρέως σε διάφορους τομείς, ενώ το τεχνητό ανοσοποιητικό σύστημα έχει λίγες εφαρμογές.

Το φυσικό ανοσοποιητικό σύστημα όπως περιγράψαμε και παραπάνω, είναι ένα πολυσύνθετο σύστημα με διάφορους μηχανισμούς για την υπεράσπιση ενάντια στους παθογόνους οργανισμούς. Ο κύριος σκοπός του, είναι να αναγνωριστούν όλα τα κύτταρα μέσα στο σώμα και να ταξινομηθούν σε γνωστά ή ξένα για τον οργανισμό. Τα κύτταρα τα οποία είναι ξένα, ταξινομούνται περαιτέρω προκειμένου να προκαλέσουν τον κατάλληλο τύπο αμυντικού μηχανισμού. Το ανοσοποιητικό σύστημα μαθαίνει, μέσω της εξέλιξης, να διακρίνει μεταξύ των επικίνδυνων ξένων αντιγόνων και των κυττάρων του ίδιου του σώματος. Είναι ένα σύστημα, που χρησιμοποιεί την εκμάθηση, τη μνήμη, και τη συνειρμική ανάκτηση για να εκτελέσει τις διαδικασίες αναγνώρισης και ταξινόμησης. Πιο συγκεκριμένα, μαθαίνει να αναγνωρίζει διάφορα μοτίβα, θυμάται τα μοτίβα που έχει συναντήσει στο παρελθόν, και χρησιμοποιεί συνδυασμούς για να κατασκευάσει σωστά και αποτελεσματικά ανιχνευτές μοτίβων. Επίσης, η γενική συμπεριφορά του συστήματος είναι μια διαδικασία που προκύπτει από πολλές τοπικές αλληλεπιδράσεις. Αυτές οι δυνατότητες του ανοσοποιητικού συστήματος παρέχουν διάφορες σημαντικές πτυχές στον τομέα του υπολογισμού. Αυτός ο τομέας είναι που αναφέρεται ως ανοσολογικός υπολογισμός ή τεχνητά ανοσοποιητικά συστήματα.[21][22][23]

3.2.1 Αρχές του ανοσοποιητικού συστήματος που υιοθετήθηκαν για τη δημιουργία του αλγορίθμου

Κατά τη διάρκεια ζωής ενός ατόμου, το σώμα εκτίθεται, όπως και προηγουμένως αναφέραμε, σε πολλούς παθογόνους μικροοργανισμούς. Για να προστατευθεί λοιπόν, το ανοσοποιητικό σύστημα, διαθέτει λεμφοκύτταρα, γνωστά ως B- και T- κύτταρα (B- and T-Cells), καθένα από τα οποία διαθέτει έναν μοναδικό μοριακό δέκτη. Ο δέκτης αυτός έχει ένα συγκεκριμένο σχήμα και συνδέεται με ένα παθογόνο στοιχείο, ή αντιγόνο όπως το ονομάζουμε, το οποίο έχει παρόμοιο σχήμα

με αυτό του δέκτη. Τα αντιγόνα που βρίσκονται στον οργανισμό και επιδιώκουν να συνδεθούν με τον μοριακό δέκτη των κυττάρων είναι πολλά, για το λόγο αυτό χρησιμοποιούμε ένα μέτρο σύγκρισης το οποίο μας δείχνει πόσο ταιριάζουν τα δυο στοιχεία. Το μέτρο αυτό ονομάζεται συγγένεια. Σε άλλες μελέτες που έχουν πραγματοποιηθεί έχει εισαχθεί ο όρος shape-space (σχήμα διαστήματος) [25], ο οποίος περιγράφει την κατανομή των δεδομένων που χρησιμοποιούμε, στο χώρο και μας βοηθάει ώστε να χρησιμοποιήσουμε την ευκλείδεια απόσταση ως μέτρο συγγένειας. Άλλος ένας όρος που χρησιμοποιείται είναι η έννοια της ARB (Artificial Recognition Ball) η οποία περιγράφει την αλληλεπίδραση των κυττάρων με ένα αντιγόνο σε ένα δίκτυο. Πιο απλά μια ARB μπορεί να θεωρηθεί ως ένα σύνολο από όμοια Β-κύτταρα και χρησιμοποιείται για να μειώσει τον πολλαπλασιασμό των κυττάρων και να ενισχύσει την επιβίωση τους στον πληθυσμό.

Απ’ τη στιγμή που θα καθοριστεί η συγγένεια μεταξύ ενός αντιγόνου και ενός Β-κυττάρου, το κύτταρο αυτό διαιρείται σε ένα κύτταρο πλάσματος και έναν κλώνο μνήμης (MC). Κατά τη διαδικασία της επέκταση το κύτταρο κλωνοποιείται με γρήγορους ρυθμούς ανάλογα με το πόσο ταιριάζει με το αντιγόνο. Οι κλώνοι του κυττάρου περνάνε από μια διαδικασία ωρίμανσης όσον αφορά τη συγγένεια και ορισμένα μεταλλάσσονται σωματικά (η μετάλλαξη εδώ αφορά τη συγγένεια του κυττάρου με το αντιγόνο) ώστε να περάσουν από μια τελευταία διαδικασία επιλογής, όπου κάποιο δοθέν κύτταρο θα γίνει κύτταρο μνήμης. Αυτά τα κύτταρα μνήμης, χρησιμεύουν ώστε να έχουμε μια πιο γρήγορη αντίδραση του οργανισμού σε περίπτωση που εμφανιστεί ξανά το ίδιο αντιγόνο στον οργανισμό, ή κάποιο παρόμοιο.

Στα συστήματα αναγνώρισης που βασίζονται στο βιολογικό ανοσοποιητικό σύστημα (AIRS), τα κύτταρα μνήμης που αναφέραμε παραπάνω, χρησιμοποιούνται στην πορεία για ταξινόμηση. Επίσης στα συστήματα αυτά, οι ARBs αφού έρθουν σε επαφή με τα δεδομένα εκπαίδευσης υπόκεινται σε μια μορφή ωρίμανσης και παραγωγής κλώνων, ανάλογη με αυτή των Β-λεμφοκυττάρων, που αναφέραμε στην προηγούμενη παράγραφο. Μετά την δημιουργία τους, τα νέα ARBs γίνονται αντικείμενα σε μια διαδικασία τυχαίας μετάλλαξης με μια συγκεκριμένη πιθανότητα και στη συνέχεια, εάν η συγγένεια τους ικανοποιεί κάποια κριτήρια, ενσωματώνονται στο σύνολο των κυττάρων μνήμης. Μέσα στο σύστημα, το οποίο είναι περιορισμένων πόρων, τα ARBs ανταγωνίζονται και όποιο έχει το μεγαλύτερο επίπεδο υποκίνησης, απαιτεί και τους περισσότερους πόρους. Τα ARBs που δεν κατάφεραν να αναδειχθούν από τον ανταγωνισμό απομακρύνονται από το σύστημα.

Αντιστοιχία	
Βιολογικό Ανοσοποιητικό Σύστημα	***Τεχνητό Ανοσοποιητικό Σύστημα***
Αντίσωμα	Διάνυσμα χαρακτηριστικών γνωρισμάτων
Β-λεμφοκύτταρα	ARB-σύνολο χαρακτηριστικών γνωρισμάτων
Ωρίμανση και Διαίρεση	Αναπαραγωγή των ARBs που ταιριάζουν με τα αντιγόνα
Αντιγόνα	Δεδομένα Εκπαίδευσης και ελέγχου
Ομοιότητα Αντιγόνου-Αντισώματος	Σχέση Συγγένειας
Ωρίμανση Συγγένειας	Τυχαία Μετάλλαξη των ARBs και αφαίρεση των λιγότερο διεγερμένων
Κύτταρα μνήμης	Σύνολο κυττάρων μνήμης του αλγόριθμου

Πίνακας 3 Αντιστοιχία Βιολογικού και Τεχνητού Ανοσοποιητικού Συστήματος

3.2.2 Παράμετροι και δομές δεδομένων του AIRS

Στο σημείο αυτό, θα παρουσιάσουμε τους όρους που χρησιμοποιήθηκαν για τον αλγόριθμο AIRS που θα παρουσιάσουμε παρακάτω.

- ***Συγγένεια (Affinity):*** Είναι ένα μέτρο ομοιότητας μεταξύ δύο αντισωμάτων ή αντιγόνων. Υπολογίζεται ως η ευκλείδεια απόσταση των χαρακτηριστικών διανυσμάτων δύο αντικειμένων. Άρα οι μικρότερες τιμές συγγένειας δείχνουν μεγαλύτερη ομοιότητα.

- ***Κατώφλι συγγένειας (Affinity threshold):*** Η μέση τιμή των τιμών συγγένειας από κάθε δυνατό ζεύγος των αντιγόνων στο σύνολο εκπαίδευσης.

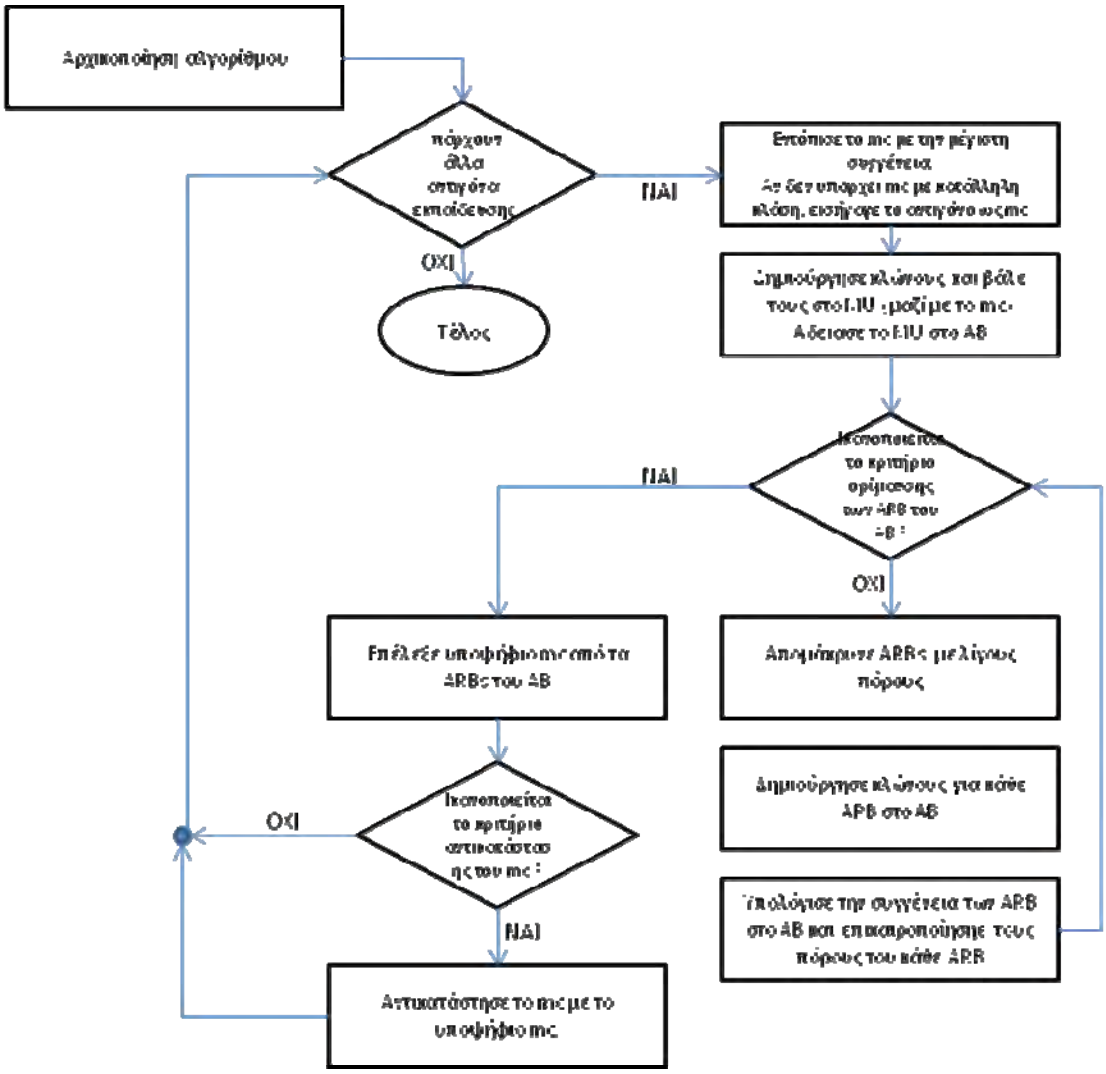
- ***Affinity threshold scalar:*** Μια τιμή στο διάστημα [0,1], η οποία πολλαπλασιασμένη με το κατώφλι συγγένειας, μας δίνει μια τιμή κατωφλίου για την αντικατάσταση των κυττάρων μνήμης στην ρουτίνα εκπαίδευσης του αλγορίθμου.
- ***Αντίσωμα (Antibody):*** Ένα διάνυσμα χαρακτηριστικών γνωρισμάτων συνοδευόμενο με την κλάση στην οποία ανήκει. Ο όρος αντίσωμα υποδεικνύει ότι πρόκειται για κανόνα αναγνώρισης, ο οποίος αποτελεί τμήμα ανός Β-κυττάρου, ή ενός κυττάρου μνήμης.
- ***Αντιγόνο (Antigen):*** Και πάλι συνδυασμός χαρακτηριστικού διανύσματος και της αντίστοιχης κλάσης. Ο όρος αντιγόνο χρησιμοποιείται σε αναλογία με τον βιολογικό όρο για να δηλώσει ότι πρόκειται για δεδομένα εκπαίδευσης ή ελέγχου τα οποία παρουσιάζονται στο σύστημα και προκαλούν ως απάντηση την κατηγοριοποίηση τους.
- ***Artificial Immune Recognition System (AIRS):*** Ο εν λόγω αλγόριθμος ταξινόμησης εμπνευσμένος από το βιολογικό ανοσοποιητικό σύστημα.
- ***Artificial Recognition Ball (ARB):*** Πρόκειται για το υπολογιστικό ανάλογο των Β-λεμφοκυττάρων. Η υπολογιστική υλοποίηση τους αποτελείται από ένα αντίσωμα (διάνυσμα χαρακτηριστικών συνοδευόμενο από την κλάσης του), τον αριθμό των πόρων που διαθέτει το κύτταρο και την τιμή απόκρισης (stimulation value) του κυττάρου στο τρέχον αντιγόνο.
- ***Β-κύτταρο (B-Cell):*** βλ. ARB.
- ***Υποψήφια κύτταρα μνήμης (Candidate memory cells):*** Ένα αντίσωμα μιας ARB, το οποίο ανήκει στην ίδια κλάση με το αντιγόνο και ήταν το πιο διεγερμένο μετά την έκθεση του σε ένα δοθέν αντιγόνο.
- ***Κλάση (class):*** Η ομάδα στην οποία ανήκει ένα δοθέν αντιγόνο.
- ***Ποσοστό κλώνων (Clonal_rate) :*** Μια ακέραια τιμή η οποία εκφράζει τον αριθμό των μεταλλαγμένων κλώνων που επιτρέπεται να παράγει ένα ARB. Επίσης χρησιμοποιείται και για την ανάθεση πόρων σε ένα ARB.
- ***Μόνιμα κύτταρα μνήμης (established memory cells):*** Το αντίσωμα ενός ARB το οποίο έχει επιβιώσει από τον ανταγωνισμό για τους πόρους, είναι το περισσότερο ερεθισμένο από ένα δοθέν αντιγόνο από το σύνολο εκπαίδευσης και έχει προστεθεί στο τελικό σύνολο των κυττάρων μνήμης.
- ***Διάνυσμα χαρακτηριστικών γνωρισμάτων (feature vector):*** Ένα παράδειγμα δεδομένων αναπαρίσταται σαν μια ακολουθία τιμών. Κάθε θέση στην ακολουθία αναπαριστά ένα διαφορετικό χαρακτηριστικό που

σχετίζεται με τα δεδομένα και κάθε χαρακτηριστικό έχει το δικό του εύρος τιμών.

- ***Hyperclonal_rate :*** Μια ακέραια τιμή η οποία χρησιμοποιείται για να καθορίσει τον αριθμό των μεταλλαγμένων κλώνων όπου ένα δοθέν κύτταρο μνήμης επιτρέπετε να εισάγει στον κυτταρικό πληθυσμό. Εδώ, ένα επιλεγμένο κύτταρο μνήμης εισάγει τουλάχιστον Hyperclonal_rate* Clonal_rate* stimulation value μεταλλαγμένους κλώνους στον πληθυσμό των κυττάρων.
- ***Κύτταρο μνήμης (memory cell):*** Το αντίσωμα όπου στο τέλος της διαδικασίας εκπαίδευσης είναι το περισσότερο ερεθισμένο από την επαφή του με το δοθέν αντιγόνο και ταιριάζει περισσότερο μαζί του.
- ***Ποσοστό μετάλλαξης (Mutation_rate):*** Μια παράμετρος στο διάστημα [0,1], η οποία μας δίνει την πιθανότητα μετάλλαξης μίας συνιστώσας του διανύσματος χαρακτηριστικών.
- ***Μέγιστοι πόροι (Max_resources):*** Μια παράμετρος που οριοθετεί το πλήθος των ARBs που επιτρέπεται να υπάρχουν στο σύστημα. Σε κάθε ARB ανατίθεται ένας αριθμός πόρων βάσει της διέγερσης του στο τρέχον αντιγόνο (stimulation value). Υπάρχει ένα όριο για τους συνολικούς πόρους του συστήματος, αν καταναλώνονται περισσότεροι πόροι από τους επιτρεπόμενους τότε απομακρύνονται πόροι από τα λιγότερο ερεθισμένα ARBs μέχρι να φτάσουμε στο επιτρεπόμενο όριο. Αν αφαιρεθούν όλοι οι πόροι από ένα ARB τότε αφαιρείται και αυτό από τον κυτταρικό πληθυσμό.
- ***Stimulation Value:*** Μια τιμή στο διάστημα [0,1], η οποία προέρχεται από μια διαδικασία μέτρησης της απόκρισης ενός ARB σε ένα αντιγόνο.
- ***Stimulation threshold:*** Παράμετρος στο διάστημα [0,1], που λειτουργεί ως κριτήριο τερματισμού για την εκπαίδευση σε κάθε αντιγόνο.
- ***Σύνολο ελέγχου (Test set):*** Το σύνολο των αντιγόνων που χρησιμοποιούνται για να αξιολογήσουν την απόδοση της ταξινόμησης από τον αλγόριθμο.
- ***Σύνολο εκπαίδευσης (Training set):*** Το σύνολο των αντιγόνων που χρησιμοποιούνται για να εκπαιδεύσουν τον αλγόριθμο.

3.2.3 Ανάλυση του αλγορίθμου AIRS

3.2.3.1 Συνοπτική περιγραφή του αλγορίθμου AIRS



Εικόνα 9 Λειτουργία του αλγορίθμου AIRS.

Στην παραπάνω εικόνα βλέπουμε διαγραμματικά τα βήματα του αλγορίθμου τα οποία έχουν ως εξής:

Σε πρώτο στάδιο γίνεται η αρχικοποίηση του αλγορίθμου. Ο αλγόριθμος εκτελείται παρουσιάζοντας κάθε ένα από τα διαθέσιμα δεδομένα του υποσυνόλου εκπαίδευσης (τα οποία θα καλούνται αντιγόνα) μία μόνο φορά στα υπολογιστικά αντίστοιχα των ανοσοποιητικών κυττάρων (κύτταρα μνήμης –mc- και Β-κύτταρα –ARBs). Για κάθε ένα αντιγόνο που εμφανίζεται, εκτελούνται τα παρακάτω βήματα:

1. Εντοπίζεται το κύτταρο μνήμης (mc_{match}) από το σύνολο MC των κυττάρων μνήμης το οποίο έχει την μέγιστη διέγερση στο τρέχον αντιγόνο.

