ΠΑΝΕΠΙΣΤΗΜΙΟ ΣΤΕΡΕΑΣ ΕΛΛΑΔΑΣ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΙΑΤΡΙΚΗ

Τεχνικές διαχείρησης Ιατρικών Δεδομένων για τον καθορισμό Νέας Γνώσης

Πτυχιακή Εργασία

Βλαχάβα Βασιλική - Δανάη

Επιβλέπων:
Παπαγεωργίου Ελπινίκη

Λαμία, 2009

ΠΑΝΕΠΙΣΤΗΜΙΟ ΣΤΕΡΕΑΣ ΕΛΑΔΑΣ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΙΑΤΡΙΚΗ

Τεχνικές διαχείρισης Ιατρικών Δεδομένων για τον καθορισμό Νέας Γνώσης

Πτυχιακή Εργασία

Βλαχάβα Βασιλική-Δανάη

Επιβλέπων:

Παπαγεωργίου Ελπινίκη

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή

.....	.....	.....
Παπαγεωργίου	**Βασιλακόπουλος**	**Μαγκλογιάννης**
Ελπινίκη	**Μιχαήλ**	**Ηλίας**

Λαμία, 2009

Ευχαριστίες

Η παρούσα πτυχιακή εργασία εκπονήθηκε το χρονικό διάστημα από τον Δεκέμβριο του 2007 έως τον Σεπτέμβριο του 2009 στο Τμήμα Πληροφορικής με εφαρμογές στη βιοϊατρική του Πανεπιστημίου Στερεάς Ελλάδας.

Θα ήθελα να ευχαριστήσω την επιβλέπουσα καθηγήτριά μου, κα Ελπινίκη Παπαγεωργίου για την σημαντική βοήθεια που μου προσέφερε για να ολοκληρώσω την πτυχιακή μου εργασία.

Τέλος θα ήθελα να εκφράσω τις ευχαριστίες μου στην οικογένειά μου και σε φίλους που η κατανόηση και η ψυχολογική τους συνδρομή ήταν απαραίτητη για την περάτωση της εργασίας αυτής.

Περίληψη

Σκοπός της παρούσας πτυχιακής εργασίας είναι η παράθεση των πιο σημαντικών αλγορίθμων και εργαλείων εξόρυξης γνώσης από δεδομένα και η εφαρμογή τους σε ιατρικά δεδομένα. Επιπλέον γίνεται μια εισαγωγή στην Ασαφή Λογική και σε τεχνικές ασαφούς ανακάλυψης γνώσης από δεδομένα. Από τα εργαλεία που είναι διαθέσιμα, επιλέχτηκε το Matlab, το οποίο υποστηρίζει πλήθος αλγορίθμων και προσφέρει πολλές δυνατότητες. Ως πεδία εφαρμογής επιλέχτηκαν ιατρικά δεδομένα που αφορούν σε ασθενείς που πάσχουν από παθήσεις του θυρεοειδή αδένα και του διαβήτη, τα οποία αποτελούν και τη βάση της μελέτης όπου εφαρμόζονται τεχνικές και αλγόριθμοι εξόρυξης γνώσης και γίνεται μια σύγκριση των αποτελεσμάτων τους.

Λέξεις Κλειδιά: Μηχανική Μάθηση, αλγόριθμοι ταξινόμησης και ομαδοποίησης δεδομένων, δέντρα απόφασης, Ασαφής Λογική, Ανακάλυψη Γνώσης σε Ιατρικά δεδομένα, k- nearest neighbor, k-means, fuzzy k-means

Abstract

The aim of this work is the presentation of the most important data mining algorithms and their application in medical data. An introduction to fuzzy logic and fuzzy data mining techniques is also given. We have chosen Matlab as a platform because it supports numerous algorithms and offers many capabilities. We applied some of them to medical data concerning thyroid and diabetes and we compare their results.

KeyWords: Machine Learning, Classification and clustering algorithms, Data Mining, Fuzzy logic, Fuzzy Data Mining, Data Mining in Medical Data, decision trees, k- nearest neighbor, k-means, fuzzy k-means.

Περιεχόμενα

ΚΕΦΑΛΑΙΟ 1

ΕΙΣΑΓΩΓΗ

Στην εποχή της πληροφορίας που βασική πεποίθηση είναι ότι η πληροφορία είναι αυτή που προσφέρει δύναμη και επιτυχία, η συλλογή τεράστιου όγκου δεδομένων και πληροφοριών είναι χαρακτηριστική. Ο ρυθμός αυτός συλλογής αυξάνεται ολοένα και περισσότερο με τις δυνατότητες που προσφέρουν οι νέες τεχνολογίες και η πληροφορική.

Τα χιλιάδες δεδομένα ποικίλουν από απλές αριθμητικές μετρήσεις και έγγραφα σε πιο περίπλοκες πληροφορίες όπως χωρικά δεδομένα, πολυμεσικά δεδομένα και έγγραφα υπερκειμένου. Ορισμένες μόνο ενδεικτικές κατηγορίες δεδομένων που συλλέγονται είναι: συναλλαγές εταιρειών που καταγράφονται για λόγους ιστορικότητας και έχουν σχέση τόσο με τις συναλλαγές των εταιρειών με άλλες εταιρείες όσο και με την εσωτερική τους λειτουργία, επιστημονικά δεδομένα, ιατρικά και προσωπικά δεδομένα που συλλέγονται από κυβερνήσεις, επιχειρήσεις και νοσηλευτικά ιδρύματα για βελτίωση της ποιότητας των παρεχόμενων υπηρεσιών, βίντεο και εικόνες, αναφορές και σημειώματα, μηνύματα ηλεκτρονικού ταχυδρομείου κτλ.

Αρχικά, η συλλογή αυτών των δεδομένων γινόταν ανεξέλεγκτα βάσει της τεράστιας δύναμης που προσέφερε η ψηφιακή αποθήκευση, αδιαφορώντας για τις δυνατότητες των υπαρχόντων αποθηκευτικών δομών σχετικά με την επεξεργασία των δεδομένων. Αυτό οδήγησε στη δημιουργία δομημένων συστημάτων διαχείρισης βάσεων δεδομένων (κυρίως σχεσιακών) που προσέφεραν υπηρεσίες αποτελεσματικής και αποδοτικής ανάκτησης πληροφοριών από τεράστιες συλλογές δεδομένων. Βέβαια, οι μόνες δυνατότητες επεξεργασίας που προσέφεραν τα παραπάνω συστήματα ήταν σχεσιακού τύπου ερωτήματα για επιλογή ενός υποσυνόλου των δεδομένων από το καθολικό σύνολο βάσει κάποιων κριτηρίων και διαδικασίες στατιστικής επεξεργασίας των δεδομένων, δυνατότητες στις οποίες ο χρήστης έχει θέσει εκ των προτέρων ένα συγκεκριμένο στόχο προς αναζήτηση [Connolly and Begg, 2004].

Σήμερα, τα δεδομένα που έχουμε είναι πολλά περισσότερα από αυτά τα οποία μπορούμε να διαχειριστούμε. Πολλές φορές μάλιστα το πλήθος των δεδομένων και το μέγεθος των βάσεων δεδομένων αυξάνονται τόσο ως προς το πλήθος των εγγράφων όσο και ως προς το πλήθος των πεδίων. Πίσω από αυτές τις τεράστιες βάσεις δεδομένων υπάρχει μη προφανής γνώση που δεν είναι ορατή και γνωστή εκ των προτέρων, αλλά μπορεί να αποδειχτεί πολύ χρήσιμη αν ανακαλυφθεί. Την απαίτηση

αυτή έρχεται να καλύψει ένα νέο επιστημονικό πεδίο η Ανακάλυψη Γνώσης σε Βάσεις Δεδομένων (Knowledge Discovery in Databases) που με την εφαρμογή μεθόδων και τεχνικών εξόρυξης γνώσης (Data Mining) στοχεύει στην ανακάλυψη προτύπων και κατασκευή μοντέλων από τα δεδομένα [Dunham, 2002], [Xie et.al 2009], [Witten and Frank, 2005], [Wu and Kumar, 2009] .

Πολλοί είναι οι τομείς των επιστημών και των επιχειρήσεων που χρησιμοποιούν, και μάλιστα σε μεγάλη έκταση, τις τεχνικές ανακάλυψης και εξόρυξης γνώσης, όπως ο τομέας του εμπορίου και της διαφήμισης, των οικονομικών και των επενδύσεων, των τηλεπικοινωνιών και της ασφάλειας των δικτύων και των υπολογιστικών συστημάτων, κλπ. [Berry and Gordon, 2004] [Giannotti and Pedreschi, 2008], [Kohavi and Provost, 2001], [Mattison 1997].

Στον τομέα της ανάλυσης βιολογικών δεδομένων τα τελευταία χρόνια γνωρίζει ιδιαίτερη ανάπτυξη. Παραδείγματα είναι η αναγνώριση και η ανάλυση του γονιδιώματος του ανθρώπου και άλλων οργανισμών, η αναζήτηση των γενετικών δικτύων και των πρωτεϊνών και η ανάπτυξη νέων φαρμάκων βασισμένων στο γενετικό προφίλ του κάθε ασθενούς. Συνεπώς, η εξόρυξη γνώσης στον τομέα της βιολογίας είναι πολύ σημαντική και οδήγησε σε ένα νέο επιστημονικό πεδίο που καλείται βιοπληροφορική. Η εξόρυξη γνώσης στον τομέα της βιολογίας αναφέρεται στην εφαρμογή τεχνικών για την σημασιολογική ολοκλήρωση ετερογενών και κατανεμημένων βάσεων γονιδίων και πρωτεϊνών, ευθυγράμμιση, ευρετηριοποίηση και ανάλυση πρωτεϊνικών ακολουθιών, ανακάλυψη δομικών προτύπων και ανάλυση των γενετικών δικτύων και πρωτεϊνικών μονοπατιών [Chen and Lonardi, 2009], [Wang et al, 2004].