2. Το mc_{match} παράγει ένα πλήθος μεταλλαγμένων κλώνων του εαυτού του, σε μία προσπάθεια να αυξήσει την συγγένεια του με το τρέχον αντιγόνο.

3. Το mc_{match} και οι κλώνοι του μεταφέρονται στο σύνολο MU και στη συνέχεια στο σύνολο AB το οποίο περιέχει τους εναπομείναντες κλώνους από προηγούμενα ενεργοποιημένα κύτταρα μνήμης, διεγερμένα από προηγούμενα αντιγόνα.

4. Το σύνολο AB εξελίσσεται με δύο τρόπους: απομακρύνοντας B-κύτταρα τα οποία δεν επιδεικνύουν αυξημένη συγγένεια με τα αντιγόνα και δημιουργώντας νέα κύτταρα μνήμης τα οποία θα προστεθούν στο σύνολο MC.

Ψάχνουμε στο σύνολο, το οποίο περιέχει κύτταρα μνήμης από προηγούμενες επαφές με το αντιγόνο και βρίσκουμε την τιμή διέγερσης για τα κύτταρα που ταιριάζουν περισσότερο με το αντιγόνο. Έπειτα υπολογίζουμε πόσοι κλώνοι θα παραχθούν και τους εισάγουμε σε ένα άλλο σύνολο. Το σύνολο αυτό περιέχει όλα εκείνα τα κύτταρα μνήμης που ταιριάζουν περισσότερο με το αντιγόνο και διεγείρονται από αυτό, αλλά και όλους τους κλώνους τους. Όλα αυτά τα στοιχεία, μεταφέρονται σε ένα νέο σύνολο το AB, το οποίο περιέχει και πιθανούς κλώνους από τα ARBs. Στο σύνολο αυτό υπολογίζονται οι πόροι και γίνεται ο έλεγχος για το ποια θα αποβληθούν από το σύστημα. Αυτά που θα απομακρυνθούν, είναι εκείνα με τις μικρότερες απαιτήσεις πόρων. Ορισμένα άλλα με βάση την ομοιότητα τους με το αντιγόνο μεταφέρονται στο MC σύνολο, ώστε να υπάρχουν σαν κύτταρα μνήμης για μια άλλη εφαρμογή του αλγορίθμου. Και τα υπόλοιπα μεταφέρονται και πάλι στο MU για να επαναληφθεί η διαδικασία, να παραχθούν δηλαδή, ξανά κλώνοι. Από το MC σύνολο, απομακρύνονται ορισμένα κύτταρα μνήμης, ώστε να μείνουν εκείνα που ταιριάζουν περισσότερο με το συγκεκριμένο αντιγόνο. Η διαδικασία αυτή περιγράφεται στο παραπάνω διάγραμμα και αναλύεται στις επόμενες παραγράφους.

3.2.3.2 Αρχικοποίηση

Σ’ αυτό το πρώτο βήμα του αλγορίθμου, βρίσκουμε την ακτίνα τιμών του διανύσματος χαρακτηριστικών των αντιγόνων του συνόλου AG, δηλαδή του συνόλου εκπαίδευσης. Η ακτίνα τιμών θα χρησιμοποιηθεί στη συνέχεια για να υπολογίσουμε την χειρότερη τιμή συγγένειας. Την χειρότερη τιμή συγγένειας την υπολογίζουμε ως την σχέση συγγένειας του μέγιστου και του ελάχιστου παράγοντα, όπως περιγράφεται στον τύπο που ακολουθεί.

Κατά την αρχικοποίηση του αλγόριθμου υπολογίζουμε το κατώφλι συγγένειας, ως την μέση συγγένεια μεταξύ όλων των δυνατών ζευγών δεδομένων του υποσυνόλου εκπαίδευσης, με βάση τον τύπο:

$$affinity\_threshold = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} affinity(agi, agj)}{n(n-1)/2} \qquad (1)$$

Η συνάρτηση $affinity(ag_i, ag_j)$ επιστρέφει την συγγένεια μεταξύ δυο αντιγόνων i και j, ως n αναφέρεται το συνολικό πλήθος των αντιγόνων. Η συγγένεια ορίζεται ως η ευκλείδεια απόσταση των δυο αντιγόνων, όπως περιγράφεται στον τύπο:

$$affinity = \sqrt{\sum_{k=1}^{N_D}\left(x_j(k) - x_i(k)\right)^2} \qquad (2)$$

όπου x_i , y_i είναι τα διανύσματα χαρακτηριστικών των αντιγόνων ag_i και ag_j με διάσταση N_D.

$$worst\_affinity = affinity(\max par, \min par) \qquad (3)$$

όπου, $affinity(\max par, \min par)$ είναι η συγγένεια μεταξύ των δυο στοιχείων του συνόλου {AG} με την μέγιστη ευκλείδεια απόσταση.

Επίσης ορίζουμε δυο νέα σύνολα, το πρώτο εκ των οποίων θα περιέχει όλα εκείνα τα στοιχεία του συνόλου ελέγχου που ανήκουν στην ίδια κλάση με το αντιγόνο και υπολογίζουμε το μήκος τους. Στο σημείο αυτό ξεκινάει η διαδικασία του αλγορίθμου για το κάθε ένα αντιγόνο ag_j του συνόλου {AG}. Για να αντιμετωπίσουμε την στοχαστική φύση του αλγορίθμου, αλλάζουμε την σειρά εμφάνισης των αντιγόνων για την κάθε επανάληψη, ώστε να έχουν νόημα τα αλλεπάλληλα τρεξίματα.

Ξεκινώντας την εκτέλεση του αλγορίθμου, παρουσιάζονται στο σύστημα όλα τα αντιγόνα του υποσυνόλου εκπαίδευσης. Για κάθε ένα από αυτά, εκτελούνται οι παρακάτω λειτουργίες.

Ελέγχεται το σύνολο MC και εντοπίζεται το κύτταρο μνήμης με την ισχυρότερη διέγερση από το τρέχον αντιγόνο. Εάν δεν υπάρχει προσθέτουμε στο MC σύνολο, το τρέχον αντιγόνο και θεωρούμε ότι έχουμε ένα ταυτόσημο αντίσωμα (mc_{match}). Σε αντίθετη περίπτωση, εντοπίζονται όλα τα στοιχεία του MC, τα οποία ανήκουν στην ίδια κλάση με το αντιγόνο. Εάν δεν υπάρχουν τέτοια στοιχεία, τότε το αντιγόνο προστίθεται στο MC . Εάν όμως υπάρχουν στοιχεία που ανήκουν στην ίδια κλάση με το αντιγόνο, υπολογίζουμε τη συγγένεια μεταξύ του αντιγόνου και των

στοιχείων αυτών και εντοπίζεται το κύτταρο μνήμης mc_{match} με την μικρότερη τιμή συγγένειας, δηλαδή, την μεγαλύτερη ομοιότητα, όπως περιγράφεται στον παρακάτω τύπο.

$$stim\_match = \frac{\left(worst\_affinity - affinity\left(mc_{match}, antigen\right)\right)}{worst\_affinity} \qquad (4)$$

όπου $affinity\left(mc_{match}, antigen\right)$ είναι η συγγένεια του αντιγόνου με το mc_{match} κύτταρο μνήμης και $worst\_affinity$ είναι η χειρότερη τιμή συγγένειας που υπολογίζεται από τον τύπο (3).

3.2.3.3 Δημιουργία κυττάρων μνήμης

Στο στάδιο αυτό του αλγορίθμου, τα στοιχεία mc_{match} του συνόλου MC κλωνοποιούνται, υφίστανται μετάλλαξη και προωθούνται στο σύνολο MU. Πιο συγκεκριμένα, υπολογίζουμε αρχικά το πλήθος των κλώνων που επιτρέπεται να παραχθούν και αρχικοποιούμε ένα νέο σύνολο MU, το οποίο περιέχει το mc_{match} που έχει ήδη επιλεγεί καθώς και τους κλώνους του mc_{match}, το πλήθος των οποίων υπολογίζεται βάσει της διέγερσης του mc_{match} από το τρέχον αντιγόνο.

3.2.3.3.1 Λειτουργία της μετάλλαξης

Την διαδικασία μετάλλαξης, την υλοποιήσαμε σαν μια συνάρτηση, η οποία δέχεται ορίσματα το κύτταρο μνήμης mc_{match} , την ακτίνα τιμών για κάθε μία από τις συνιστώσες του δυανύσματος χαρακτηριστικών, την τιμή της μεταβλητής mutation_rate, η οποία είναι η πιθανότητα μετάλλαξης κάθε συνιστώσας του διανύσματος χαρακτηριστικών και τον αριθμό των κλάσεων. Ως έξοδος επιστρέφεται το μεταλλαγμένο κύτταρο mc_{match} , καθώς και μια λογική μεταβλητή που επισημαίνει αν έγινε επιτυχώς η μετάλλαξη.

Πιο αναλυτικά, η συνάρτηση αυτή λειτουργεί ως εξής. Ορίζω δύο στοιχεία change και change_to, τα οποία είναι διανύσματα με μήκος ίσο με το διάνυσμα χαρακτηριστικών που παίρνουν τυχαίες τιμές στο διάστημα [0,1] . Εντοπίζονται όλα εκείνα τα στοιχεία του change, τα οποία είναι μικρότερα από το ποσοστό μετάλλαξης (mutation_rate). Εάν υπάρχουν τέτοια στοιχεία πραγματοποιείται η μετάλλαξη επιλέγοντας τυχαία μία τιμή στο επιτρεπόμενο διάστημα τιμών, με ομοιόμορφη κατανομή. Στο σημείο αυτό πρέπει να επισημανθεί ότι ο αλγόριθμος AIRS επιτρέπει μετάλλαξη και της κλάσης του διανύσματος χαρακτηριστικών, η οποία υπολογιστικά θεωρείται ως η πρώτη συνιστώσα του διανύσματος. Για την απόφαση μετάλλαξης της κλάσης χρησιμοποιείται ξεχωριστή τιμή πιθανότητας από ότι για τις λοιπές συνιστώσες του διανύσματος χαρακτηριστικών.

Περιγραφή

Είσοδοι:

- Τρέχον αντιγόνο, μέγιστο και ελάχιστο στοιχείο, ποσοστό μετάλλαξης και πλήθος κλάσεων

Έξοδοι:

- Το μεταλλαγμένο αντιγόνο και μια λογική μεταβλητή

Βήματα της μεθόδου:

1. Δημιουργία δυο συνόλων με τυχαίες τιμές.
2. Δημιουργία συνόλου που περιέχει τα στοιχεία εκείνα με τιμή μικρότερη από το ποσοστό μετάλλαξης
3. Εάν υπάρχουν τέτοια στοιχεία
   i. Πραγματοποιείται η μετάλλαξη
   ii. Και το νέο αντιγόνο προκύπτει με βάση τον μέγιστο και ελάχιστο παράγοντα
4. Σε αντίθετη περίπτωση δεν πραγματοποιείται η μετάλλαξη

Σύνταξη

```
[y,mut] = mutate(x,minparam,maxparam,mutation_rate,nclass)
```

3.2.3.4 Ανταγωνισμός για πόρους

Ο ανταγωνισμός για πόρους του συστήματος λαμβάνει χώρα στο σύνολο AB, το οποίο περιλαμβάνει το mc_{match}, τις μεταλλάξεις αυτού και πιθανόν και άλλα ARBs από προηγούμενα ερεθίσματα από αντιγόνα.
Ο στόχος του αλγορίθμου είναι να εξελίσσει το σύνολο κυττάρων μνήμης πεπερασμένου πλήθους, ώστε να είναι επιτυχέστερα στην ταξινόμηση των αντιγόνων. Ο τρόπος να το πετύχει, είναι μέσω του ανταγωνισμού των ARBs για τους πόρους του συστήματος. Οι πόροι διατίθενται σε ένα ARB βάσει της τιμής διέγερσης του από το τρέχον αντιγόνο, η οποία χρησιμοποιείται σαν ένδειξη για την αναγνώριση του αντιγόνου. Εάν με τη εκχώρηση των πόρων ξεπεράσουμε τον μέγιστο επιτρεπόμενο αριθμό πόρων του συστήματος, απομακρύνονται πόροι από τα λιγότερο διεγερμένα ARB.

3.2.3.4.1 Υπολογισμός της διέγερσης ενός B-κυττάρου

Για να πραγματοποιήσουμε την διαδικασία αυτή, υλοποιήσαμε μια συνάρτηση, η οποία παίρνει ορίσματα το σύνολο AB, το τρέχον αντιγόνο και την χειρότερη τιμή συγγένειας. Ως έξοδο, μας επιστρέφει έναν πίνακα με τις τιμές διέγερσης των στοιχείων του AB. Η συνάρτηση λειτουργεί ως εξής. Ορίζω δυο

σύνολα, το $ind1$ και το $ind2$, τα οποία περιέχουν τα στοιχεία του AB που ανήκουν στην ίδια κλάση με το τρέχον αντιγόνο και τα στοιχεία του AB που ανήκουν σε διαφορετική κλάση από το τρέχον αντιγόνο αντίστοιχα. Για το καθένα σύνολο, η τιμή διέγερσης υπολογίζεται με διαφορετικό τρόπο, όπως φαίνεται στον τύπο (5).

$$stim\_AB = \begin{cases} \dfrac{\left(worst\_affinity - affinity\left(AB(ind_1), antigen\right)\right)}{worst\_affinity}, & \text{αν η κλάση του AB είναι ίδια με την κλάση του αντιγόνου} \\ \dfrac{affinity\left(AB(ind_2), antigen\right)}{worst\_affinity}, & \text{αλλιώς} \end{cases} \quad (5)$$

όπου $affinity(AB(ind1).antigen)$ είναι η σχέση συγγένειας μεταξύ των στοιχείων της πρώτης ομάδας και του αντιγόνου και $affinity(AB(ind2), antigen)$ είναι η συγγένεια μεταξύ των στοιχείων της δεύτερης ομάδας και του αντιγόνου.

Περιγραφή

Είσοδοι:

- Σύνολο AB, τρέχον αντιγόνο, χειρότερη τιμή συγγένειας

Έξοδοι:

- Τιμή διέγερσης για τα στοιχεία του AB

Βήματα της μεθόδου:

1. Δημιουργία δυο συνόλων. Το ένα περιέχει τα στοιχεία που ανήκουν στην ίδια κλάση με το αντιγόνο και το δεύτερο περιέχει αυτά που δεν ανήκουν στην ίδια κλάση.
2. Αρχικοποίηση της τιμής διέγερσης
3. Εύρεση της τιμής διέγερσης με διαφορετικό τρόπο για την κάθε μια ομάδα

Σύνταξη

```
stim_AB = Stimulation(AB, antigen, worst_affinity)
```

3.2.3.4.2 Υπολογισμός των απαιτούμενων πόρων για κάθε ARB

Μετά τον υπολογισμό της τιμής διέγερσης (5), υπολογίζουμε τους πόρους που απαιτούνται για κάθε ένα από τα στοιχεία του AB. Ο υπολογισμός γίνεται με τον

τύπο (6), όπου $stim\_AB$ είναι η τιμή που έχουμε σαν έξοδο από την προηγούμενη συνάρτηση και $clonal\_rate$ το ποσοστό των κλώνων που θα παραχθούν.

$$resources = clonal\_rate * stim\_AB \qquad (6)$$

3.2.3.4.3 Απομάκρυνση των μη διεγερμένων ARBs

Το μέγιστο επιτρεπόμενο μέγεθος πόρων για κάθε Β-κύτταρο ορίζεται σύμφωνα με τον ακόλουθο τύπο, ανάλογα με την κλάση του:

$$res\_allowed = \begin{cases} \frac{\max\_resources}{2}, \text{για ARB κλάση ίδια με του αντιγόνου} \\ \frac{\max\_resources}{2*(nclass-1)}, \text{ για ARB κλάση διάφορη του αντιγόνου} \end{cases} \qquad (7)$$

όπου, $\max\_resources$ είναι οι μέγιστοι πόροι του συστήματος και $nclass$ το πλήθος των κλάσεων.

Υπολογίζεται το απαιτούμενο ύψος των πόρων σύμφωνα με την (6). Στη συνέχεια δημιουργούνται 2 πίνακες που περιέχουν τα ARBs διατεταγμένα σε φθίνουσα σειρά ως προς τους πόρους τους, ως εξής: ο πρώτος πίνακας περιέχει τα ARBs με κλάση ίδια με αυτή του αντιγόνου και ο δεύτερος πίνακας τα λοιπά ARBs. Με μία επαναληπτική δομή αφαιρούνται οι πόροι που υπερβαίνουν τους μέγιστους επιτρεπόμενους από το τέλος του κάθε πίνακα (δηλ. από τα ARBs με το μικρότερο πλήθος πόρων). Η συνάρτηση που υλοποιεί την παραπάνω διαδικασία παίρνει ως ορίσματα το σύνολο AB, τους πόρους των ARBs, τον μέγιστο αριθμό πόρων του συστήματος, το πλήθος των κλάσεων και το τρέχον αντιγόνο. Έξοδος της συνάρτησης αυτής είναι το τροποποιημένο AB σύνολο, το πλήθος των πόρων που έχουν παραμείνει και οι τιμές διέγερσης των στοιχείων του AB που έχουν παραμείνει.

Περιγραφή

Είσοδοι:

- Σύνολο AB, τρέχον αντιγόνο, τιμή διέγερσης για τα στοιχεία του AB, πόροι που απαιτούνται, μέγιστοι επιτρεπόμενοι πόροι και πλήθος των κλάσεων

Έξοδοι:

- Το νέο σύνολο AB, οι πόροι που παρέμειναν και η τιμή διέγερσης των στοιχείων που παρέμειναν στο σύνολο AB

Βήματα της μεθόδου:

1. Υπολογισμός των επιτρεπόμενων πόρων για κάθε κλάση
2. Δημιουργία συνόλου που περιέχει τα στοιχεία του AB που ανήκουν στην ίδια κλάση με το αντιγόνο
3. Εάν υπάρχουν τέτοια στοιχεία
    i. Ταξινομούνται οι πόροι αθροιστικά
    ii. Υπολογίζονται οι πόροι που πρέπει να αφαιρεθούν
    iii. Δημιουργία συνόλου που περιέχει τους πόρους με τιμή μικρότερη ή ίση από αυτούς που πρέπει να αφαιρεθούν
    iv. Αφαιρούνται οι πόροι με την μικρότερη αθροιστική τιμή.

Σύνταξη

```
[AB,resources,stim]=candidate_mc(AB,ag,stim,resources,max_reso
urces,nclass)
```

3.2.3.4.4 Ορισμός κριτηρίου τερματισμού της εκπαίδευσης στο τρέχον αντιγόνο

Η διαδικασία εκπαίδευσης των Β-κυττάρων στο τρέχον αντιγόνο, η οποία περιλαμβάνει τα παραπάνω βήματα, επαναλαμβάνεται έως να ικανοποιηθεί ένα κριτήριο τερματισμού. Το κριτήριο αυτό περιλαμβάνει τον ικανοποιητικό βαθμό διέγερσης κάθε ενός από τα Β-κύτταρα του συνόλου AB, στο τρέχον αντιγόνο. Ο ορισμός της διέγερσης (5) εξασφαλίζει ότι τα Β-κύτταρα ίδιας κλάσης με το τρέχον αντιγόνο θα το αναγνωρίζουν, ενώ αυτά με διαφορετική κλάση δεν θα είναι σε θέση να το αναγνωρίζουν, έτσι ώστε να μην εμφανίζονται ψευδώς θετικές, ή ψευδώς αρνητικές κατηγοριοποιήσεις. Έτσι αν ag το αντιγόνο, nc το πλήθος των κλάσεων, N_i το πλήθος των Β-κυττάρων του συνόλου AB με κλάση i, s_0 το κατώφλι διέγερσης και $stim\_AB_j$ η τιμή της διέγερσης του j-οστού Β-κυττάρου με το τρέχον αντιγόνο, το κριτήριο τερματισμού της εκπαίδευσης (ωρίμανσης) των Β-κυττάρων ορίζεται ως εξής:

$$s_i > s_0, i = 1,...,nc$$
$$s_i = \frac{1}{N_i}\sum_{j=1}^{N_i} stim\_AB_j \qquad (8)$$

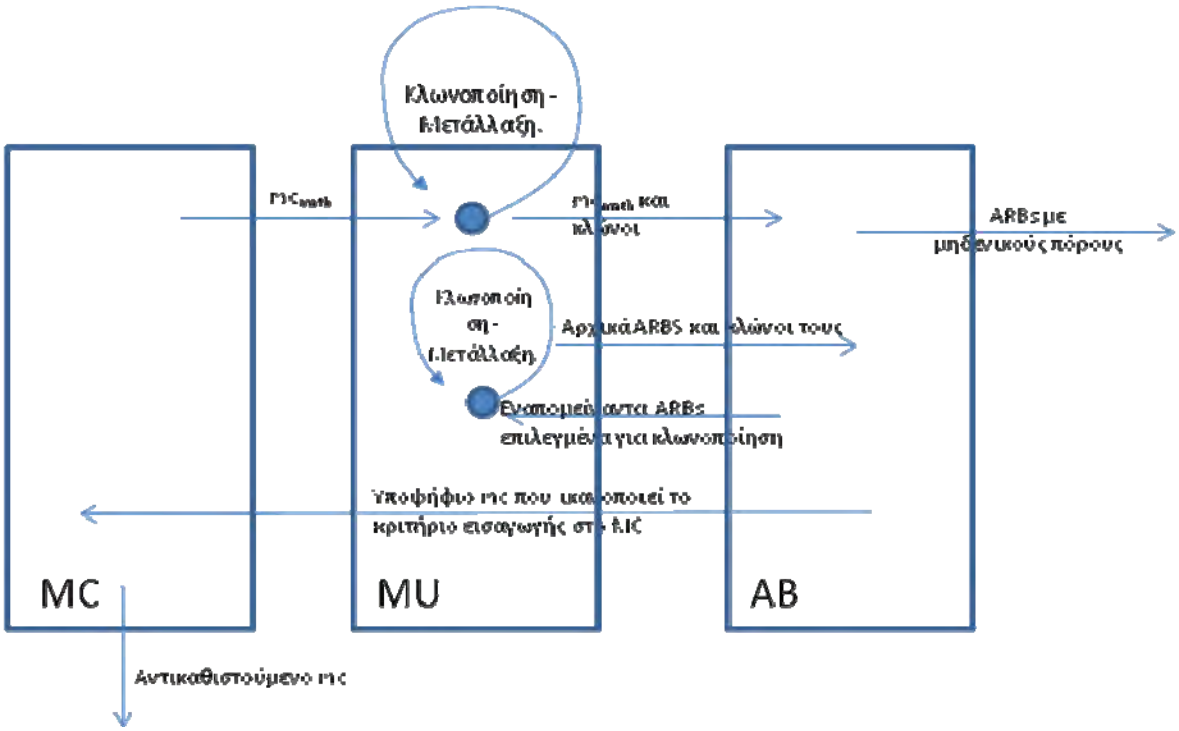
3.2.3.4.5 Παραγωγή κλώνων από τα στοιχεία του AB

Σε περίπτωση που το κριτήριο (8) δεν ικανοποιείται, επόμενη ενέργεια είναι η παραγωγή κλώνων από τα στοιχεία του AB αυτή τη φορά. Ορίζω λοιπόν μια νέα μεταβλητή η οποία θα είναι ένας πίνακας με τυχαίες τιμές και μέγεθος όσο το μέγεθος του $stim\_AB$ που βρήκαμε προηγουμένως. Επίσης δημιουργώ έναν πίνακα ο οποίος θα περιέχει όλα τα στοιχεία του AB, που έχουν τιμή διέγερσης, μεγαλύτερη από την παραπάνω τυχαία μεταβλητή. Για κάθε στοιχείο του πίνακα αυτού, υπολογίζω το πλήθος των κλώνων για τα αντίστοιχα στοιχεία του AB. Στη συνέχεια κάθε ένα στοιχείο από αυτά μεταλλάσσεται με την συνάρτηση mutate που περιγράψαμε νωρίτερα. Εάν η μετάλλαξη πραγματοποιήθηκε επιτυχώς στο MU σύνολο, το οποίο μηδενίσαμε, προσθέτουμε τους κλώνους. Στην πορεία υπολογίζουμε την τιμή διέγερσης για τα στοιχεία του MU με την ίδια συνάρτηση που χρησιμοποιήσαμε πιο πάνω, την Stimulation. Το AB σύνολο με το τέλος αυτής της διαδικασίας περιέχει και το νέο MU σύνολο. Επίσης υπολογίζουμε τους πόρους και την τιμή διέγερσης για το MU σύνολο.

3.2.3.5 Επιλογή υποψήφιων κυττάρων μνήμης από το AB και εισαγωγή τους στο MC

Σε περίπτωση που το κριτήριο (8) ικανοποιείται, ακολουθεί το τελευταίο στάδιο στην διαδικασία εκπαίδευσης, το οποίο είναι η επιλογή των υποψήφιων κυττάρων μνήμης και η εισαγωγή αυτών που ικανοποιούν το σχετικό κριτήριο, στο σύνολο των ήδη υπαρχόντων κυττάρων μνήμης, δηλαδή στο MC. Θεωρώ ως υποψήφιο κύτταρο το στοιχείο του AB, το οποίο ανήκει στην ίδια κλάση με το τρέχον αντιγόνο και η τιμή διέγερσής του (stimulation) είναι μεγαλύτερη από τις τιμές διέγερσης των άλλων στοιχείων του AB που ανήκουν στην ίδια κλάση με αυτό. Εάν η τιμή του υποψήφιου είναι μεγαλύτερη από την τιμή της διέγερσης του mc_{match}, τότε υπολογίζουμε την συγγένεια μεταξύ των δυο αυτών κυττάρων. Εάν το αποτέλεσμα της συγγένειας βρεθεί μικρότερο από το κατώφλι που ορίζει η παράμετρος *Affinity threshold scalar* (βλ. παρ. 3.2.2), τότε το mc_match απομακρύνεται από το MC και στη θέση του προστίθεται το υποψήφιο κύτταρο μνήμης. Σε αντίθετη περίπτωση, το υποψήφιο κύτταρο μνήμης προστίθεται στο σύνολο MC, χωρίς να απομακρυνθεί κάποιο υπάρχον κύτταρο μνήμης

Εφόσον το υποψήφιο κύτταρο μνήμης αξιολογήθηκε για να προστεθεί στο σύνολο των καθιερωμένων κυττάρων μνήμης, ολοκληρώνεται η διαδικασία εκπαίδευσης για το αντιγόνο. Σειρά έχει το δεύτερο αντιγόνο από το σύνολο ελέγχου και η διαδικασία επαναλαμβάνεται έως ότου όλα τα αντιγόνα να έχουν παρουσιαστεί στο σύστημα.



Εικόνα 10 Λειτουργία Αλγορίθμου. Η μετακίνηση των κυττάρων και η παραγωγή των κλώνων τους, από το ένα σύνολο στο άλλο, μέχρι να καταλήξουμε στο καλύτερο σύνολο κυττάρων μνήμης.

3.2.3.6 Εφαρμογή του εκπαιδευμένου αλγόριθμου AIRS

Αφού ο αλγόριθμος AIRS έχει εκπαιδευτεί, εφαρμόζεται για νέα δεδομένα (αντιγόνα του συνόλου ελέγχου) ως εξής:
Για κάθε νέο αντιγόνο: εντοπίζεται το κύτταρο μνήμης το οποίο έχει την ελάχιστη απόσταση από το αντιγόνο. Στο αντιγόνο ανατίθεται η κλάση του κοντινότερου κυττάρου μνήμης και στη συνέχεια υπολογίζεται η ακρίβεια της σωστής ταξινόμησης για κάθε κλάση, με βάση τα στοιχεία του συνόλου ελέγχου, που ανήκουν στην πρώτη κλάση και έχουν την μικρότερη απόσταση, από τα κύτταρα μνήμης της κλάσης αυτής και με βάση τα στοιχεία του συνόλου ελέγχου, που ανήκουν στην δεύτερη κλάση αντίστοιχα.

3.1.1 Μεταβολές στις τιμές των μεταβλητών

Ο αλγόριθμος AIRS που χρησιμοποιήσαμε, εξαρτάται από κάποιες μεταβλητές, τις οποίες παίρνει σαν ορίσματα κατά την κλήση του. Οι μεταβλητές αυτές είναι οι εξής:

1. *Max_resources*, μια παράμετρος που οριοθετεί τα ARBs βάσει των stimulation value τους και των Clonal_rate τους. Αντιστοιχεί στο όριο των συνολικών πόρων του συστήματος. Αν όλοι οι πόροι ενός ARB αφαιρεθούν, τότε αφαιρείται και αυτό από τον κυτταρικό πληθυσμό.

2. *Hyperclonal_rate*, μια ακέραια τιμή η οποία αντιστοιχεί στον αριθμό των μεταλλαγμένων κλώνων που ένα δοθέν κύτταρο μνήμης επιτρέπεται να εισάγει στον κυτταρικό πληθυσμό.

3. *Clonal_rate*, μια ακέραια τιμή η οποία εκφράζει τον αριθμό των μεταλλαγμένων κλώνων που επιτρέπεται να παράγει ένα χαρακτηριστικό. Επίσης χρησιμοποιείται και για την ανάθεση πόρων σε μια ARB.

4. *Mutation_rate*, μια παράμετρος στο διάστημα [0,1], η οποία μας δίνει την πιθανότητα μετάλλαξης (τυχαίας αλλαγής της τιμής) ενός χαρακτηριστικού.

5. *Stim_thres*, παράμετρος στο διάστημα [0,1] που λειτουργεί ως κριτήριο τερματισμού για την εκπαίδευση σε κάθε αντιγόνο

Οι αλλαγές στις τιμές των παραπάνω μεταβλητών αλλάζουν το αποτέλεσμα της εξόδου του αλγορίθμου. Θα ξεκινήσουμε αναλύοντας την πρώτη που αναφέραμε, την max_resources. Όπως αναφέραμε αντιστοιχεί στον μέγιστο αριθμό πόρων που μπορούν να υπάρξουν στο σύστημα. Αλλάζοντας λοιπόν την τιμή της μεταβλητής αυτής, αλλάζει και ο αριθμός των επιτρεπόμενων πόρων του συστήματος, οι οποίοι υπολογίζονται από τον τύπο (7). Κατά συνέπεια, αύξηση της τιμής της μεταβλητής αναμένεται να οδηγήσει στην αύξηση του πλήθους των ARB και της υπολογιστικής πολυπλοκότητας του αλγόριθμου. Διαδοχικά αλλάζει και το σύνολο AB που δημιουργείται μετά την απομάκρυνση των λιγότερο διεγερμένων στοιχείων.

Στην μελέτη μας αυτή έχουμε δώσει στην max_resources τιμές από 200 έως 1000 με βήμα 200 και οι μεταβολές φαίνονται στον πίνακα 4, όπου βλέπουμε τις διάφορες τιμές για την ακρίβεια της κάθε κλάσης, αλλά και της συνολικής ακρίβειας. Επίσης παρατηρούμε ότι ο χρόνος που απαιτείται για να ολοκληρωθεί η διαδικασία εκπαίδευσης αυξάνεται όσο αυξάνεται και ο αριθμός των μέγιστων πόρων, καθώς όσο μεγαλύτερος είναι ο αριθμός αυτός τόσο λιγότερα θα είναι τα στοιχεία που θα βγάλουμε από το AB σύνολο. Για την μεταβλητή max_resources, με βάση τον πίνακα 4, βλέπουμε ότι υπάρχουν δυο καλές τιμές, η μια είναι 600 και η άλλη 1000. Εμείς θα χρησιμοποιήσουμε την τιμή 600 για τους ελέγχους μας, καθώς έχει υψηλή ακρίβεια

για τις κλάσεις και την συνολική τους ακρίβεια, αλλά απαιτείται και λιγότερος χρόνος εκτέλεσης του αλγορίθμου.

Επόμενη μεταβλητή είναι η hyperclonal_rate, η οποία, όπως αναφέραμε και πιο πάνω ελέγχει τον αριθμό των κλώνων που θα δημιουργηθούν από τα κύτταρα μνήμης του MC πληθυσμού, για να περάσουν στον MU πληθυσμό. Η αλλαγή στην τιμή της μεταβλητής αυτής, οδηγεί στην αλλαγή του αριθμού των κλώνων, ο οποίος με τη σειρά του επηρεάζει τον MU πληθυσμό που θα δημιουργηθεί. Αύξηση της τιμής της μεταβλητής οδηγεί σε καλύτερη εξερεύνηση του χώρου των διανυσμάτων χαρακτηριστικών από τον αλγόριθμο. Η επίδραση στην ακρίβεια της κατηγοριοποίησης πιθανόν να μην είναι μονοτονική, καθώς πολύ αυξημένες τιμές μπορεί να οδηγήσουν σε «overtraining» του αλγορίθμου στα δεδομένα εκπαίδευσης και απώλεια της γενικότητας με αποτέλεσμα μη επιτυχή κατηγοριοποίηση των δεδομένων ελέγχου. Η επίδραση στο χρόνο εκτέλεσης δεν είναι δυνατό να προβλεφθεί, καθώς αναμένεται να επηρεαστεί η ταχύτητα σύγκλισης της φάσης εκπαίδευσης των B-κυττάρων. Εδώ θα δώσουμε στο hyperclonal_rate τιμές από το 5 έως το 20 με βήμα 5. Στον πίνακα 5 φαίνονται οι αλλαγές στις τιμές της ακρίβειας των κλάσεων για τις διάφορες τιμές που δώσαμε στη μεταβλητή. Η καλύτερη τιμή που προκύπτει από τον πίνακα αυτό είναι η 15, καθώς έχει καλή τιμή απόδοσης για την ακρίβεια των κλάσεων, μικρή τυπική απόκλιση όσον αφορά την επαναληψημότητα των αποτελεσμάτων της και σχετικά καλό χρόνο εκτέλεσης.

Σειρά έχει η μεταβλητή *Clonal_rate*, η οποία είναι η πιο σημαντική θα λέγαμε, καθώς επηρεάζει τόσο τον αριθμό των κλώνων όσο και τον αριθμό των πόρων που θα ανατεθούν σε κάποιο B-κύτταρο (ARB). Αύξηση της τιμής της παραμέτρου (χωρίς ταυτόχρονη αύξηση του μέγιστου πλήθους των διαθέσιμων πόρων του συστήματος) οδηγεί σε αυστηροποίηση της επιλογής των B-κυττάρων, μείωση του αριθμού τους και κατά συνέπεια σε ταχύτερη εκτέλεση του αλγορίθμου. Η επίπτωση στην ακρίβεια της κατηγοριοποίησης δεν είναι ασφαλές να προβλεθφεί. Το πεδίο τιμών της παραμέτρου είναι από το 5 έως το 20 με βήμα 5. Οι επιπτώσεις των αλλαγών των τιμών της, στην ακρίβεια των κλάσεων φαίνεται στον πίνακα 6. Όπως παρατηρούμε από τον πίνακα αυτόν, όσο αυξάνουμε την τιμή της μεταβλητής, τόσο αυξάνεται και η απόδοση του αλγορίθμου. Έχουμε περισσότερους πόρους με μεγαλύτερο clonal_rate αλλά και περισσότερους κλώνους για το πληθυσμό τόσο του MU όσο και του AB. Η καλύτερη της τιμή είναι η 20 με την καλύτερη απόδοση στην ακρίβεια των κλάσεων, τον καλύτερο χρόνο και με μικρή τυπική απόκλιση.

Τιμές max_resources	Ακρίβεια πρώτης κλάσης	Ακρίβεια δεύτερης κλάσης	Συνολική μέση ακρίβεια +- τυπική απόκλιση		Χρόνος εκτέλεσης
200	0.880	0.879	0.880	0.015	636
400	0.887	0.865	0.876	0.019	807
600	***0.861***	***0.882***	***0.870***	***0.014***	***907***
800	0.858	0.871	0.864	0.032	1008
1000	0.884	0.888	0.886	0.012	1175

Πίνακας 4 Πίνακας ακρίβειας των κλάσεων για τις διάφορες τιμές της μεταβλητής max_resources. Εφαρμογή του AIRS στα συνθετικά δεδομένα (δ).

Τιμές hyper_clonal_rate	Ακρίβεια πρώτης κλάσης	Ακρίβεια δεύτερης κλάσης	Συνολική μέση ακρίβεια +- τυπική απόκλιση		Χρόνος εκτέλεσης
5	0.904	0.874	0.888	0.015	840
10	0.866	0.863	0.864	0.023	768
15	***0.903***	***0.870***	***0.887***	***0.012***	***770***
20	0.866	0.856	0.861	0.029	894

Πίνακας 5 Πίνακας ακρίβειας των κλάσεων για τις διάφορες τιμές της μεταβλητής hyperclonal_rate. Εφαρμογή του AIRS στα συνθετικά δεδομένα (δ).

Τιμές clonal_rate	Ακρίβεια πρώτης κλάσης	Ακρίβεια δεύτερης κλάσης	Συνολική μέση ακρίβεια +- τυπική απόκλιση		Χρόνος εκτέλεσης
5	0.871	0.877	0.874	0.023	1007
10	0.866	0.863	0.864	0.023	858
15	0.884	0.875	0.880	0.027	763
20	***0.877***	***0.884***	***0.881***	***0.021***	***771***

Πίνακας 6 Πίνακας ακρίβειας των κλάσεων για τις διάφορες τιμές της μεταβλητής clonal rate. Εφαρμογή του AIRS στα συνθετικά δεδομένα (δ).

Μένουν λοιπόν δύο μεταβλητές ακόμα για να μελετήσουμε. Η μία είναι η Mutation_rate και η δεύτερη είναι η Stim_thres. Η πρώτη αναφέρεται στην πιθανότητα που έχει ένα αντιγόνο να υποστεί μετάλλαξη και η δεύτερη αντιστοιχεί σε ένα όριο τερματισμού για την εκπαίδευση του κάθε αντιγόνου. Επιπλέον η πρώτη, καθότι πιθανότητα, παίρνει τιμές στο διάστημα [0,1 , 1] με βήμα 0,1, ενώ η δεύτερη παίρνει τιμές 0,4 και 0,5. Η μεταβλητή Mutation rate, επηρεάζει την απόδοση του αλγορίθμου, καθώς από αυτήν εξαρτάται κατά πόσο θα πραγματοποιηθεί μετάλλαξη στους πληθυσμούς MU και AB. Οι μεταβολές στις τιμές της φαίνονται στον Πίνακας 7, από όπου διαπιστώνουμε ότι η καλύτερη τιμή της είναι η τιμή 0,5.

Όσον αφορά τώρα την μεταβλητή Stim_thres, οι μεταβολές στις τιμές της φαίνονται στον πίνακα 8. Η μεταβλητή αυτή επηρεάζει τον αλγόριθμο, καθώς το αν θα σταματήσει η διαδικασία εκπαίδευσης του τρέχοντος αντιγόνου, εξαρτάται από το αν το κριτήριο τερματισμού της διαδικασίας είναι μικρότερο ή όχι από την τιμή αυτής. Ως καλύτερη τιμή της μεταβλητής αυτής θεωρήσαμε την τιμή 0,5, για τους ίδιους λόγους που επιλέξαμε και παραπάνω.