Στο τομέα της υγείας, πολλοί οργανισμοί παροχής υπηρεσιών ιατρικής περίθαλψης διατηρούν αποθηκευμένα πλήθος κλινικών, δημογραφικών, οικονομικών και κοινωνικοοικονομικών δεδομένων που αφορούν τόσο σε ασθενείς όσο και στους ίδιους τους οργανισμούς. Η εφαρμογή τεχνικών εξόρυξης γνώσης μπορεί να φανεί χρήσιμη για την ανακάλυψη κρυμμένης ιατρικής γνώσης από τα δεδομένα που διατηρούνται στους ηλεκτρονικούς ιατρικούς φακέλους των ασθενών και που μπορεί να είναι πολύτιμη για την ισχυροποίηση κάποιων ιατρικών συμπερασμάτων, αλλά και την αύξηση ήδη υπάρχουσας γνώσης. Παράλληλα, μπορεί να φανούν χρήσιμα εργαλεία για την διοικητικοοικονομική διαχείριση τέτοιων οργανισμών [Berka et al, 2009].

Είναι αποδεκτό ότι ο όγκος των δεδομένων που συλλέγεται φτάνει σε τεράστιο μέγεθος. Ο όγκος αυτός είναι ασφαλώς απαγορευτικός για την επεξεργασία του από ανθρώπους χωρίς την χρήση κατάλληλων εργαλείων και μεθόδων. Επίσης, εργαλεία ερωταποκρίσεων δεν επαρκούν για την επεξεργασία και την αξιοποίηση των δεδομένων.

Για την αντιμετώπιση του ανωτέρω προβλήματος έχουν αναπτυχθεί πολλοί αλγόριθμοι και τεχνικές καθώς και εργαλεία που ενσωματώνουν πληθώρα τεχνικών και αλγορίθμων για την αναζήτηση και εξόρυξη νέων πληροφοριών από αυτές τις βάσεις δεδομένων. Κάποια από τα εργαλεία αυτά είναι π.χ. το SPSS, ο DBMiner, ο Intelligent Miner κ.τ.λ. Το εργαλείο που χρησιμοποιήθηκε για την ολοκλήρωση της

παρούσας εργασίας είναι το Matlab, το οποίο αποτελεί προϊόν ερευνητικού επιπέδου, μια παρουσίαση του οποίου δίδεται στα επόμενα κεφάλαια

Δομή της Εργασίας

Το παρόν κείμενο είναι χωρισμένο σε πέντε κεφάλαια.

Το πρώτο κεφάλαιο είναι μια γενικού σκοπού εισαγωγή σχετικά με την Ανακάλυψη Γνώσης σε Βάσεις Δεδομένων και των τομέων εφαρμογής της.

Το δεύτερο κεφάλαιο, με τίτλο Μηχανική Μάθηση και Ανακάλυψη Γνώσης, παρέχει τις βασικές έννοιες σχετικά με την κατανόηση της επιστημονικής περιοχής που κινείται η εργασία. Αναφέρονται δηλαδή, οι έννοιες της ανακάλυψης γνώσης και της εξόρυξης δεδομένων, η σχέση τους με την μηχανική μάθηση και την στατιστική, η διαδικασία της ανακάλυψης γνώσης, τα προβλήματα που τυχόν θα προκύψουν κατά την εκτέλεση της διαδικασίας, τα αποτελέσματα της διαδικασίας καθώς και οι μέθοδοι και οι τεχνικές εξόρυξης σε δεδομένα, αλλά και οι δομές στις οποίες μπορούν οι προηγούμενες να εφαρμοστούν.

Στο τρίτο κεφάλαιο, με τίτλο Ασαφείς Τεχνικές Ανακάλυψης Γνώσης, γίνεται μια σύντομη παρουσίαση της θεωρίας ασαφών συνόλων και ενδεικτικών αλγορίθμων ασαφούς μάθησης. Παρουσιάζεται επίσης συνοπτικά το εργαλείο Matlab.

Στο τέταρτο κεφάλαιο, με τίτλο Ανακάλυψη Γνώσης σε Ιατρικά Δεδομένα, παρουσιάζονται και περιγράφονται τα ιατρικά δεδομένα που αφορούν σε ασθενείς που πάσχουν από παθήσεις του θυρεοειδή αδένα και διαβήτη τα οποία αποτελούν κα τη βάση της μελέτης και εφαρμόζονται τεχνικές και αλγόριθμοι ανακάλυψης και εξόρυξης γνώσης και γίνεται μια σύγκριση των αποτελεσμάτων τους.

Τέλος, το πέμπτο κεφάλαιο παρουσιάζει τα Συμπεράσματα και γίνεται μια συνολική επισκόπηση όλων όσων παρουσιάστηκαν καθώς και κάποιες προτάσεις για πιθανή μελλοντική εργασία.

ΚΕΦΑΛΑΙΟ 2

ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ ΚΑΙ ΑΝΑΚΑΛΥΨΗ ΓΝΩΣΗΣ

2.1 Εισαγωγή

Η τεράστια πρόοδος της επιστήμης της πληροφορικής ενθάρρυνε την μαζική συλλογή και αποθήκευση δεδομένων σε όλους τους τομείς της ανθρώπινης δραστηριότητας. Η τεράστια πρόοδος των βάσεων δεδομένων, όλων των μεγεθών και τύπων, είναι ενδεικτική της ικανότητας για συλλογή δεδομένων, αλλά ταυτόχρονα αυξάνει την αναγκαιότητα για καλύτερες μεθόδους πρόσβασης και ανάλυσης των δεδομένων με σκοπό την ανακάλυψη νέας γνώσης.

Στην δεκαετία του 1980 εμφανίστηκε ο όρος Εξόρυξη σε Δεδομένα (data mining), που χρησιμοποιήθηκε από τους στατιστικολόγους και αναλυτές των δεδομένων και περιέγραφε την εφαρμογή αλγορίθμων για την ανεύρεση προτύπων σε συλλογές δεδομένων. Λίγο αργότερα, το 1989, ένας νέος όρος, Ανακάλυψη Γνώσης σε Βάσεις Δεδομένων (Knowledge Discovery in Databases) χρησιμοποιήθηκε για να αντικαταστήσει όλους τους παλιούς όρους που σκοπός τους ήταν η ανακάλυψη προτύπων και ομοιοτήτων σε δεδομένα. Το νέο αυτό επιστημονικό πεδίο περιέχει στοιχεία από πολλούς άλλους επιστημονικούς κλάδους όπως την τεχνητή νοημοσύνη, μηχανική μάθηση, στατιστική, βάσεις δεδομένων και οπτική αναπαράσταση εννοιών [Dunham, 2002], [Maimon and Rokach 2005], [Witten and Frank, 2005], [Wu and Kumar, 2009].

Πολύ γρήγορα υιοθετήθηκε ως πρακτική από πολλούς επιστήμονες της τεχνητής νοημοσύνης γενικότερα και της **μηχανικής μάθησης** ειδικότερα και χρησιμοποιήθηκε για να περιγράψει την συνολική διαδικασία εξαγωγής γνώσης από βάσεις δεδομένων, από τον προσδιορισμό των στόχων της επιχείρησης ως την τελική ανάλυση των αποτελεσμάτων.

Παράλληλα, σύμφωνα με την θέση που υιοθετήθηκε στο Πρώτο Διεθνές Συνέδριο Ανακάλυψης Γνώσης σε Βάσεις Δεδομένων, που πραγματοποιήθηκε το 1995 στο Μόντρεαλ, ο όρος Εξόρυξη σε Δεδομένα (data mining) περιορίστηκε στην περιγραφή ενός μόνο βήματος της όλης διαδικασίας και πιο συγκεκριμένα αυτό της εφαρμογής των αλγορίθμων εξόρυξης.

2.2 Μηχανική μάθηση

Η *Μηχανική Μάθηση* (machine learning) είναι μια από τις βασικές περιοχές της *Τεχνητής Νοημοσύνης* [Vlahavas et al. 2006] που ασχολείται με υπολογιστικές μεθόδους για την κατασκευή συστημάτων με ικανότητα μάθησης. Ο όρος «*Μηχανική Μάθηση*» έκανε την εμφάνισή του στις αρχές της δεκαετίας του 1980.

Η *Μηχανική Μάθηση* αποτελεί ένα επιστημονικό πεδίο που μελετά την σχεδίαση υπολογιστικών προγραμμάτων ικανών να μαθαίνουν, δηλαδή ικανών να βελτιώνουν την απόδοση τους μέσω της αξιοποίησης προηγούμενης γνώσης και εμπειρίας. Η απόκτηση αυτής της γνώσης και εμπειρίας δεν προκύπτει έπειτα από την αλληλεπίδραση του συστήματος με το περιβάλλον, αλλά από ένα σύνολο κωδικοποιημένων δεδομένων που προέκυψαν από δειγματοληψία στο σύνολο της βάσης δεδομένων και αποτελούν το **σύνολο των δεδομένων εκπαίδευσης** (training set) [Mitchell 1997].