Όλες οι παραπάνω τιμές δοκιμάστηκαν στα συνθετικά δεδομένα για να επιλέξουμε την καλύτερη τιμή για το κάθε ένα και έπειτα να την χρησιμοποιήσουμε στα πιο σύνθετα δεδομένα.

Τιμές mutation_rate	**Ακρίβεια κλάσης A**	**Ακρίβεια κλάσης B**	**Συνολική μέση ακρίβεια +- τυπική απόκλιση**		**Χρόνος εκτέλεσης**
0.1	0.894	0.871	0.879	0.018	658
0.2	0.869	0.886	0.878	0.016	637
0.3	0.885	0.854	0.869	0.018	648
0.4	0.861	0.884	0.873	0.018	691
0.5	***0.883***	***0.880***	***0.881***	***0.011***	***617***
0.6	0.878	0.868	0.873	0.019	695
0.7	0.854	0.881	0.868	0.032	709
0.8	0.862	0.873	0.868	0.026	819
0.9	0.868	0.857	0.862	0.026	871
1	0.854	0.863	0.859	0.021	1095

Πίνακας 7 Πίνακας ακρίβειας των κλάσεων για τις διάφορες τιμές της μεταβλητής mutation_rate. Εφαρμογή του AIRS στα συνθετικά δεδομένα (δ).

Τιμές stimulation_threshold	Ακρίβεια πρώτης κλάσης	Ακρίβεια δεύτερης κλάσης	Συνολική μέση ακρίβεια	+- τυπική απόκλιση	Χρόνος εκτέλεσης
0.4	0.841	0.888	0.865	0.021	683
0.5	***0.866***	***0.863***	***0.864***	***0.023***	***620***

Πίνακας 8 Πίνακας ακρίβειας των κλάσεων για τις διάφορες τιμές της μεταβλητής stim_thres. Εφαρμογή του AIRS στα συνθετικά δεδομένα (δ).

Παράμετρος αλγορίθμου	Βέλτιστη τιμή
max_resources	600
hyper_clonal_rate	15
clonal_rate	20
mutation_rate	0,5
stimulation_threshold	0,5

Πίνακας 9 Πίνακας βέλτιστων τιμών των παραμέτρων, όπως καταλήξαμε από τα παραπάνω αποτελέσματα.

4 Αποτελέσματα και συμπεράσματα

4.1 Αποτελέσματα

4.1.1 Αποτελέσματα από την χρήση των ΤΝΔ

Χρησιμοποιήσαμε τον αλγόριθμο των τεχνικών νευρωνικών δικτύων, για να μπορέσουμε να συγκρίνουμε την απόδοση της μεθόδου που μας ενδιαφέρει με την απόδοση των ΤΝΔ. Καθότι τα ΤΝΔ είναι πιο διαδεδομένα και χρησιμοποιούνται ευρέως για προβλήματα ταξινόμησης, τα θεωρήσαμε ένα αξιόπιστο μέτρο σύγκρισης.

Πραγματοποιήσαμε την εκτέλεση του αλγορίθμου, για τα συνθετικά δεδομένα, τα δεδομένα από το φυτό Iris και φυσικά τα κλινικά δεδομένα από βιοψίες. Και για τις τρείς περιπτώσεις, πριν γίνει η ταξινόμηση, χωρίσαμε τα δεδομένα, σε ξένα υποσύνολα ελέγχου και εκπαίδευσης με τυχαίο διαμοιρασμό σε διαφορές τιμές ποσοστών, με τον τρόπο που περιγράψαμε στην παράγραφο 2.2.

Λόγω της στοχαστικής φύσης του αλγορίθμου εκτελέσαμε πολλές επαναλήψεις για κάθε ρύθμιση των ΤΝΔ. Ο αλγόριθμος μας έδινε σαν έξοδο την ακρίβεια των δύο κλάσεων και στη συνέχεια υπολογίστηκε η συνολική ακρίβεια της κανονικοποίησης με χρήση του παρακάτω τύπου

$$Pcc = \frac{(accA * N1) + (accB * N2)}{N1 + N2} \qquad (11)$$

όπου Pcc είναι η συνολική ακρίβεια των δυο κλάσεων, $accA$ η ακρίβεια της πρώτης κλάσης και $accB$ η ακρίβεια της δεύτερης κλάσης. Επίσης $N1$ είναι το πλήθος των στοιχείων ελέγχου της πρώτης κλάσης και $N2$ το πλήθος των στοιχείων ελέγχου της δεύτερης κλάσης.

Για τα δύο πρώτα είδη δεδομένων, συνθετικά (δ) και Iris εκτελέσαμε τον αλγόριθμο για διαφορετικές τιμές νευρώνων, ώστε να διερευνήσουμε πότε επιτυγχάνεται η καλύτερη απόδοση του αλγορίθμου, ώστε να τον χρησιμοποιήσουμε στα πιο περίπλοκα κλινικά δεδομένα. Για τα δεδομένα από το φυτό Iris, χρειάστηκε να κάνουμε κανονικοποίηση του διανύσματος των χαρακτηριστικών, που έπαιρναν τιμές μεγαλύτερες της μονάδας, στο διάστημα [0,1].

Στον παρακάτω πίνακα 10, φαίνονται οι διάφορες τιμές ακρίβειας των κλάσεων για διαφορετικές τιμές νευρώνων. Όπως μπορούμε να συμπεράνουμε, καλύτερη απόδοση του αλγορίθμου αυτού, επιτυγχάνεται με την χρήση 7 νευρώνων σε ένα ενδιάμεσο στρώμα (hidden layer), αναφορικά με αμφότερα τα δεδομένα στα οποία τον εφαρμόσαμε. Διατηρώντας το πλήθος των νευρώνων στην τιμή 7, διαπιστώσαμε ότι αύξηση του πληθάριθμου του υποσυνόλου ελέγχου από 0.1 έως 0.4 των συνολικών δεδομένων οδηγεί (όπως είναι αναμενόμενο) σε αύξηση της ακρίβειας κατηγοριοποίησης (Πίνακας 11).

	ΙΡΙΣ			Συνθετικά Δεδομένα (δ)		
Αριθμός νευρώνων	**Μέση τιμή ακρίβειας κλάσης Α**	**Μέση τιμή ακρίβειας κλάσης Β**	**Μέση τιμή συνολικής ακρίβειας**	**Μέση τιμή ακρίβειας κλάσης Α**	**Μέση τιμή ακρίβειας κλάσης Β**	**Μέση τιμή συνολικής ακρίβειας**
1	0.690	1.000	0.815	0.908	0.911	0.910
3	0.812	0.906	0.854	0.916	0.921	0.918
5	0.838	0.793	0.805	0.924	0.920	0.922
7	**0.911**	**0.845**	**0.866**	**0.940**	**0.930**	**0.935**

Πίνακας 10 Πίνακας ακρίβειας των κλάσεων για διαφορετικό αριθμό νευρώνων με υλοποίηση ΤΝΔ, στα Iris και τα συνθετικά δεδομένα (δ) με ποσοστό test-train→0,6-0,4.

Test-train	**Μέση τιμή ακρίβειας πρώτης κλάσης**	**Μέση τιμή ακρίβειας δεύτερης κλάσης**	**Μέση τιμή συνολικής ακρίβειας**
0.9 – 0.1	0.945	0.940	0.943
0.8 – 0.2	0.933	0.942	0.938
0.7 – 0.3	0.936	0.939	0.938
0.6 – 0.4	**0.991**	**0.863**	**0.983**

Πίνακας 11 Πίνακας ακρίβειας των κλάσεων για διαφορετικό ποσοστό διαχωρισμού test και train συνόλου για τα non-suspicious δεδομένα, με εφαρμογή της μεθόδου ΤΝΔ (7 νευρώνες σε ένα ενδιάμεσο στρώμα).

Για την καλύτερη λοιπόν τιμή όσον αφορά τους νευρώνες, εφαρμόσαμε τον αλγόριθμο στα πλέον σημαντικά δεδομένα, τα κλινικά δεδομένα, από τις βιοψίες FNA θυρεοειδούς. Όπως αναφέραμε και κατά την περιγραφή τους, τα πραγματικά δεδομένα, χωρίζονται σε δυο κατηγορίες, τα suspicious (ύποπτα) και τα non-suspicious (μη ύποπτα). Εμείς αρχικά θα εργαστούμε με τα δεύτερα γιατί είναι ευκολότερο να ταξινομηθούν.

Στον πίνακα 13 φαίνονται οι διάφορες τιμές ακρίβειας των κλάσεων για τα non-suspicious δεδομένα, σε ποσοστό test-train→0.6-0.4 και πλήθος κρυφών νευρώνων 7. Όπως μπορούμε να διαπιστώσουμε η απόδοση του αλγορίθμου αυτού

είναι αρκετά ικανοποιητική με την ακρίβεια των κλάσεων να είναι σε αρκετά υψηλά. Η σύνθεση των υποσυνόλων εκπαίδευσης και ελέγχου δίνεται στον Πίνακα 12.

Πλήθος Επαναλήψεων	**Δεδομένα Ελέγχου**			**Δεδομένα Εκπαίδευσης**		
	Συνολικά	Κλάση Α	Κλάση Β	Συνολικά	Κλάση Α	Κλάση Β
1η	1174	1111	63	789	739	50
2η	1202	1130	72	761	720	41
3η	1157	1094	63	806	756	50
4η	1208	1144	64	755	706	49
5η	1189	1120	69	774	730	44
6η	1159	1092	67	804	758	46
7η	1188	1120	68	775	730	45
8η	1173	1103	70	790	747	43
9η	1131	1051	80	832	799	33
10η	1195	1123	72	768	727	41

Πίνακας 12 Πίνακας πλήθους στοιχείων των συνόλων ελέγχου και εκπαίδευσης, όπως αυτά δημιουργήθηκαν με τον διαχωρισμό, για τα non-suspicious δεδομένα.

Πλήθος επαναλήψεων	Ακρίβεια πρώτης κλάσης	Ακρίβεια δεύτερης κλάσης	Συνολική ακρίβεια
1η	0.995	0.930	0.992
2η	0.997	0.794	0.986
3η	0.993	0.871	0.986
4η	0.990	0.871	0.983
5η	0.981	0.921	0.978
6η	0.988	0.836	0.979
7η	0.995	0.855	0.986
8η	0.995	0.855	0.987
9η	0.981	0.850	0.975
10η	0.991	0.849	0.982
Μέση τιμή για τις 10 επαναλήψεις	0.991	0.863	0.983

Πίνακας 13 Πίνακας ακρίβειας των κλάσεων για 10 επαναλήψεις των ΤΝΔ για τα non_suspicious δεδομένα με ποσοστό διαχωρισμού test-train ίσο με 0,6-0,4 και 7 κρυφούς νευρώνες.

4.1.2 Αποτελέσματα από την χρήση του kNN

Με παρόμοιο τρόπο, όπως και με τα ΤΝΔ, εργαστήκαμε και με αυτόν τον αλγόριθμο. Κάναμε ελέγχους για όλα τα δεδομένα και εστιάσαμε και πάλι στα πραγματικά δεδομένα και την κατηγορία τους non_suspicious. Ο αλγόριθμος που δημιουργήσαμε για τον kNN, μας επιστρέφει ως έξοδο τις αληθώς θετικές και αρνητικές τιμές της μήτρας αληθείας. Για να υπολογίσουμε την ακρίβεια της κάθε κλάσης, διαιρέσαμε τις τιμές αυτές με το πλήθος των στοιχείων από το σύνολο ελέγχου που ανήκουν σε κάθε κλάση. Δηλαδή η ακρίβεια της πρώτης κλάσης υπολογίστηκε ως το πηλίκο της αληθώς σωστής τιμής (TP), προς το πλήθος των στοιχείων του συνόλου ελέγχου που ανήκουν στην πρώτη κλάση (N1) και η ακρίβεια της δεύτερης κλάσης, υπολογίστηκε ως το πηλίκο της αληθώς αρνητικής τιμής (TN), προς το πλήθος των στοιχείων του συνόλου ελέγχου που ανήκουν στη δεύτερη κλάση. Στη συνέχεια για να υπολογίσουμε την συνολική ακρίβεια των δυο κλάσεων χρησιμοποιήσαμε και πάλι τον τύπο (11).

Στον πίνακα 14 φαίνονται οι διάφορες τιμές ακρίβειας για τα συνθετικά δεδομένα μας αλλά και για τα δεδομένα από το φυτό Iris, για διάφορες τιμές στο πλήθος γειτόνων. Βλέπουμε ότι η απόδοση του αλγορίθμου αυτού για τα Iris δεδομένα δεν είναι αρκετά ικανοποιητική, σε αντίθεση βέβαια με αυτή των συνθετικών δεδομένων, η οποία φαίνεται να βρίσκεται σε αρκετά υψηλά επίπεδα. Με βάση τις τιμές ακρίβειας για τα συνθετικά δεδομένα θεωρούμε ως την καλύτερη απόδοση εκείνη για αριθμό γειτόνων k=3.

Ένας άλλος παράγοντας, που επηρεάζει την απόδοση του αλγορίθμου αυτού, πέραν από το πλήθος των γειτόνων, είναι και το ποσοστό διαχωρισμού των δεδομένων σε σύνολο ελέγχου (test data) και εκπαίδευσης (train data). Στον πίνακα 15, φαίνονται οι μέσες τιμές της ακρίβειας των κλάσεων για δέκα επαναλήψεις. Όπως μπορούμε να συμπεράνουμε, ο αλγόριθμος kNN έχει καλύτερη απόδοση όταν έχουμε μικρό ποσοστό δεδομένων εκπαίδευσης και μεγάλο ποσοστό δεδομένων ελέγχου. Ως καλύτερη τιμή θα θεωρήσουμε την 0.8 – 0.2, η οποία έχει υψηλή απόδοση αλλά έχει και μεγαλύτερο ποσοστό δεδομένων εκπαίδευσης σε σύγκριση με την δεύτερη καλύτερη τιμή, την 0.9 – 0.1.

	Iris			**Συνθετικά δεδομένα (δ)**			
Αριθμός γειτόνων	**Μέση ακρίβεια**		**Συνολική ακρίβεια**	**Μέση ακρίβεια**		**Συνολική ακρίβεια**	**Τυπική από-κλιση**
	Κλάση Α	**Κλάση Β**		**Κλάση Α**	**Κλάση Β**		
1	0.782	0.816	0.798	0.964	0.958	0.961	0.003
3	***0.708***	***0.868***	***0.785***	***0.962***	***0.966***	***0.964***	***0.001***
5	0.266	0.977	0.618	0.964	0.963	0.964	0.003
7	0.492	0.959	0.726	0.960	0.961	0.960	0.004

Πίνακας 14 Πίνακας μέσης τιμής ακρίβειας των κλάσεων για τα συνθετικά (δ) και τα Iris δεδομένα, για διαφορετικό αριθμό γειτόνων, με χρήση kNN και ποσοστό διαχωρισμού test-train→0,8-0,2.

Test-train	Μέση τιμή ακρίβειας		Μέση τιμή συνολικής ακρίβειας
	Κλάση Α	Κλάση Β	
0.9 – 0.1	0.830	0.876	0.868
0.8 – 0.2	***0.863***	***0.916***	***0.867***
0.7 – 0.3	0.479	0.948	0.505
0.6 – 0.4	0.192	0.977	0.235

Πίνακας 15 Πίνακας τιμών ακρίβειας των κλάσεων για διαφορετικά ποσοστά διαχωρισμού των συνόλων Test-train, για τον kNN με πλήθος γειτόνων k=3, για τα δεδομένα non-suspicious.

Πλήθος επαναλήψεων	Μέση ακρίβεια 1ης κλάσης	Μέση ακρίβεια 2ης κλάσης	Συνολική μέση ακρίβεια
1η	0.957	0.911	0.955
2η	0.963	0.905	0.960
3η	0.959	0.897	0.956
4η	0.959	0.902	0.956
5η	0.962	0.915	0.959
6η	0.959	0.891	0.959
7η	0.958	0.912	0.955
8η	0	1.000	0.061
9η	0.958	0.912	0.956
10η	0.958	0.915	0.956
Μέση τιμή για τις 10 επαναλήψεις	0.863	0.916	0.867

Πίνακας 16 Πίνακας ακρίβειας των κλάσεων για 10 επαναλήψεις του kNN για τα non_suspicious δεδομένα με ποσοστό διαχωρισμού 0,8-0,2 και πλήθος γειτόνων k=3.

Από τον παραπάνω πίνακα μπορούμε να δούμε ότι και ο kNN αλγόριθμος, έχει αξιοσημείωτη απόδοση με τις τιμές της ακρίβειας των κλάσεων να φτάνει σε επίπεδο 96%. Υπάρχουν βέβαια και χαμηλές τιμές αλλά στο συγκεκριμένο ποσοστό διαχωρισμού των δεδομένων σε δεδομένα ελέγχου και εκπαίδευσης, οι τιμές είναι ικανοποιητικές.

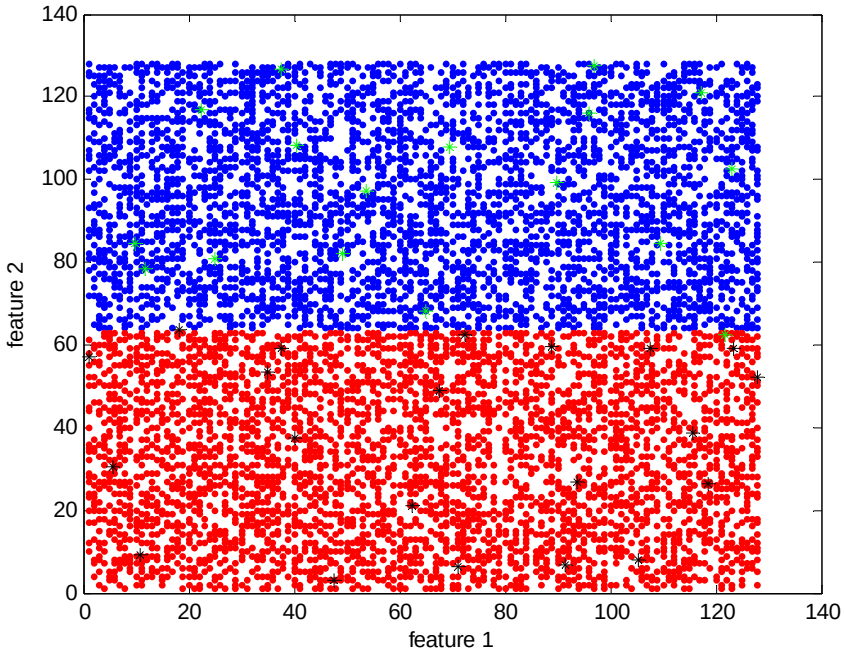
4.1.3 Αποτελέσματα από τη χρήση της προτεινόμενης μεθόδου

4.1.3.1 Εφαρμογή αλγορίθμου στα συνθετικά δεδομένα

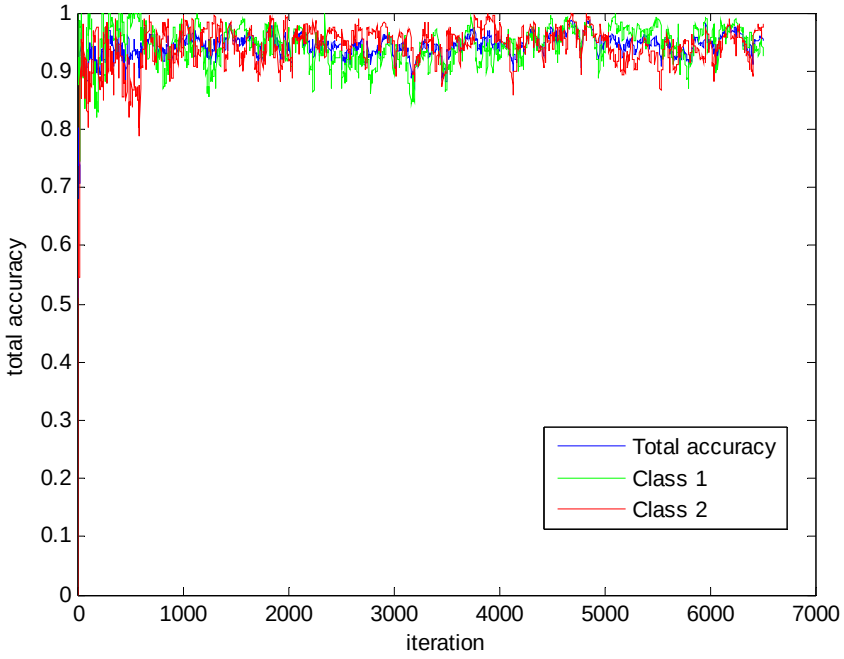
Όπως και στις δυο προηγούμενες μεθόδους, έτσι εργαστήκαμε και για την βασική μέθοδο, που μελετάμε. Αρχικά εφαρμόσαμε τον αλγόριθμο στα πιο απλά δεδομένα για να μπορέσουμε να δούμε την απόδοση του. Ως πιο απλά, αναφερόμαστε στα συνθετικά δεδομένα (α), (β) και (γ). Πριν εφαρμόσουμε τον αλγόριθμο, χωρίσαμε τα δεδομένα σε δεδομένα ελέγχου και δεδομένα εκπαίδευσης όπως περιγράψαμε στην παράγραφο 2.2. Στη συνέχεια εφαρμόσαμε τον αλγόριθμο, με σταθερές τιμές για τις μεταβλητές που περιγράψαμε στην ενότητα 3.1.1.