2.3 Μηχανική Μάθηση και Ανακάλυψη Γνώσης

Υπάρχει ισχυρή σχέση ανάμεσα στην Μηχανική Μάθηση και στην Ανακάλυψη Γνώσης σε Βάσεις Δεδομένων εφόσον η δεύτερη αποτελεί μια ειδική περίπτωση της πρώτης, στην οποία όμως ο χώρος αναζήτησης γνώσης περιορίζεται σε μια βάση δεδομένων.

Η ανακάλυψη γνώσης σε βάσεις δεδομένων (Knowledge Discovery in Databases - KDD) είναι μία σύνθετη διαδικασία για τον προσδιορισμό έγκυρων, νέων, χρήσιμων και κατανοητών σχέσεων-προτύπων σε δεδομένα. Αν και ως όρος είναι σχετικά πρόσφατος, αποτελεί μια σημαντική εφαρμογή σε πραγματικές συνθήκες και σε μεγάλη κλίμακα των ερευνητικών αποτελεσμάτων της μηχανικής μάθησης και της στατιστικής.

Η διαδικασία ανακάλυψης γνώσης είναι μια ολοκληρωμένη διαδικασία που περιλαμβάνει την επεξεργασία των δεδομένων, την εφαρμογή των αλγορίθμων ανακάλυψης γνώσης και τέλος την ερμηνεία των αποτελεσμάτων. Χρησιμοποιεί τεχνικές από πολλούς τομείς, όπως στατιστική, μηχανική μάθηση, βάσεις δεδομένων, αναγνώριση προτύπων, πράκτορες, επεξεργασία φυσικής γλώσσας, κτλ.

Οι πιο σημαντικές διαφορές μεταξύ ανακάλυψης γνώσης και μηχανικής μάθησης οφείλονται στο γεγονός ότι η πρώτη εφαρμόζεται σε έναν μεγάλο όγκο δεδομένων τα οποία οργανώνονται σε βάσεις δεδομένων που συνήθως έχουν σχεδιαστεί για άλλο σκοπό. Αντίθετα, στη μηχανική μάθηση τα δεδομένα είναι πολύ λιγότερα και προσεκτικά επιλεγμένα ώστε να εξυπηρετούν καλύτερα τον εκάστοτε σκοπό.

2.4 Εξόρυξη σε Δεδομένα

Ο όρος Ανακάλυψη Γνώσης σε Βάσεις Δεδομένων συχνά ταυτίζεται με τον όρο Εξόρυξη σε δεδομένα (data mining) και αναφέρεται στην εφαρμογή τεχνικών και μεθόδων ανακάλυψης γνώσης σε μεγάλες βάσεις δεδομένων.

Συνεπώς, η Ανακάλυψη Γνώσης σε Βάσεις Δεδομένων αναφέρεται σε μια συνολική, πολλαπλών βημάτων διαδικασία ανακάλυψης γνώσης από τα δεδομένα, συμπεριλαμβανομένου του τρόπου αποθήκευσης και ανάκαμψης των δεδομένων, του τρόπου εφαρμογής των αλγορίθμων σε μαζικά σύνολα δεδομένων που εξακολουθούν όμως να εκτελούνται αποδοτικά, του τρόπου ερμηνείας και οπτικοποίησης των αποτελεσμάτων και του τρόπου μοντελοποίησης της αλληλεπίδρασης ανθρώπου – μηχανής. Η εξόρυξη σε δεδομένα είναι ένα μόνο βήμα της Ανακάλυψης Γνώσης σε Βάσεις Δεδομένων, που περιλαμβάνει την εφαρμογή της ανάλυσης των δεδομένων και των αλγορίθμων που παράγουν πρότυπα ή μοντέλα για τα δεδομένα υπό το πρίσμα αποδεκτών υπολογιστικών περιορισμών αποδοτικότητας.

2.5 Μέθοδοι Μηχανικής Μάθησης

Τα τελευταία χρόνια έχει γίνει πολύ ερευνητική δουλειά στην περιοχή της μηχανικής μάθησης και έχουν γίνει διάφορες προσπάθειες ταξινόμησης αυτής της δουλειάς με βάση διαφορετικά κριτήρια [Michalski 1983] [Michalski et.al 1986]. Μια ταξινόμηση των διάφορων μεθόδων μηχανικής μάθησης οργανώνεται με βάση το πρόβλημα που αντιμετωπίζουν. Τα βασικά προβλήματα είναι:

- Η προσέγγιση μιας συνάρτησης από δεδομένα με παρατηρήσεις τιμών εισόδου και εξόδου της. Αν η συνάρτηση έχει ως έξοδο διακριτές τιμές, τότε το πρόβλημα ονομάζεται *ταξινόμηση* ή *κατηγοριοποίηση* (classification), ενώ αν έχει συνεχείς τιμές ονομάζεται *παρεμβολή* (regression).
- Η εύρεση φυσικών οργανώσεων των δεδομένων σε ομάδες, έτσι ώστε δεδομένα της ίδιας ομάδας να μοιάζουν όσο το δυνατόν περισσότερο και δεδομένα διαφορετικών ομάδων να διαφέρουν όσο το δυνατόν περισσότερο. Το πρόβλημα αυτό ονομάζεται ομαδοποίηση (clustering).
- Η εύρεση κανόνων συσχέτισης μεταξύ αντικειμένων σε συναλλακτικές (transactional) βάσεις δεδομένων. Το πρόβλημα αυτό ονομάζεται εξόρυξη κανόνων συσχέτισης (association rule mining) και προέκυψε τη δεκαετία του `90 από τον τομέα της ανάλυσης καλαθιών αγορών.
- Η εύρεση της βέλτιστης συμπεριφοράς ενός πράκτορα με βάση την ανταμοιβή που παίρνει σε μια τελική κατάσταση σε κάποιο περιβάλλον έχοντας ξεκινήσει από μια αρχική κατάσταση στο ίδιο περιβάλλον και ακολουθώντας μια σειρά από ενέργειες και ενδιάμεσες καταστάσεις. Το πρόβλημα αυτό ονομάζεται *ενισχυτική μάθηση* (reinforcement learning).

Η εξαγωγή πληροφορίας από τις βάσεις δεδομένων μπορεί να γίνει με δυο τεχνικές:

- την *παραγωγή* (deduction), όπου η πληροφορία που συμπεραίνεται είναι λογικό επακόλουθο της πληροφορίας που είναι αποθηκευμένη στην βάση δεδομένων και
- την *επαγωγή* (induction), όπου έχει μεγαλύτερη αξία γιατί η πληροφορία είναι γενίκευση της πληροφορίας που βρίσκεται στην βάση δεδομένων.

Η τελευταία, η επαγωγική μάθηση και κατ' επέκταση η μηχανική μάθηση διακρίνεται σε:

Μάθηση με επίβλεψη (supervised learning) ή *μάθηση με παραδείγματα* (learning from examples) στην οποία το σύστημα τροφοδοτείται με διάφορα παραδείγματα αντικειμένων που ανήκουν σε μια κατηγορία. Το ίδιο το σύστημα καλείται να ανακαλύψει τις κοινές ιδιότητες των αντικειμένων αυτών. Η μάθηση με επίβλεψη ταυτίζεται με την πρώτη κατηγορία προβλημάτων, δηλαδή της ταξινόμησης και παρεμβολής. Το όνομα προέρχεται από το γεγονός ότι σε αυτά τα προβλήματα υπάρχει κάποιος "επιβλέπων", ο οποίος μας παρέχει την τιμή εξόδου της συνάρτησης για τα δεδομένα που εξετάζουμε [Kotsiantis 2007].

Μάθηση χωρίς επίβλεψη (unsupervised learning) ή *μάθηση από παρατήρηση*, όπου το σύστημα μόνο του, βασισμένο στις δικές του ιδιότητες καλείται να ανακαλύψει κλάσεις/κατηγορίες αντικειμένων. Η μάθηση χωρίς επίβλεψη ταυτιζόταν παραδοσιακά με το πρόβλημα της ομαδοποίησης. Ο λόγος είναι ότι στην ομαδοποίηση δεν υπάρχει κάποιος ``επιβλέπων", αφού δε γνωρίζουμε πόσες, ποιες και αν υπάρχουν ομάδες [Ghahramani 2004].

Η εξόρυξη κανόνων συσχέτισης εμφανίστηκε αρκετά αργότερα από την μηχανική μάθηση, και έχει περισσότερες επιρροές από την ερευνητική περιοχή των βάσεων δεδομένων. Ωστόσο θα μπορούσαμε να την εντάξουμε στη μάθηση χωρίς επίβλεψη, αφού και πάλι δε γνωρίζουμε εκ των προτέρων αν υπάρχουν κάποιες συσχετίσεις στα δεδομένα και ποιες είναι αυτές.

Τέλος, η ενισχυτική μάθηση έχει επιρροές και από τα δύο είδη μάθησης. Όπως στη μάθηση με επίβλεψη, υπάρχει κάποιος εξωτερικός παράγων (το περιβάλλον), που δίνει μια αριθμητική ανταμοιβή στον πράκτορα για κάθε ενέργειά του. Ωστόσο, η συμπεριφορά του περιβάλλοντος είναι άγνωστη στον πράκτορα, και πρέπει να την ανακαλύψει μέσω δοκιμής και αποτυχίας, κάτι που θυμίζει την μάθηση χωρίς επίβλεψη.

Στη μάθηση με επίβλεψη το σύστημα καλείται να "μάθει" μια έννοια ή συνάρτηση από ένα σύνολο δεδομένων, η οποία αποτελεί περιγραφή ενός **μοντέλου (model)**. Αντίθετα, στη μάθηση χωρίς επίβλεψη το σύστημα πρέπει μόνο του να ανακαλύψει συσχετίσεις ή ομάδες σε ένα σύνολο δεδομένων, δημιουργώντας **πρότυπα (patterns)**, χωρίς να είναι γνωστό αν υπάρχουν, πόσα και ποια είναι.

Τα μοντέλα περιγράφουν το σύνολο των δεδομένων και χαρακτηρίζονται και ως μοντέλα πρόβλεψης (predictive models) επειδή προβλέπουν την τιμή μιας μεταβλητής. Εκτός από τις δυνατότητες πρόβλεψης επιπλέον έχουν και κάποιες δυνατότητες πληροφόρησης επειδή δίνουν και ποιοτικές πληροφορίες για τα δεδομένα. Αντίθετα, τα πρότυπα έχουν τοπικό χαρακτήρα, δηλαδή το καθένα περιγράφει ένα μέρος των δεδομένων και χαρακτηρίζονται ως πρότυπα πληροφόρησης (informative patterns) επειδή περιγράφουν συσχετίσεις μεταξύ των δεδομένων. Για παράδειγμα, μοντέλο αποτελεί ένα σύνολο κανόνων ταξινόμησης, ένα δένδρο απόφασης, ένα νευρωνικό δίκτυο, ή μια γραμμική συνάρτηση ενώ πρότυπο αποτελεί ένας κανόνας συσχέτισης, ένας κανόνας ταξινόμησης ή μια ομάδα.

2.6 Μάθηση με Επίβλεψη

Στη μάθηση με επίβλεψη το σύστημα πρέπει να "μάθει" επαγωγικά μια συνάρτηση που ονομάζεται συνάρτηση στόχος (target function) και αποτελεί έκφραση του μοντέλου που περιγράφει τα δεδομένα. Η συνάρτηση στόχος (συμβολίζεται συνήθως με *c*) χρησιμοποιείται για την πρόβλεψη της τιμής μιας μεταβλητής, που ονομάζεται εξαρτημένη μεταβλητή ή μεταβλητή εξόδου, βάσει των τιμών ενός συνόλου μεταβλητών, που ονομάζονται ανεξάρτητες μεταβλητές ή μεταβλητές εισόδου ή χαρακτηριστικά.

Το σύνολο των διαφορετικών δυνατών τιμών εισόδου της συνάρτησης, δηλαδή το πεδίο ορισμού της, ονομάζεται σύνολο των περιπτώσεων ή στιγμιότυπων (instances) και συμβολίζεται με *X*. Κάθε περίπτωση (ή στιγμιότυπο) περιγράφεται από ένα σύνολο χαρακτηριστικών (attributes ή features). Ένα υποσύνολο του συνόλου των περιπτώσεων για τα οποία γνωρίζουμε την τιμή της μεταβλητής εξόδου, ονομάζεται σύνολο δεδομένων εκπαίδευσης ή παραδείγματα και συμβολίζεται με *D*.

Για να προσεγγίσει το σύστημα όσο το δυνατόν καλύτερα τη συνάρτηση στόχο εξετάζει διάφορες εναλλακτικές συναρτήσεις οι οποίες ονομάζονται *υποθέσεις* και συμβολίζονται με h. Το σύνολο όλων των δυνατών υποθέσεων που το πρόγραμμα μάθησης πρέπει να εξετάσει προκειμένου να βρει τη συνάρτηση στόχο ονομάζεται σύνολο υποθέσεων και συμβολίζεται με *H*. Κάθε υπόθεση $h \in H$, αναπαριστά είτε μια λογική συνάρτηση $h: X \rightarrow \{0, 1\}$ ή μια πραγματική συνάρτηση $h: X \rightarrow R$.

Η επαγωγική μάθηση στηρίζεται στην υπόθεση επαγωγικής μάθησης (inductive learning hypothesis), σύμφωνα με την οποία κάθε υπόθεση *h* που έχει βρεθεί να προσεγγίζει καλά τη συνάρτηση στόχο για ένα αρκετά μεγάλο σύνολο παραδειγμάτων, θα προσεγγίζει το ίδιο καλά τη συνάρτηση στόχο και για περιπτώσεις που δεν έχει εξετάσει.