Ως έξοδο από τον αλγόριθμο, λαμβάνουμε την ακρίβεια των κλάσεων, καθώς και τον αριθμό των στοιχείων του συνόλου ελέγχου που ανήκουν στην πρώτη κλάση $N1$ και των στοιχείων του συνόλου ελέγχου που ανήκουν στη δεύτερη κλάση $N2$. Ακόμα ένα στοιχείο που λαμβάνουμε ως έξοδο, είναι ο χρόνος που χρειάστηκε για να ολοκληρωθεί η εκπαίδευση και η κατηγοριοποίηση όλων των στοιχείων. Στην συνέχεια υπολογίσαμε, όπως και στις άλλες μεθόδους, την συνολική ακρίβεια των κλάσεων με τον τύπο (11). Τα αποτελέσματα της εφαρμογής του αλγορίθμου στα απλά συνθετικά δεδομένα φαίνονται στους πίνακες που ακολουθούν.

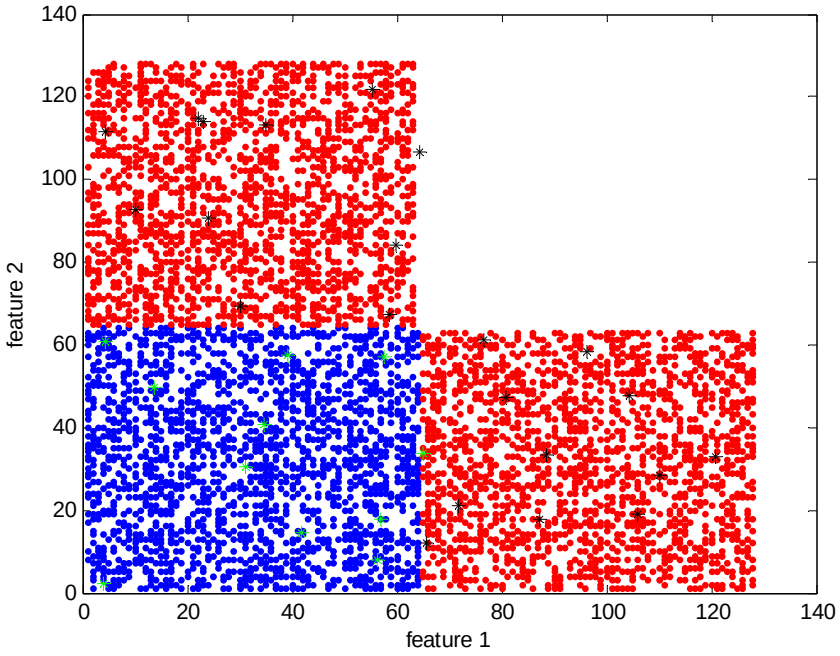
Εφαρμόσαμε στη συνέχεια, τον αλγόριθμο και στα συνθετικά δεδομένα (δ). Τα αποτελέσματα τους φαίνονται στους πίνακες που ακολουθούν.



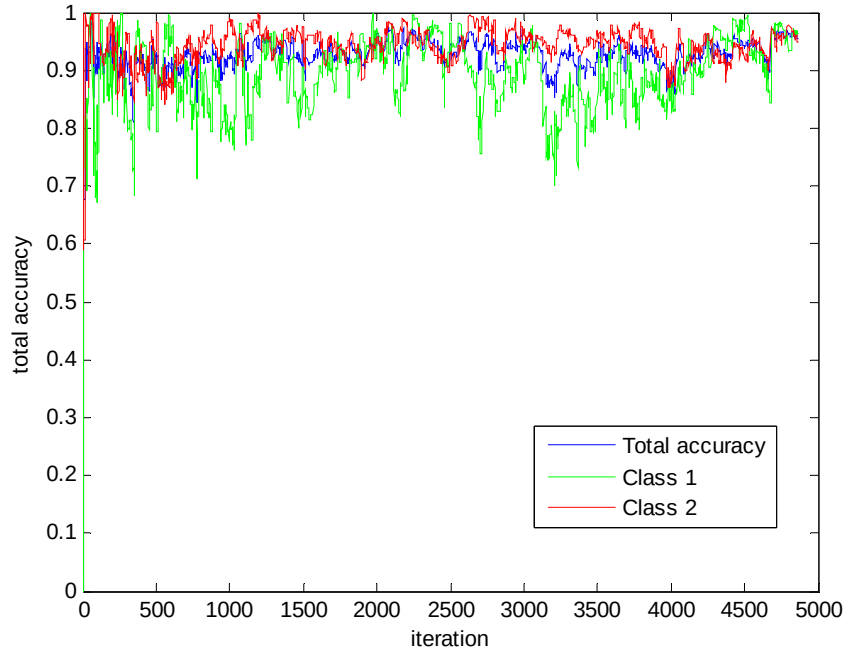
Εικόνα 11 Κατανομή όλων των στοιχείων των δύο κλάσεων (κόκκινο – μπλε) και των στοιχείων του MC συνόλου στο χώρο (μαύρο - πράσινο), για τα συνθετικά (α) δεδομένα, με χρήση του αλγορίθμου AIRS και τις καλύτερες τιμές των παραμέτρων του.



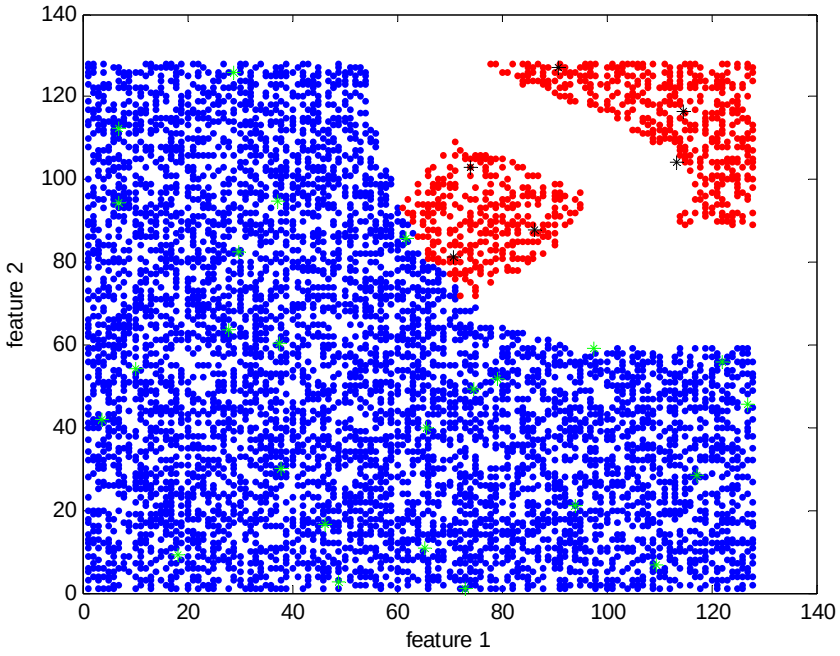
Εικόνα 12 Διάγραμμα ακρίβειας των κλάσεων και συνολική ακρίβεια για μια δεδομένη επανάληψη του αλγορίθμου AIRS και τις καλύτερες τιμές των παραμέτρων του, για τα συνθετικά (β) δεδομένα.



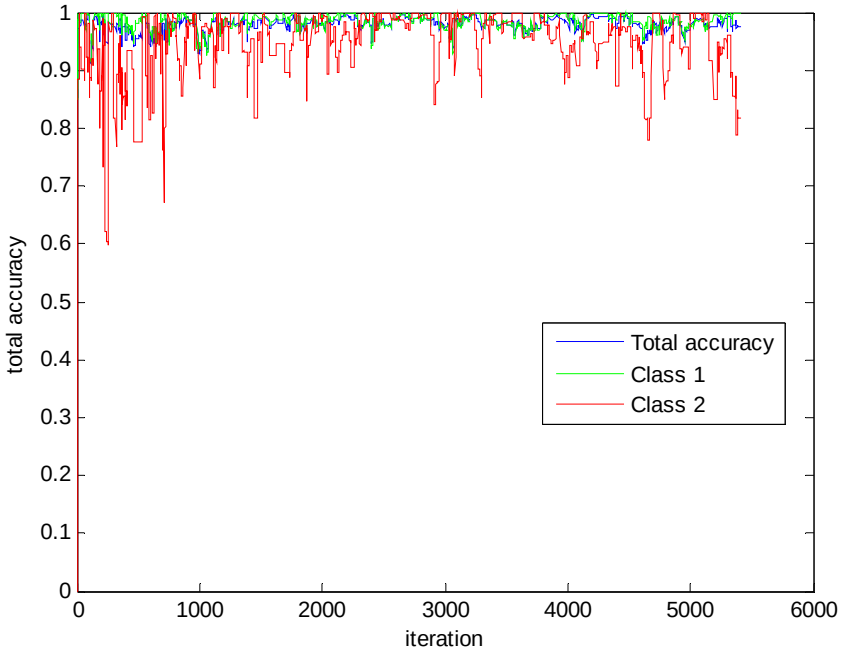
Εικόνα 13 Κατανομή όλων των στοιχείων των δύο κλάσεων (κόκκινο – μπλε) και των στοιχείων του MC συνόλου στο χώρο (μαύρο - πράσινο), για τα συνθετικά (β) δεδομένα, με χρήση του αλγορίθμου AIRS και τις καλύτερες τιμές των παραμέτρων του.



Εικόνα 14 Διάγραμμα ακρίβειας των κλάσεων και συνολική ακρίβεια για μια δεδομένη επανάληψη του αλγορίθμου AIRS και τις καλύτερες τιμές των παραμέτρων του, για τα συνθετικά (β) δεδομένα.



Εικόνα 15 Κατανομή όλων των στοιχείων των δύο κλάσεων(κόκκινο – μπλε) και των στοιχείων του MC συνόλου στο χώρο(μαύρο - πράσινο), για τα συνθετικά (γ) δεδομένα, με χρήση του αλγορίθμου AIRS και τις καλύτερες τιμές των παραμέτρων του.



Εικόνα 16 Διάγραμμα ακρίβειας των κλάσεων και συνολική ακρίβεια για μια δεδομένη επανάληψη του αλγορίθμου AIRS και τις καλύτερες τιμές των παραμέτρων του, για τα συνθετικά (γ) δεδομένα.

Πλήθος επαναλήψεων	Μέση ακρίβεια 1ης κλάσης	Μέση ακρίβεια 2ης κλάσης	Συνολική μέση ακρίβεια	Χρόνος εκτέλεσης
1η	0.965	0.961	0.963	1140
2η	0.896	0.987	0.941	1036
3η	0.947	0.976	0.961	932
4η	0.932	0.969	0.950	1309
5η	0.945	0.917	0.931	1189
6η	0.977	0.940	0.959	1145
7η	0.945	0.967	0.956	1051
8η	0.988	0.918	0.954	865
9η	0.943	0.983	0.963	1087
10η	0.953	0.952	0.952	1116
Μέση τιμή για τις 10 επαναλήψεις	0.949	0.957	0.953	1087

Πίνακας 17 Πίνακας ακρίβειας των κλάσεων για τα συνθετικά δεδομένα (α) με την χρήση AIRS για δέκα επαναλήψεις, για τις καλύτερες τιμές των παραμέτρων του αλγορίθμου και διαχωρισμό test-train→0.6-0.4.

Πλήθος επαναλήψεων	Μέση ακρίβεια 1ης κλάσης	Μέση ακρίβεια 2ης κλάσης	Συνολική μέση ακρίβεια	Χρόνος εκτέλεσης
1η	0.921	0.980	0.960	585
2η	0.942	0.935	0.937	592
3η	0.953	0.975	0.968	608
4η	0.896	0.929	0.918	655
5η	0.876	0.946	0.923	546
6η	0.920	0.958	0.945	639
7η	0.929	0.941	0.937	578
8η	0.912	0.938	0.929	614

9η	0.946	0.968	0.961	573
10η	0.881	0.997	0.959	584
Μέση τιμή για τις 10 επαναλήψεις	0.918	0.957	0.944	597

Πίνακας 18 Πίνακας ακρίβειας των κλάσεων για τα συνθετικά δεδομένα (β) με την χρήση AIRS για δέκα επαναλήψεις, για τις καλύτερες τιμές των παραμέτρων του αλγορίθμου και διαχωρισμό test-train→0.6-0.4.

Πλήθος επαναλήψεων	**Μέση ακρίβεια 1ης κλάσης**	**Μέση ακρίβεια 2ης κλάσης**	**Συνολική μέση ακρίβεια**	**Χρόνος εκτέλεσης**
1η	0.989	0.892	0.976	1172
2η	0.982	1.000	0.984	862
3η	0.988	0.961	0.984	853
4η	0.995	0.876	0.979	895
5η	0.991	0.986	0.990	876
6η	0.968	0.996	0.972	1004
7η	1.000	0.915	0.989	873
8η	0.989	1.000	0.986	1060
9η	1.000	0.897	0.986	1196
10η	0.996	0.905	0.985	902
Μέση τιμή για τις 10 επαναλήψεις	0.990	0.943	0.983	969

Πίνακας 19 Πίνακας ακρίβειας των κλάσεων για τα συνθετικά δεδομένα (γ) με την χρήση AIRS για δέκα επαναλήψεις, για τις καλύτερες τιμές των παραμέτρων του αλγορίθμου και διαχωρισμό test-train→0.6-0.4.

Τα τρία πρώτα σύνολα των συνθετικών δεδομένων, τα θεωρήσαμε πιο εύκολα για ταξινόμηση και εστιάσαμε στο τέταρτο και τελευταίο σύνολο, το οποίο είναι πιο περίπλοκο καθώς ο διαχωρισμός των κλάσεων δεν είναι γραμμικός, αλλά γίνεται από μια ημιτονοειδή καμπύλη με αυξανόμενη χωρική συχνότητα. Τα αποτελέσματα για τις δοκιμές στα δεδομένα αυτά φαίνονται στους παρακάτω πίνακες και διαγράμματα.

Πλήθος επαναλήψεων	Μέση ακρίβεια 1ης κλάσης	Μέση ακρίβεια 2ης κλάσης	Συνολική μέση ακρίβεια	Χρόνος εκτέλεσης
1η	0.876	0.882	0.879	237
2η	0.850	0.880	0.865	221
3η	0.847	0.888	0.868	230
4η	0.864	0.836	0.850	257
5η	0.845	0.894	0.870	237
6η	0.857	0.881	0.869	239
7η	0.865	0.876	0.871	222
8η	0.859	0.879	0.869	245
9η	0.869	0.855	0.862	228
10η	0.884	0.848	0.865	223
Μέση τιμή για τις 10 επαναλήψεις	0.8616	0.8719	0.8668	233

Πίνακας 20 Πίνακας ακρίβειας των κλάσεων για τα συνθετικά δεδομένα (δ) με την χρήση AIRS για δέκα επαναλήψεις, για τις καλύτερες τιμές των παραμέτρων του αλγορίθμου και διαχωρισμό test-train→0.6-0.4.

Στον παραπάνω πίνακα, φαίνονται οι τιμές της μέσης ακρίβειας των δύο κλάσεων καθώς και η συνολική μέση ακρίβεια των κλάσεων για δέκα ανεξάρτητες επαναλήψεις του αλγορίθμου. Ο πίνακας προκύπτει από την εφαρμογή του αλγορίθμου AIRS με σταθερές αρχικές τιμές για τις παραμέτρους της παραγράφου 3.1.1 και για ποσοστό test-train ίσο με 0,6-0,4. Στον πίνακα αυτόν, παρατηρούμε επίσης ότι η απόδοση του αλγορίθμου δεν βρίσκεται στα ίδια επίπεδα με την απόδοση για τα προηγούμενα συνθετικά δεδομένα. Αυτό ήταν αναμενόμενο καθώς ο βαθμός δυσκολίας διάκρισης των δυο κλάσεων έχει αυξηθεί.

Επόμενος πίνακας είναι ο πίνακας 21, στον οποίο παρουσιάζεται το πλήθος των στοιχείων της πρώτης και δεύτερης κλάσης αλλά και το πλήθος των στοιχείων των κυττάρων μνήμης που δημιουργήθηκε από την εφαρμογή του αλγορίθμου.

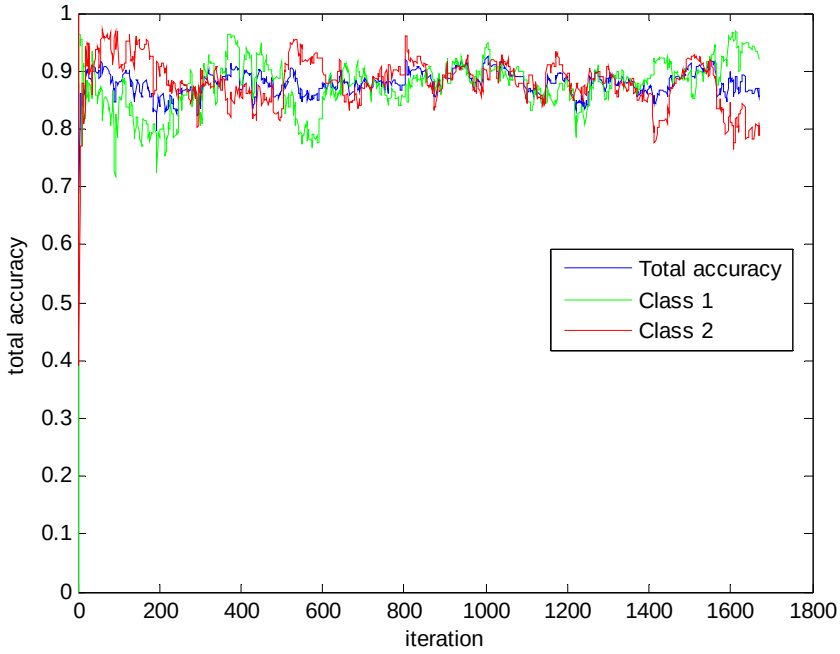
Επαναλήψεις	Πλήθος πρώτης κλάσης	Πλήθος δεύτερης κλάσης	Πλήθος MC συνόλου
1η	7191	7522	29
2η	7159	7557	27
3η	7200	7551	34
4η	7151	7522	31
5η	7205	7547	26
6η	7205	7563	27
7η	7199	7530	26
8η	7154	7609	29
9η	7141	7506	31
10η	7184	7520	26

Πίνακας 21 Πλήθος στοιχείων ελέγχου που ανήκουν στην πρώτη και την δεύτερη κλάση και πλήθος του τελικού MC συνόλου κατά τη διάρκεια των δέκα επαναλήψεων, για την εφαρμογή του AIRS στα συνθετικά (δ) δεδομένα.

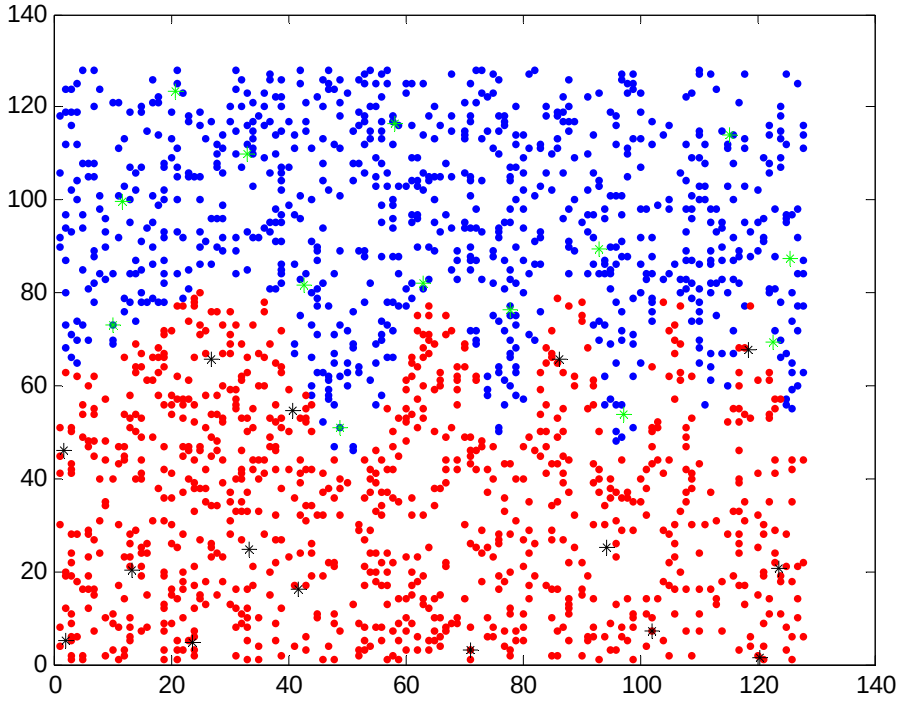
Όπως αναφέραμε οι πρώτες δοκιμές για την απόδοση του αλγορίθμου έγιναν με σταθερές αρχικές τιμές και ποσοστό test-train→0,6-0,4. Στον παρακάτω πίνακα φαίνεται η ακρίβεια των κλάσεων και για τα υπόλοιπα ποσοστά διαχωρισμού των συνόλων εκπαίδευσης και ελέγχου. Όπως συμπεραίνουμε, όσο το ποσοστό του συνόλου εκπαίδευσης αυξάνεται, ο αλγόριθμος έχει καλύτερη ακρίβεια κατηγοριοποίησης, αλλά ο χρόνος που απαιτείται για να ολοκληρωθεί η διαδικασία εκπαίδευσης και ταξινόμησης αυξάνεται. Ως καλύτερη τιμή για το ποσοστό διαχωρισμού, θεωρούμε το ποσοστό 0,6-0,4 το οποίο έχει την καλύτερη απόδοση σε σχέση με τα υπόλοιπα ποσοστά. Ο χρόνος για το ποσοστό αυτό δεν είναι ικανοποιητικός.

Μέγεθος train / test set	Μέση Ακρίβεια 1ης κλάσης	Μέση Ακρίβεια 2ης κλάσης	Συνολική μέση ακρίβεια +- τυπική απόκλιση		Χρόνος εκτέλεσης (sec)
0,1/0,9	0.866	0.863	0.864	0.023	234
0,2/0,8	0.832	0.904	0.869	0.024	432
0,3/0,7	0.862	0.867	0.865	0.028	627
0,4/0,6	***0.877***	***0.888***	***0.882***	***0.021***	***905***
0,5/0,5	0.879	0.883	0.881	0.013	1037

Πίνακας 22 Μέση τιμή ακρίβειας των κλάσεων για τις διαφορετικές τιμές των ποσοστών test και train, για τα συνθετικά (δ) δεδομένα, για τις καλύτερες τιμές των παραμέτρων του αλγορίθμου AIRS.



Εικόνα 17 Διάγραμμα ακρίβειας πρώτης και δεύτερης κλάσης και συνολική ακρίβεια για τα συνθετικά δεδομένα με χρήση του αλγορίθμου AIRS.



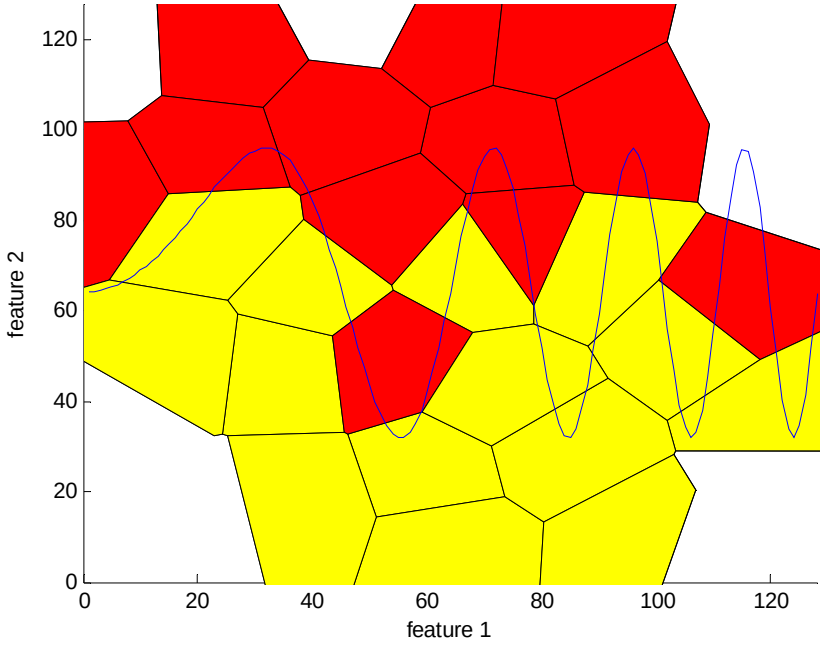
Εικόνα 18 Κατανομή όλων των στοιχείων των δύο κλάσεων (κόκκινο – μπλε) και των στοιχείων του MC συνόλου στο χώρο (μαύρο - πράσινο), για τα συνθετικά δεδομένα (δ), με χρήση του αλγορίθμου AIRS και τις βέλτιστες τιμές των παραμέτρων του.

Στην επόμενη εικόνα παρουσιάζεται η διαμέριση του χώρου των χαρακτηριστικών που επιτυγχάνεται από τον εκπαιδευμένο αλγόριθμο AIRS για τα συνθετικά δεδομένα (δ). Πιο συγκεκριμένα, χρησιμοποιούμε τις συντεταγμένες των mcs ώστε να παραχθεί το διάγραμμα Voronoi. Το διάγραμμα Voronoi ενός συνόλου σημείων διαμερίζει τον χώρο δημιουργώντας ένα πολύγωνο γύρω από κάθε δεδομένο σημείο, έτσι ώστε το δεδομένο σημείο να είναι το κοντινότερο από κάθε σημείο εντός του πολυγώνου.

Τα διαγράμματα Voronoi είναι γεωμετρικές κατασκευές που παρουσιάζουν ιδιαίτερο ενδιαφέρον καθώς διαθέτουν ένα ευρύ φάσμα εφαρμογών στην αναγνώριση προτύπων. Στην συγκεκριμένη περίπτωση, μπορούμε να θεωρήσουμε ότι η κατηγοριοποίηση είναι ισοδύναμη με την διαμέριση του χώρου των διανυσμάτων χαρακτηριστικών (εν προκειμένω το $\square^2$ αφού πρόκειται για τα συνθετικά δεδομένα - δ) σε τόσες περιοχές όσες και το πλήθος των κλάσεων (εν προκειμένω 2). Αν κατασκευάσουμε το διάγραμμα Voronoi που προκύπτει από τις θέσεις των κυττάρων μνήμης, τότε διαμερίζουμε τον χώρο των διανυσμάτων χαρακτηριστικών σε τόσα χωρία όσα και τα κύτταρα μνήμης. Κάθε χωρίο είναι ο γεωμετρικός τόπος των σημείων του χώρου των χαρακτηριστικών που έχουν ελάχιστη ευκλείδεια απόσταση από το κύτταρο μνήμης του χωρίου τους, σε σχέση με οποιοδήποτε άλλο κύτταρο μνήμης. Έτσι, για κάθε σημείο του χώρου των χαρακτηριστικών, μπορούμε να

οπτικοποιήσουμε την έξοδο του αλγορίθμου AIRS, ανάλογα με την κλάση του κυττάρου μνήμης στο χωρίο του οποίου βρίσκεται το εν λόγω σημείο [36][37][38]

Επί του διαγράμματος Voronoi υπερτίθεται η καμπύλη διαχωρισμού των δύο κλάσεων (Εικόνα 19). Ανάλογα με την κλάση του κάθε κυττάρου μνήμης κατηγοριοποιούμε τα σημεία του χώρου των χαρακτηριστικών (κόκκινα κλάση 1 και κίτρινα κλάση 2). Ιδανικά θα έπρεπε οι κόκκινες περιοχές να βρίσκονται κάτω από την καμπύλη διαχωρισμού και οι κίτρινες πάνω από αυτήν. Παρατηρούμε ότι υπάρχει γενική συμφωνία μεταξύ της διαμέρισης που επιτυγχάνεται με το διάγραμμα Voronoi και της αναλυτικής καμπύλης, αν και συγκεκριμένα σημεία του χώρου κατηγοριοποιούνται εσφαλμένα από τα κύτταρα μνήμης που παρήγαγε ο αλγόριθμος AIRS.



Εικόνα 19 Διαγραμματική απεικόνιση της διαμέρισης του χώρου των χαρακτηριστικών που επιτυγχάνεται από τον εκπαιδευμένο αλγόριθμο AIRS για τα συνθετικά δεδομένα (δ).

4.1.3.2 Εφαρμογή αλγορίθμου στα δεδομένα από το φυτό Iris

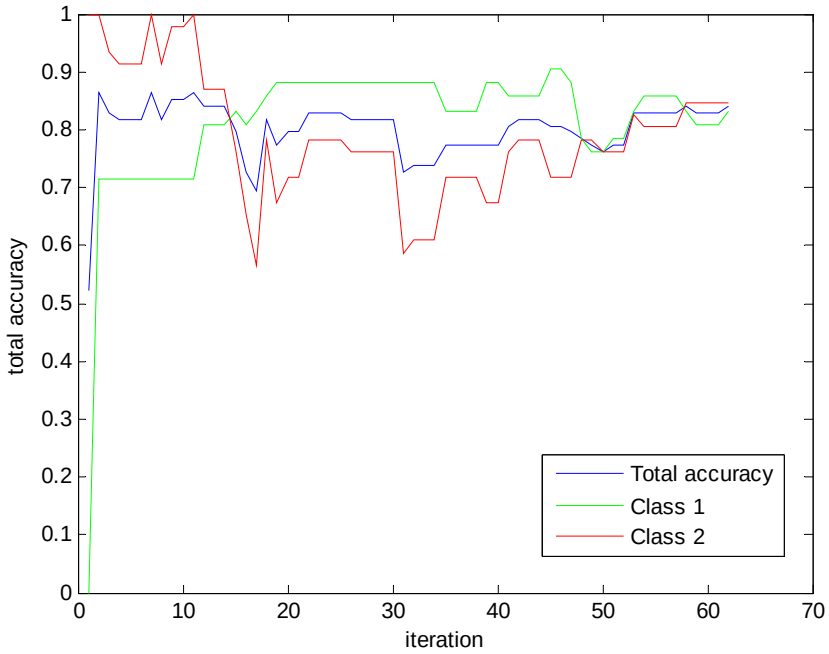
Όπως εργαστήκαμε με τα δεδομένα αυτά, στις προηγούμενες μεθόδους, έτσι εργαστήκαμε και στην μέθοδο που μελετάμε. Αφού πρώτα κάναμε τις απαραίτητες μετατροπές για να κανονικοποιήσουμε τις τιμές των δεδομένων αυτών, τα χωρίσαμε σε δυο σύνολα. Το πρώτο είναι το σύνολο ελέγχου και το δεύτερο είναι το σύνολο εκπαίδευσης. Στον πίνακα που ακολουθεί παρουσιάζονται οι τιμές της μέσης

ακρίβειας των κλάσεων για δέκα διαδοχικές επαναλήψεις. Η εφαρμογή του αλγόριθμου στα δεδομένα αυτά έγινε με τις βέλτιστες τιμές των παραμέτρων, όπως αυτές καθορίστηκαν στην προηγούμενη παράγραφο βάσει των συνθετικών δεδομένων. Επίσης ο διαχωρισμός σε ποσοστό test-train έγινε με βάση τα αποτελέσματα που πήραμε για τα συνθετικά δεδομένα, από τα οποία θεωρήσαμε ως καλύτερο ποσοστό το ποσοστό 0,6-0,4 αντίστοιχα.

Από τον πίνακα αυτόν, μπορούμε να δούμε ότι οι τιμές της ακρίβειας των κλάσεων είναι σε ένα ικανοποιητικό επίπεδο, πιο κάτω όμως από τις τιμές που έδωσαν τα συνθετικά δεδομένα. Το πλήθος των δεδομένων αυτών είναι αρκετά μικρότερο από το πλήθος των στοιχείων των συνθετικών. Για τον λόγο αυτό, η καμπύλες των τιμών ακρίβειας που φαίνονται στην Εικόνα 20 είναι περισσότερο διακριτές. Επιπλέον τόσο από τον πίνακα 23 όσο και από την εικόνα 20, βλέπουμε ότι οι τιμές τις ακρίβειας των κλάσεων έχουν μεγαλύτερη τυπική απόκλιση σε σύγκριση με τα προηγούμενα δεδομένα.

Πλήθος επαναλήψεων	**Μέση ακρίβεια 1ης κλάσης**	**Μέση ακρίβεια 2ης κλάσης**	**Συνολική μέση ακρίβεια**
1η	0.947	0.619	0.775
2η	0.778	0.904	0.842
3η	0.750	0.844	0.807
4η	0.850	0.717	0.779
5η	0.732	0.867	0.802
6η	0.860	0.794	0.826
7η	0.744	0.809	0.778
8η	0.795	0.936	0.872
9η	0.833	0.848	0.841
10η	0.833	0.800	0.818
Μέση τιμή για τις 10 επαναλήψεις	0.812	0.814	0.814

Πίνακας 23 Πίνακας ακρίβειας των κλάσεων για τα δεδομένα Iris με την χρήση AIRS για δέκα επαναλήψεις, με τις καλύτερες τιμές για τις παραμέτρους και διαχωρισμό test-train→0.6-0.4.



Εικόνα 20 Διάγραμμα ακρίβειας πρώτης και δεύτερης κλάσης και συνολική ακρίβεια (πράσσινη, κόκκινη και μπλε καμπύλη αντίστοιχα) για τα Iris δεδομένα με χρήση του αλγορίθμου AIRS για μια από τις 10 επαναλήψεις.

4.1.3.3 Εφαρμογή αλγορίθμου στα κλινικά δεδομένα

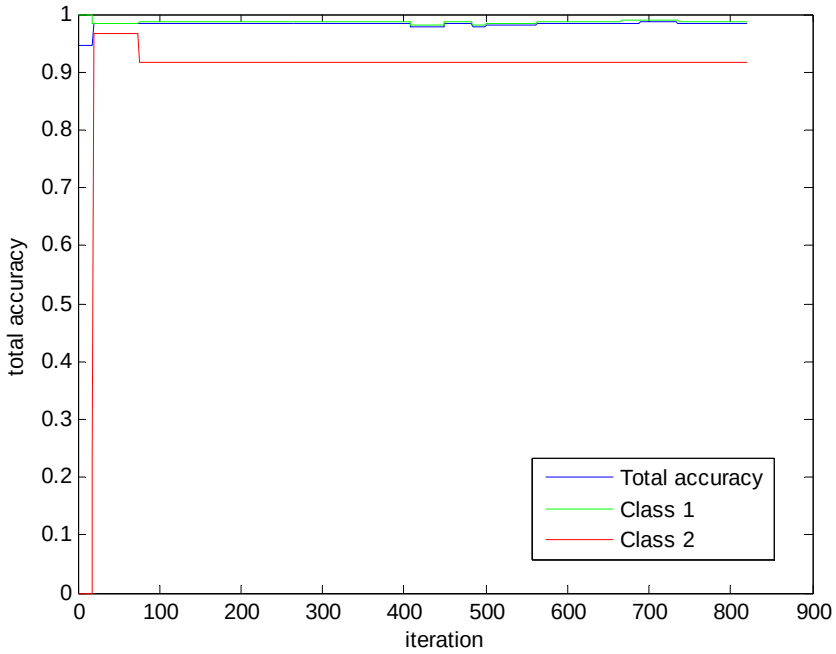
Τελευταία και σημαντικότερα είναι τα δεδομένα που προέκυψαν από βιοψίες FNA θυρεοειδούς. Αρχικά θα αναφερθούμε στην πρώτη ομάδα των δεδομένων αυτών, στα non-suspicious (μη ύποπτα δεδομένα), καθώς όπως αναφέραμε και παραπάνω είναι πιο εύκολο να ταξινομηθούν. Εφαρμόσαμε και σε αυτά τα δεδομένα, τον αλγόριθμο AIRS με τις καλύτερες τιμές των παραμέτρων και το καλύτερο ποσοστό διαχωρισμού test-train.

Στον παρακάτω πίνακα φαίνονται οι τιμές τις ακρίβειας των κλάσεων καθώς και η συνολική ακρίβεια. Η συνολική ακρίβεια υπολογίστηκε και πάλι με τον τύπο (11). Όπως μπορούμε να δούμε η απόδοση του αλγορίθμου για τα δεδομένα αυτά είναι σε αρκετά υψηλό επίπεδο, το οποίο φτάνει πάνω από το 90% για τις περισσότερες επαναλήψεις.

Πλήθος επαναλήψεων	Μέση ακρίβεια 1ης κλάσης	Μέση ακρίβεια 2ης κλάσης	Συνολική μέση ακρίβεια
1η	0.987	0.918	0.983
2η	0.998	0.855	0.991
3η	0.998	0.901	0.992
4η	0.992	0.917	0.987
5η	1.000	0.836	0.991
6η	0.993	0.894	0.987
7η	0.998	0.848	0.990
8η	0.999	0.848	0.991
9η	0.996	0.881	0.990
10η	0.999	0.895	0.992
Μέση τιμή για τις 10 επαναλήψεις	0.996	0.879	0.989

Πίνακας 24 Πίνακας ακρίβειας των κλάσεων για τα non-suspicious δεδομένα με την χρήση AIRS για δέκα επαναλήψεις, με τις καλύτερες τιμές των παραμέτρων και διαχωρισμό test-train→0.6-0.4.

Στην παρακάτω εικόνα φαίνονται διαγραμματικά οι τιμές της ακρίβειας και καθίσταται σαφές ότι στην κάθε επανάληψη η τυπική απόκλιση στις τιμές είναι πολύ μικρή. Όπως βλέπουμε, συγκεκριμένα για την δεύτερη κλάση (Class 2) η τιμή της ακρίβειας φαίνεται να είναι σταθερή μετά από ένα σημείο. Αυτό μας δηλώνει ότι ο διαχωρισμός των δεδομένων στις δυο ομάδες, καλοήθη και κακοήθη κύτταρα, έγινε με επιτυχία.



Εικόνα 21 Διάγραμμα ακρίβειας πρώτης και δεύτερης κλάσης και συνολική ακρίβεια για τα non-suspicious δεδομένα με χρήση του αλγορίθμου AIRS κατά την 3η επανάληψη του Πίνακα 24.

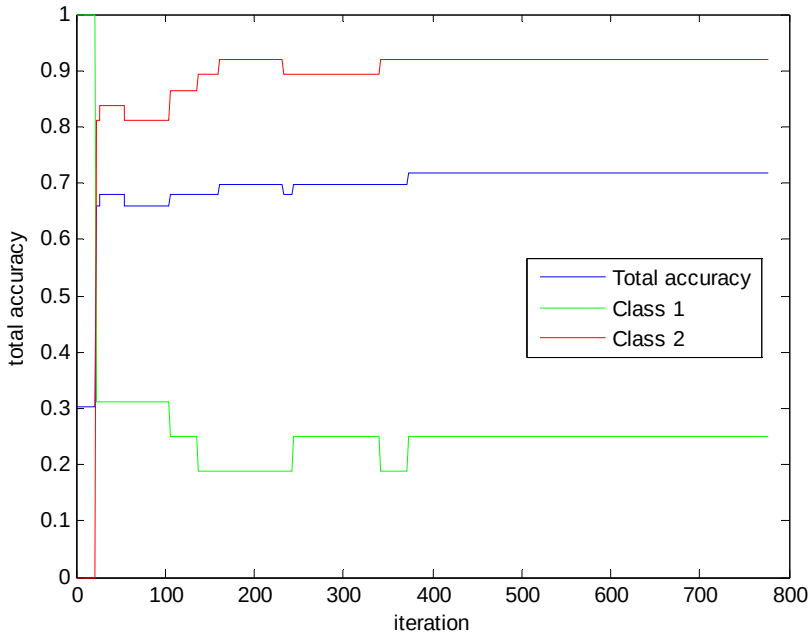
Στη συνέχεια εφαρμόσαμε και πάλι τον αλγόριθμο στα non-suspicious δεδομένα, με τη διαφορά ότι αυτή την φορά ορίσαμε ως δεδομένα ελέγχου τα suspicious δεδομένα. Θέλουμε με αυτόν τον τρόπο να διερευνήσουμε την κλινική χρησιμότητα του εν λόγω αλγορίθμου κατηγοριοποίησης αναφορικά με τα ύποπτα δεδομένα, τα οποία όπως αναφέρθηκε και προηγουμένως, η βιοψία FNA δεν μπορεί να κατηγοριοποιήσει και τα θεωρεί ως κακοήθη (κλάση 2).

Τα αποτελέσματα που παίρνουμε από αυτόν τον έλεγχο, μας δείχνουν τιμές για την ακρίβεια της πρώτης κλάσης που δεν ξεπερνούν 31% κατά τη διάρκεια των δέκα επαναλήψεων, ενώ οι τιμές για την δεύτερη κλάση κυμαίνονται από το 78% έως και το 91%. Η διαφορά αυτή ανάμεσα στις τιμές ακρίβειας των κλάσεων δίνουν μια χαμηλή τιμή στην συνολική ακρίβεια η οποία κυμαίνεται μεταξύ 71% και 60%.

Πλήθος επαναλήψεων	**Μέση ακρίβεια 1ης κλάσης**	**Μέση ακρίβεια 2ης κλάσης**	**Συνολική μέση ακρίβεια**
1η	0.250	0.811	0.642
2η	0.188	0.919	0.698
3η	0.188	0.838	0.642
4η	0.250	0.919	0.717

5η	0.250	0.892	0.698
6η	0.313	0.838	0.679
7η	0.250	0.784	0.623
8η	0.250	0.838	0.660
9η	0.250	0.892	0.698
10η	0.188	0.919	0.698
Μέση τιμή για τις 10 επαναλήψεις	0.238	0.865	0.675

Πίνακας 25 Πίνακας ακρίβειας των κλάσεων για τα non-suspicious δεδομένα και δεδομένα ελέγχου τα suspicious, με την χρήση AIRS για δέκα επαναλήψεις.



Εικόνα 22 Διάγραμμα ακρίβειας πρώτης και δεύτερης κλάσης και συνολική ακρίβεια όπως υπολογίζεται χρησιμοποιώντας τμήμα των non-suspicious δεδομένων ως υποσύνολο εκπαίδευσης και ως δεδομένα ελέγχου τα suspicious δεδομένα, με χρήση του αλγορίθμου AIRS κατά την 4η επανάληψη του πίνακα 25.

4.1.4 Σύγκριση μεθόδων

Στην ενότητα αυτή ελέγχουμε τα αποτελέσματα από όλες τις μεθόδους και τα συγκρίνουμε ώστε να μπορέσουμε να καταλήξουμε σε κάποια γενικά συμπεράσματα σχετικά με την απόδοση της προτεινόμενης μεθόδου. Όπως έχουμε ήδη αναφέρει, η μέθοδος των νευρωνικών δικτύων καθώς και η μέθοδος των k πλησιέστερων

γειτόνων, είναι περισσότερο διαδεδομένες και χρησιμοποιούνται συχνά για την επίλυση προβλημάτων κατηγοριοποίησης. Έτσι λοιπόν, μετά από την σύγκριση της μεθόδου με αυτές τις δύο, η απόδοση της καθώς και τα αποτελέσματα της, μπορούν να θεωρηθούν αξιόπιστα. Στον πίνακα 26 που ακολουθεί καταγράφονται οι τιμές των παραμέτρων από τις οποίες εξαρτάται η κάθε μέθοδος. Σύμφωνα με τις ρυθμίσεις αυτές έχουμε τα αποτελέσματα που βλέπουμε στον πίνακα 27 για την εφαρμογή της κάθε μεθόδου στα non-suspicious δεδομένα. Και στον πίνακα 28 καταγράφονται οι τιμές ακρίβειας των κλάσεων, κατά μέσο όρο για δέκα επαναλήψεις στις μεθόδους, για κάθε ποσοστό διαχωρισμού των δεδομένων σε σύνολο ελέγχου και σύνολο εκπαίδευσης.

Παράμετρος αλγορίθμου	**AIRS**	**Neural Networks**	**kNN**
max_resources	600	-	-
hyper_clonal_rate	15	-	-
clonal_rate	20	-	-
mutation_rate	0,5	-	-
stimulation_threshold	0,5	-	-
Ποσοστό test-train	0,6-0,4	0,6-0,4	0,8-0,2
Πλήθος νευρώνων	-	7	-
Πλήθος γειτόνων	-	-	3

Πίνακας 26 Πίνακας τιμών των παραμέτρων για την κάθε μέθοδο ταξινόμησης.

Μέθοδος ταξινόμησης	Μέση τιμή ακρίβειας πρώτης κλάσης	Μέση τιμή ακρίβειας δεύτερης κλάσης	Μέση τιμή συνολικής ακρίβειας
AIRS	0.996	0.879	0.989
Neural Networks	0.991	0.863	0.983
kNN	0.863	0.916	0.867

Πίνακας 27 Μέση τιμή της ακρίβειας της κάθε κλάσης και της συνολικής για την κάθε μέθοδο με τις καλύτερες τιμές των παραμέτρων κάθε μιας για τα non-suspicious δεδομένα.

	Μέση τιμή ακρίβειας πρώτης κλάσης			Μέση τιμή ακρίβειας δεύτερης κλάσης			Μέση τιμή συνολικής ακρίβειας		
Train Test	**AIRS**	**Neural Net**	**kNN**	**AIRS**	**Neural Net**	**kNN**	**AIRS**	**Neural Net**	**kNN**
10 - 90	0.998	0.992	0.830	0.876	0.816	0.876	0.990	0.982	0.868
20 – 80	0.999	0.988	0.863	0.847	0.825	0.916	0.990	0.979	0.867
30 - 70	0.999	0.990	0.479	0.838	0.844	0.948	0.989	0.982	0.505
40 - 60	0.996	0.991	0.192	0.879	0.963	0.977	0.989	0.983	0.235

Πίνακας 28 Πίνακας μέσων τιμών ακρίβειας των κλάσεων και συνολική ακρίβεια για τις τρείς μεθόδους, για τις καλύτερες τιμές των παραμέτρων τους, για διαφορετικά ποσοστά test-train διαχωρισμού, για τα non-suspicious δεδομένα.

Από τους δυο παραπάνω πίνακες, διαπιστώνουμε ότι η μέθοδος που προτείνουμε για την κατηγοριοποίηση των non-suspicious δεδομένων, έχει τα καλύτερα αποτελέσματα σε σύγκριση με τις άλλες δυο μεθόδους με το ποσοστό επιτυχούς ταξινόμησης να φτάνει στο 99% κατά μέσο όρο, για την συνολική ακρίβεια των κλάσεων. Αξίζει στο σημείο αυτό να γίνει η σύγκριση με την ακρίβεια που επιτυγχάνει ο ειδικός με χρήση της μεθόδου FNA. Χρησιμοποιώντας τον Πίνακα 1 και αφαιρώντας τα ύποπτα δεδομένα, εύκολα διαπιστώνουμε ότι ο ειδικός επιτυγχάνει για την κλάση Α (καλοήθη) ακρίβεια ίση με 1848/1850=99.89% και για την κλάση Β (κακοήθη) ακρίβεια ίση με 97/113=85,84%. Η συνολική ακρίβεια του ειδικού (χωρίς τα ύποπτα δεδομένα) ανέρχεται σε 99,08%. Παρατηρούμε ότι και οι τρεις υπό σύγκριση μέθοδοι επιτυγχάνουν ακρίβεια για κάθε μία κλάση, αλλά και συνολικά ίση ή και οριακά καλύτερη από αυτή του ειδικού.

Ακολουθούν δυο πίνακες που περιγράφουν την απόδοση των μεθόδων χρησιμοποιώντας αυτή τη φορά τα suspicious δεδομένα, ως δεδομένα ελέγχου. Ο διαχωρισμός γίνεται ως εξής:

Διαχωρίζουμε τα non-suspicious δεδομένα με τον τρόπο που έχει περιγραφεί σε παραπάνω παραγράφους και στην πορεία ορίζω ως σύνολο ελέγχου τα suspicious δεδομένα που έχουμε λάβει από την βιοψία λεπτής βελόνας FNA. Αυτό που συμπεραίνουμε από τους παρακάτω πίνακες είναι ότι και οι τρείς μέθοδοι κατηγοριοποίησης που μελετάμε δίνουν παρόμοια αποτελέσματα, κατηγοριοποιώντας σωστά την δεύτερη κλάση που αναφέρεται στα κακοήθη δείγματα. Πιό συγκεκριμένα, η μέθοδος ΤΝΔ δίνει ποσοστό ακρίβειας της τάξης του 23,8% για την πρώτη κλάση κατά μέσο όρο, 76,5% για την δεύτερη κλάση κατά μέσο όρο και 61,9% για την συνολική μέση ακρίβεια. Ο αλγόριθμος kNN δίνει ποσοστό ακρίβειας της τάξης του18,8% για την πρώτη κλάση κατά μέσο όρο, 93,2% για την δεύτερη κλάση κατά μέσο όρο και 70,8% για την συνολική μέση ακρίβεια.

Μέθοδος ταξινόμησης	**Μέση τιμή ακρίβειας πρώτης κλάσης**	**Μέση τιμή ακρίβειας δεύτερης κλάσης**	**Μέση τιμή συνολικής ακρίβειας**
AIRS	0.238	0.865	0.675
Neural Networks	0.281	0.765	0.619
kNN	0.188	0.932	0.708

Πίνακας 29 Μέση τιμή της ακρίβειας της κάθε κλάσης και της συνολικής για την κάθε μέθοδο με τις καλύτερες τιμές των παραμέτρων με σύνολο εκπαίδευσης τμήμα των non-suspicious δεδομένων και σύνολο ελέγχου τα suspicious δεδομένα.

Πλήθος επαναλήψεων	**Μέση ακρίβεια 1ης κλάσης**	**Μέση ακρίβεια 2ης κλάσης**	**Συνολική μέση ακρίβεια**
1η	0.188	0.892	0.679
2η	0.250	0.838	0.660
3η	0.250	0.784	0.623
4η	0.250	0.838	0.660
7η	0.125	0.784	0.585
8η	0.438	0.973	0.774

9η	0.438	0.811	0.698
10η	0.250	0.676	0.547
Μέση τιμή για τις 10 επαναλήψεις	0.281	0.765	0.619

Πίνακας 30 Πίνακας ακρίβειας των κλάσεων για τα non-suspicious δεδομένα και δεδομένα ελέγχου τα suspicious, με την χρήση ΤΝΔ για δέκα επαναλήψεις, με εφαρμογή των καλύτερων τιμών των παραμέτρων.

Πλήθος επαναλήψεων	**Μέση ακρίβεια 1ης κλάσης**	**Μέση ακρίβεια 2ης κλάσης**	**Συνολική μέση ακρίβεια**
1η	0	1	0.698
2η	0.375	0.865	0.717
3η	0	1	0.698
4η	0	1	0.698
5η	0.375	0.865	0.717
6η	0.375	0.865	0.717
7η	0.375	0.865	0.717
8η	0	1	0.698
9η	0	1	0.698
10η	0.375	0.865	0.717
Μέση τιμή για τις 10 επαναλήψεις	0.188	0.932	0.708

Πίνακας 31 Πίνακας ακρίβειας των κλάσεων με σύνολο εκπαίδευσης τμήμα των non-suspicious δεδομένων και σύνολο ελέγχου τα suspicious δεδομένα, με την χρήση kNN για δέκα επαναλήψεις, με χρήση των καλύτερων τιμών των παραμέτρων.

Διαπιστώνουμε ότι οι τρεις μέθοδοι συμπεριφέρονται με παρόμοιο τρόπο, επιτυγχάνοντας υψηλή ακρίβεια για τα ύποπτα δεδομένα που είναι πράγματι κακοήθη (κλάση Β) βάσει της μετέπειτα ιστολογικής εξέτασης και χαμηλή ακρίβεια για τα ύποπτα δεδομένα που αποδείχτηκαν καλοήθη. Αυτό σε σχέση με την επίδοση του ειδικού με χρήση της FNA αποτελεί μία βελτίωση των αποτελεσμάτων (αφού ο

ειδικός επιτυγχάνει 0% ακρίβεια για την κλάση Α και 100% ακρίβεια για την κλάση Β). Παρόλα αυτά η κλινική χρησιμότητα των μεθόδων κατηγοριοποίησης για τα ύποπτα δεδομένα είναι περιορισμένη, καθώς η ακρίβεια της κλάσης Α, παρότι καλύτερη αυτής του ειδικού εξακολουθεί να είναι απαγορευτικά χαμηλή.

4.2 Συμπεράσματα

Από το σύνολο των δοκιμών στο πλαίσιο της παρούσας πτυχιακής, μπορούμε να συνοψίσουμε τα τελικά συμπεράσματα, τα οποία έχουν ήδη αναλυθεί διεξοδικά στο κείμενο, στις παραγράφους σύγκρισης μεθόδων.

Ο αλγόριθμος AIRS αποτελεί μία ιδιαίτερη προσέγγιση του προβλήματος της εποπτευόμενης κατηγοριοποίησης δεδομένων. Στα περισσότερα πειράματα μας απεδείχθη το ίδιο ή και περισσότερο αποτελεσματικός με τις άλλες δύο υπό σύγκριση μεθόδους, αναφορικά με την ακρίβεια κατηγοριοποίησης που πέτυχε.

Δεδομένης της πολυπλοκότητας του ανοσοποιητικού συστήματος και της μεγάλης ικανότητας του να αναγνωρίζει τεράστιο πλήθος αντιγόνων, πολλές τάξεις μεγέθους μεγαλύτερο από το πλήθος των κυττάρων μνήμης του, θεωρούμε ότι υπάρχει χώρος για έρευνα στην βελτίωση της μεθόδου αυτής.

Αναφορικά με το κλινικό πρόβλημα της ορθής διάγνωσης του καρκίνου του θυρεοειδούς βάσει των αποτελεσμάτων της FNA, είναι σαφές ότι αναφορικά με τα μη ύποπτα δεδομένα, οι μέθοδοι κατηγοριοποίησης θα μπορούσαν να προσφέρουν στην ακρίβεια της κλάσης B (κακοήθη). Στην περίπτωση των ύποπτων δεδομένων, επιτυγχάνεται κάποια βελτίωση στην ανίχνευση των καλοηθών περιπτώσεων (που χαρακτηρίστηκαν ως ύποπτες από την FNA). Παρ' όλα αυτά, η χαμηλή επιτυγχανόμενη ακρίβεια και η κρισιμότητα της εφαρμογής καθιστούν αμφίβολη την κλινική εφαρμογή της αυτόματης κατηγοριοποίησης, στα δεδομένα αυτά.

Παρατηρώντας τους παραπάνω πίνακες και διαγράμματα μπορεί κάποιος να καταλάβει ότι κάνοντας χρήση του αλγορίθμου των τεχνητών νευρωνικών δικτύων, με 7 νευρώνες και διαχωρισμό δεδομένων σε σύνολα εκπαίδευσης και ελέγχου 40% και 60% αντίστοιχα, επιτυγχάνεται συνολική ακρίβεια που ανέρχεται στο 98,3%. Επίσης ο αλγόριθμος kNN με πλήθος γειτόνων k=3 και διαχωρισμό δεδομένων σε ποσοστό 80% ελέγχου και 20% εκπαίδευσης, επιτυγχάνει ακρίβεια 86,7%, κατά μέσο όρο.

Όσον αφορά τον αλγόριθμο τεχνητού ανοσοποιητικού συστήματος, συμπεραίνουμε ότι αν τον εφαρμόσουμε χρησιμοποιώντας τις καλύτερες τιμές για τις παραμέτρους Clonal_rate, Hyperclonal_rate, Max_resources, Mutation-rate και Stimulation_threshold, και το καλύτερο ποσοστό διαχωρισμού των δεδομένων, σε δεδομένα ελέγχου και δεδομένα εκπαίδευσης, μπορεί να μας δώσει επιτυχή αποτελέσματα στο 98,9%. Με βάση τα αποτελέσματα από τα πειράματα, καλύτερες τιμές για τις παραπάνω παραμέτρους θεωρήθηκαν οι εξής: Clonal_rate = 20, Hyperclonal_rate = 15, Max_resources = 600, Mutation-rate = 0,5 και Stimulation_threshold=0,5. Επίσης καλύτερο ποσοστό διαχωρισμού θεωρήθηκε το ποσοστό test – train = 60 - 40.

Συνοψίζοντας λοιπόν, συμπεραίνουμε ότι, η μέθοδος που προτείνουμε έχει τα καλύτερα αποτελέσματα για την κατηγοριοποίηση των δεδομένων σε σύγκριση με τις άλλες δύο μεθόδους ταξινόμησης, με ποσοστό επιτυχίας στο 98,9% κατά μέσο όρο. Ακολουθεί η μέθοδος των τεχνιτών νευρωνικών δικτύων και τέλος η μέθοδος των k πλησιέστερων γειτόνων, με ποσοστά 98,3% και 86,7% αντίστοιχα.

Όσον αφορά τα suspicious δεδομένα βλέπουμε πως τα αποτελέσματα δεν βρίσκονται στο ίδιο υψηλό επίπεδο με τα non-suspicious, ήταν όμως αναμενόμενα. Όπως αναφέραμε και στην παράγραφο περιγραφής των δεδομένων, μικρό ποσοστό των δεδομένων που η FNA έκρινε ως ύποπτα ήταν στην πραγματικότητα κακοήθη. Βλέπουμε όμως ότι με την μέθοδο που μελετάμε, η ακρίβεια της δεύτερης κλάσης, δηλαδή τα κακοήθη, είναι αρκετά υψηλή, φτάνοντας την τάξη του 91% και με τις μεθόδους ΤΝΔ και kNN, να φτάνει σε ποσοστό 76,5% και 93,2% αντίστοιχα. Έτσι μπορούμε να πούμε για ένα δείγμα αν είναι στην πραγματικότητα κακοήθη, σύμφωνα με το αν χαρακτηρίστηκε ύποπτο από την FNA.

Η χρήση ενός αλγορίθμου ταξινόμησης λοιπόν, ο οποίος θα μπορεί να κατηγοριοποιεί με επιτυχία τα δεδομένα έχει μεγάλη σημασία για το πρόβλημα που μελετάμε καθώς είναι απαραίτητο για την διάγνωση των δειγμάτων της βιοψίας λεπτής βελόνας και με την εξέλιξη του τομέα και την ανησυχία των ερευνητών, νέες μέθοδοι θα προκύπτουν συνεχώς με καλύτερα αποτελέσματα.

5 Βιβλιογραφία

1. Rogers WF, Asper SP, Williams RH. Clinical significance of malignant neoplasms of the thyroid gland. N Engl J Med. 1947.
2. Miller JM. Carcinoma and thyroid nodules. Problem in endemic goiter. N EnglJ Med.
3. Crockford PM, Bain GO. Fine-needle aspiration biopsy of the thyroid. Can Med Assoc J.
4. Crile G Jr, Hawk WA Jr. Aspiration biopsy of thyroid nodules. Surg Gynecol Obstet.(8 8.Wang C, Vickery AL Jr, Maloof F. Needle biopsy of the thyroid. Surg Gynecol Obstet.
5. Miller JM, Hamburger JI, Kini SR. The impact of needle biopsy on the preoperative diagnosis of thyroid nodules. Henry Ford Hosp Med J.
6. Young J., Fine Needle Aspiration Cytopathology Blackwell, 1993.
7. G. Kocjan, Fine Needle Aspiration Cytopathology, Diagnostic Principles and Dilemmas, 2006.
8. Ι. Βλαχάβας, Π. Κεφαλάς, Ν. Βασιλειάδης, Φ. Κόκορας, Η. Σακελλαρίου, Τεχνητή Νοημοσύνη, Γκιούρδας Εκδοτική, Ελλάδα 2006 (Γ' Έκδοση).
9. Ιωάννης Ε. Ψαρουδάκης, Ανάπτυξη και Αξιολόγηση Γενετικών Αλγορίθμων Ανακάλυψης Κανόνων για την Κατηγοριοποίηση Δεδομένων, Διπλωματική Εργασία, 2008
10. Ιάσονας Γ. Διγαλάκης, Παράλληλοι Μιμητικοί Αλγόριθμοι, 2005.
11. Frank, A. & Asuncion, A. (2010). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
12. Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems. Ann. Eugenics 7, Pt. II, 179-188
13. Καραγίαννης Αντώνης, κατάταξη δειγμάτων σε κατηγορίες με αυτό-οργανούμενους χάρτες
14. Konstantinos K. Delibasis, Pantelis A. Asvestas, George K.Mastopoulos, Computer-Aided Diagnosis of Thyroid Malignancy Using an Artificial Immune System Classification Algorithm
15. Βλαχάβας Ι, Κεφαλάς Π., Βασιλειάδης Ν, Κόκκορας Φ. και Σακελλαρίου Η.,2006
16. Δημητράτζου Αναστασία, Αναγνώριση χαρακτηριστικών μεγεθών σε Προκλητά Δυναμικά εγκεφάλου και αυτόματη κατηγοριοποίησή τους, 2009
17. Margaret H. Dunham, Data Mining, Εισαγωγικά και Προηγμένα Θέματα Εξόρυξης Γνώσης από Δεδομένα
18. Χρονάκης Ιωάννης, Επεκτάσεις και Περαιτέρω Αξιολόγηση Συστήματος Αναγνώρισης Μερών του Λόγου για Ελληνικά Κείμενα, 2006
19. Μιχάλης-Γεράσιμου Στρίντζη, Αναγνώριση Προτύπων
20. Βιολογία Γενικής Παιδείας Γ' Λυκείου, Υπουργείο Εθνικής Παιδείας και Θρησκευμάτων, Παιδαγωγικό Ινστιτούτο

21. Steve Cayzer, HP Labs Bristol, Artificial Immune Systems
22. Donald E. Goodman, JR. Lois C. Boggess, Andrew B. Wakins, Artificial Immune System Classification of MultipleclassProblems.
23. Steven A. Hofmeyr, An Interpretative Introduction to the Immune System, Dept. of Computer Science University of Mexico, April 2000
24. De Castro, L. N. and J. Timmis: 2002b, Artificial Immune Systems: A New Computational Intelligence Approach. Spinger-Verlag.
25. Zhou Ji, Dipankar Dasgupra, Zhiling Yang and Honmei Teng, Analysis of Dental Images using Artificial Immune Systems
26. Dipankar Dasgupra, Fabio Gonzalez, An Immunity-Based Technique to characterize Instructions in Computer Networks
27. Jon Timmis and Camilla Edmonds, A comment on ot-Ainet: An Immune Network Algorithm for optimization.
28. Liangpei ZHANG, Yanfei ZHONG, Pingxiang LI, Applications of Artificial Immune Systems in Remote sensing image Classification.
29. Jon Timmis and Thomas Knight, Artificial Immune Systems: Using the Immune System as Inspiration for Data Mining.
30. Leonardo Nunes de Castro, Fernando Jose von Zuben, Artificial Immune Systems, Part II – A Servey of Applications.
31. Fatma Latifoglu, Kemal Polat, Sadık Kara, Salih Gunes, Medical diagnosis of atherosclerosis from Carotid Artery Doppler Signals using principal component analysis (PCA), k-NN4 based weighting pre-processing and Artificial Immune Recognition System (AIRS), 2006
32. Kemal Polat, Seral S_ahan, Salih Gunes, A novel hybrid method based on artificial immune recognition system (AIRS) with fuzzy weighted pre-processing for thyroid disease diagnosis, 2007
33. Seral, Sahana, Kemal Polata, Halife Kodazb, Salih Gunes, A new hybrid method based on fuzzy-artificial immune system and k-nn algorithm for breast cancer diagnosis, 2006
34. Giuseppe Nicosia Cineca, High Performance Systems, Computational Immunology and Immunological Computation.
35. Aurenhammer, F. and Klein, R. "Voronoi Diagrams." Ch. 5 in Handbook of Computational Geometry (Ed. J.-R. Sack and J. Urrutia). Amsterdam, Netherlands: North-Holland, pp. 201-290, 2000.
36. Κωνσταντίνος Π. Τσιρογιάννης , Διάγραμμα Voronoi κύκλων και υλοποιήσεις στην CGAL, Διπλωματική Εργασία 2007
37. Karasick Michael, Lieber Derek, Nackman Lee, Efficient Delaunay Triangulation Using Rational Arithmetic, In Proceedings of the Symposium on Computational Geometry, ACM, 1991
38. Liotta Giuseppe, Preparata Franco P., Tamassia Roberto, Robust Proximity Queries in Implicit Voronoi Diagrams, 1996
39. D. Goldberg, Genetic Algorithms in Search, Optimization, and Machine Learning, Adisson – Wesley, 1989.

Παράρτημα

Αλγόριθμος AIRS

```
function [MC,pcc,N1,N2] = AIRS(AG, nclass, Test, max_resources,
hyperclonal_rate, clonal_rate, mutation_rate, stim_thres )
% AG:  δεδομένα εκπαίδευσης (train data)
% nclass: πλήθος κλάσεων
% Test : δεδομένα ελέγχου
% max_resources: μέγιστος αριθμός πόρων
% hyperclonal_rate: ποσοστό κλώνων που εισάγονται στον πληθυσμό
% clonal_rate: ποσοστό κλώνων που δημιουργούνται
% mutation_rate: πιθανότητα μετάλλαξης
% stim_thres: κατώφλι διέγερσης(κριτήριο τερματισμού )

MC = [];
AB = [];
MChist=[];
maxpar=max(AG);
minpar=min(AG);
nparam = size(AG,2)-1;                  % αριθμός στοιχείων
worst_affinity = affinity(maxpar,minpar);

% υπολογισμός κατώτατου ορίου συγγένειας
NAG = size(AG,1);
affinity_thres = sum(sum(triu(affinity(AG, AG)))) / (NAG*(NAG-1)/2);
ATS = 0.5;
stop_criterion = zeros(1, nclass);
ind1 = find(Test(:,1) == 1);
ind2 = find(Test(:,1) == 2);
N1 = length(ind1);
N2 = length(ind2);

% έναρξη επαναλήψεων
perm = randperm(NAG);    %γίνεται αλλαγή της σειράς των στοιχείων του NAG
pcc = zeros(NAG,2);
for ag_i = 1 : NAG,
  antigen = AG(perm(ag_i), :);

% εύρεση περισσότερο όμοιου κυττάρου μνήμης, το οποίο είναι αυτό με την μικρότερη
%τιμή συγγένειας
 if isempty(MC)
    MC=[MC ; antigen];
    i_match = 1;
    afmin = 0;
```

```
  else
  ind = find(MC(:,1) == antigen(1)); % εύρεση MC ίδιας κλάσης με το αντιγόνο
    if isempty(ind)
      MC=[MC ; antigen];
      i_match = size(MC,1);
      afmin = 0;
    else
      af = affinity(antigen, MC(ind,:));
      [afmin, imin] = min(af);
      i_match = ind(imin);
    end
  end
  mc_match = MC(i_match, :);
  stim_match = (worst_affinity - affinity(mc_match , antigen)) /
worst_affinity;

% δημιουργία mc κλώνων στο MU. Το MU περιέχει το MC(i_match,:) και τους
%κλώνους του
 NumClones = hyperclonal_rate * clonal_rate * stim_match;
  MU = mc_match;
  while size(MU,1) < NumClones
    [mc_clone, mut] = mutate(mc_match, minpar, maxpar,
mutation_rate, nclass);
    if mut
      MU= [MU; mc_clone];
    end
  end
  AB = [AB; MU];

 % υπολογισμός stimulation τιμής για το AB και υπολογισμός πόρων για το AB
 stim_AB = Stimulation(AB, antigen, worst_affinity);
  resources = clonal_rate * stim_AB;
  iter = 1;
  while 1,
    [AB,resources,stim_AB] = candidate_mc(AB, antigen, stim_AB,
resources, max_resources, nclass);

 % εάν είναι η πρώτη επανάληψη, δημιουργώ AB clones, αλλιώς ελέγχω το κριτήριο
%τερματισμού
    if iter > 1
      for i = 1 : nclass,
        ind = find(AB(:,1) == i);
        stop_criterion(i) = sum(stim_AB(ind)) / length(ind);
      end
      if sum(stop_criterion < stim_thres) == 0
        break;
      end
    end

% δημιουργία AB κλώνων στο AB, ενημέρωση της stim_AB και των πόρων
```

```
    MU = [];
    rd = rand(size(stim_AB));
    ind = find(stim_AB > rd);
    for k = 1 : length(ind),
      NumClones = stim_AB(ind(k)) * clonal_rate;
      for i = 1 : NumClones,
        [ab_clone, mut] = mutate(AB(ind(k),:), minpar, maxpar,
mutation_rate, nclass);
        if (mut)
          MU = [MU; ab_clone];
        end
      end
    end
    stim_MU = Stimulation(MU, antigen, worst_affinity);
    AB = [AB; MU];
    stim_AB = [stim_AB stim_MU];
    resources = [resources clonal_rate * stim_MU];
    iter = iter + 1;
  end

% επιλογή υποψήφιου mc από το AB και εισαγωγή του στο MC
 ind = find(AB(:,1) == antigen(1));
  if ~isempty(ind)
    [stim_cand, ind_max] = max(stim_AB(ind));
    mc_candidate = AB(ind(ind_max), :);
    if stim_cand > stim_match
      cell_aff = affinity(mc_candidate, mc_match);
% εάν το υποψήφιο mc έχει αξιοσημείωτη συγγένεια με το matching memory cell,
%απομακρύνω το matching memory cell
      if cell_aff < affinity_thres * ATS
        MC(i_match,:) = [];
      end
      MC = [MC; mc_candidate];
      MChist=[MChist;[ag_i*ones(size(MC,1),1),MC]];
    end
  end
  if ~isempty(MC)
    if size(MC,1) > 1
      [aff_min, imin] = min(affinity(Test, MC));
    else
      imin = ones(1,size(Test,1));
    end
    pcc(ag_i,:) = [sum(Test(ind1,1) == MC(imin(ind1),1))
         sum(Test(ind2,1) == MC(imin(ind2),1))] ./ [N1 N2];
    fprintf('Antigen: %d / %d, memory cels: %d Performance = [%g
         %g]\n', ag_i, NAG, size(MC,1), pcc(ag_i,1),
         pcc(ag_i,2));
  end
  end
csvwrite('MChist.txt',MChist);
```

Συνάρτηση affinity

```
function af = affinity(y, X)
af = distfcm(X(:,2:end), y(:,2:end));
```

Συνάρτηση mutate

```
function [y,mut] = mutate(x,minparam,maxparam,mutation_rate,nclass)

N = length(x);
y = x;
change = rand(1,N);
change_to = rand(1,N);
ind = find(change(2:end) < mutation_rate);
if ~isempty(ind)
  ind = ind + 1;
  mut = true;
  y(ind) = minparam(ind) + change_to(ind) .* (maxparam(ind) –
       minparam(ind));
  if(change(1) < mutation_rate)
    y(1) = randint(1, 1 ,[1,nclass]);
  end
else
  mut = false;
end
```

Συνάρτηση Stimulation

```
function stim_AB = Stimulation(AB, antigen, worst_affinity)

ind1 = find(AB(:,1) == antigen(1));
ind2 = find(AB(:,1) ~= antigen(1));
stim_AB = zeros(1,size(AB,1));
if ~isempty(ind1)
  stim_AB(ind1) = (worst_affinity - affinity(AB(ind1,:),antigen)) /
worst_affinity;
end
if ~isempty(ind2)
  stim_AB(ind2) = affinity(AB(ind2,:),antigen) / worst_affinity;
end
```

Συνάρτηση candidate_mc

```
function
[AB,resources,stim]=candidate_mc(AB,ag,stim,resources,max_resources,n
class)
% επιλέγει τα στοιχεία του AB που πρέπει να αφαιρεθούν ώστε resources=max_resources
for i=1:nclass
  if i == ag(1)
    res_allowed = max_resources/2;
  else
    res_allowed = max_resources/(2*(nclass-1));
  end
  ind = find(AB(:,1) == i);
  if ~isempty(ind)
    [sort_res, sort_ind] = sort(resources(ind));
    res_alloc = sum(sort_res);
    res_remove = res_alloc - res_allowed;
    cs = cumsum(sort_res);
    i1 = find( cs <= res_remove);
    if~isempty(i1)        % αφαιρεί αυτό με τους λιγότερους πόρους
      irem = ind(sort_ind(i1));
      d = res_remove - cs(i1(end));
      if(d > 0)
        j = ind(sort_ind(i1(end)+1));
        resources(j) = resources(j) - d;
      end
      resources(irem) = [];
      AB(irem,:)   = [];
      stim(irem)   = [];
    end
  end
end
```

Αλγόριθμος kNN

kNN για τα Iris δεδομένα

Διαχωρισμός δεδομένων

```
function [TP,TN,N1,N2,TEST1,TRAIN1]=spl_data
% C: πίνακας δεδομένων
C=csvread('irisdata.txt');
%επειδή το label στην τελευταία στήλη κάνουμε εναλλαγή πρώτης και τελευταίας.
c=C(:,1);
c1=C(:,end);
C(:,1)=c1;
```

```
C(:,end)=c;

%κανονικοποιούμε τις τιμές των δεδομένων
for n=2:5
  a=1/(max(C(:,n))-min(C(:,n)));
  b=-a*min(C(:,n));
  C1(:,n)=a.*C(:,n)+ b;
end
 C1(:,1)=C(:,1);
 C=C1;

%διαχωρισμός των κλάσεων
class1=[];
class2=[];
N=size(C);
for i=1:N
  if C(i,1)==1
    class1=[class1; C(i,:)];
  else
    class2=[class2; C(i,:)];
  end
end

%διαχωρισμός σε test-train
s=[];
sosto=0;
TP=0;
TN=0;
FP=0;
FN=0;
TEST1=[];
TRAIN2=[];

t0=cputime

TEST1=[];
TRAIN1=[];
for i=1:N
  r=rand;
  if r<0.9
    TEST1=[TEST1;C(i,:)];
  else
    TRAIN1=[TRAIN1;C(i,:)];
  end

end

[class_out,TP,TN]=newknn2(TEST1,TRAIN1,7,2);   %καλείται ο βασικός
αλγόριθμος
ind1 = find(TEST1(:,1) == 1);
ind2 = find(TEST1(:,1) == 2);
```

```
N1 = length(ind1);
N2 = length(ind2);
accA=TP/N1;
accB=TN/N2;
pcc=((accA*N1)+(accB*N2))/(N1+N2);

t1=cputime;
fprintf('CPU time for knn algolithm = %g\n',t1-t0);
```

kNN για τα συνθετικά δεδομένα

Διαχωρισμός δεδομένων

```
I1=imread('εικόνα');
D=I1;
[Nrow,Ncol]=size(I1);
class1=[];
class2=[];
% διαχωρισμός κλάσεων
for i=1:Nrow
  for j=1:Ncol
    if D(i,j)==0
      class1=[class1;[i,j]];
    end
    if D(i,j)==255
      class2=[class2;[i,j]];
    end
  end
end

label1=ones(size(class1,1),1);
class1=[label1, class1];
label2=2*ones(size(class2,1),1);
class2=[label2, class2];
C=[class1; class2];

s=[];
sosto=0;
TP=0;
TN=0;
FP=0;
FN=0;
CTEST=[];
CTRAIN=[];

t0=cputime

% διαχωρισμός σε test-train
for i=1:length(C)
  r=rand;
```

```
  if r<0.9
    CTEST=[CTEST;C(i,:)];
  else
    CTRAIN=[CTRAIN;C(i,:)];
  end

end
[class_out,Confusion]=newknn2(CTEST,CTRAIN,9,2);

ind1 = find(CTEST(:,1) == 1);
ind2 = find(CTEST(:,1) == 2);
N1 = length(ind1);
N2 = length(ind2);

accA=TP/N1;
accB=TN/N2;
pcc=((accA*N1)+(accB*N2))/(N1+N2);
t1=cputime;
fprintf('CPU time for knn algolithm = %g\n',t1-t0);
```

Βασικός αλγόριθμος kNN

```
function [class_out,TP,TN]=newknn2(CTEST,CTRAIN,K,nclass)
% K: αριθμός γειτόνων
TP=0;
TN=0;
FP=0;
FN=0;

Ntest=size(CTEST,1);
Ntrain=size(CTRAIN,1);
for i=1:Ntest
  D=[];
  for j=1:Ntrain
    c=0;
%υπολογισμός απόστασης στοιχείων και γειτόνων
      c=c+((CTEST(i,2)-CTRAIN(j,2)).^2+((CTEST(i,3)-
CTRAIN(j,3)).^2));
    D=[D,c];
  end
  [Dsort,Jsort]=sort(D);
  p=Jsort(2:K+1);
  h=hist(CTRAIN(p,1),[0:nclass]);

  [m,most_class]=max(h);
  most_class=most_class-1;

  class_out(i)=most_class;  % η κλάση στην οποία ανήκει το στοιχείο δίνεται ως
```

```
                                %έξοδος
end
for i=1:length(class_out)
if CTEST(i,1)==1
  if class_out(i)==1
    TP=TP+1;
  else
    FN=FN+1;
  end
else
  if class_out(i)==2
    TN=TN+1;
  else
    FP=FP+1;
  end
end
end
Confusion=[TP,FN;FP,TN];
```