Στην μάθηση με επίβλεψη διακρίνονται δυο είδη προβλημάτων (learning tasks), τα προβλήματα ταξινόμησης και τα προβλήματα παρεμβολής. Η ***ταξινόμηση*** (***classification***) αφορά στη δημιουργία μοντέλων πρόβλεψης διακριτών τάξεων (κλάσεων/κατηγοριών) όπως για παράδειγμα η ομάδα αίματος, ενώ η ***παρεμβολή*** (***regression***) αφορά στη δημιουργία μοντέλων πρόβλεψης αριθμητικών τιμών. Οι κυριότερες τεχνικές μηχανικής μάθησης με επίβλεψη είναι:

Δένδρα ταξινόμησης ή απόφασης (Classification or Decision Trees)

Μάθηση Κανόνων (Rule Learning)

Μάθηση κατά Περίπτωση (Instance Based Learning)

Μάθηση κατά Bayes

Γραμμική παρεμβολή (Linear Regression)

Νευρωνικά Δίκτυα (Neural Networks)

Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines, SVMs)

2.6.1 Δένδρα Ταξινόμησης/Απόφασης

Οι αλγόριθμοι μάθησης ή επαγωγής δέντρων ταξινόμησης/απόφασης (classification/decision trees) είναι από τους πιο δημοφιλείς αλγόριθμους μάθησης και έχουν εφαρμοστεί αποτελεσματικά σε διάφορους τομείς, όπως διάγνωση ιατρικών περιστατικών, αξιολόγηση ρίσκου αποδοχής αίτησης για πιστωτική κάρτα, πρόβλεψη συμπεριφοράς καταναλωτή, κτλ. Είναι μια μέθοδος για την προσέγγιση συναρτήσεων στόχος, που έχουν ως έξοδο διακριτές τιμές. Το αποτέλεσμά τους είναι μία δενδροειδής δομή που με γραφικό τρόπο περιγράφει τα δεδομένα και εναλλακτικά, για τη βελτίωση της αναγνωσιμότητάς του, μπορεί να αναπαρασταθεί και ως σύνολο κανόνων *if-then*, που ονομάζονται *κανόνες ταξινόμησης* (*classification rules*).

Κάθε κόμβος στο δένδρο ορίζει μια συνθήκη ελέγχου της τιμής κάποιου χαρακτηριστικού (attribute ή feature) των περιπτώσεων (instances) και κάθε κλαδί που φεύγει από τον κόμβο αυτό αντιστοιχεί σε μια διαφορετική διακριτή τιμή του χαρακτηριστικού αυτού. Μια περίπτωση ταξινομείται αρχίζοντας από την ρίζα και ακολουθώντας τα κλαδιά του δένδρου προς κάποιο φύλλο, το οποίο περιέχει και μια διακριτή τιμή της κατηγορίας. Σε κάθε κόμβο ελέγχεται η τιμή της περίπτωσης για το χαρακτηριστικό του κόμβου και ακολουθείται το αντίστοιχο κλαδί.

Τα δένδρα ταξινόμησης χρησιμοποιούνται για να προβλέψουν με κάποιο βαθμό ακρίβειας την τιμή της μεταβλητής που μοντελοποιούν με βάση τις τιμές των θεωρούμενων ανεξάρτητων μεταβλητών (χαρακτηριστικών). Ένα σημαντικό πλεονέκτημα τους είναι η ευκολία με την οποία ερμηνεύονται.

Οι περισσότεροι αλγόριθμοι που έχουν αναπτυχθεί για μάθηση δένδρων ταξινόμησης είναι παραλλαγές ενός βασικού αλγορίθμου. Παράδειγμα του βασικού αυτού αλγορίθμου αποτελούν ο αλγόριθμος ID3 και ο απόγονος του C4.5.

Ο αλγόριθμος ID3

Είναι ο πιο γνωστός αλγόριθμος μάθησης δένδρων ταξινόμησης. Είναι αναδρομικός και στη γενική του μορφή περιγράφεται ως εξής:

1. **Βρες την ανεξάρτητη μεταβλητή η οποία αν χρησιμοποιηθεί ως κριτήριο διαχωρισμού των δεδομένων εκπαίδευσης θα οδηγήσει σε κόμβους κατά το δυνατό διαφορετικούς σε σχέση με την εξαρτημένη μεταβλητή.**
2. **Κάνε το διαχωρισμό.**
3. **Επανέλαβε τη διαδικασία για κάθε έναν από τους κόμβους που προέκυψαν μέχρι να μην είναι δυνατός περαιτέρω διαχωρισμός.**

Δηλαδή ο αλγόριθμος ID3 κατασκευάζει το δένδρο άπληστα (greedy) από πάνω προς τα κάτω επιλέγοντας αρχικά το πιο κατάλληλο χαρακτηριστικό για έλεγχο στη ρίζα. Η επιλογή βασίζεται σε κάποιο στατιστικό μέτρο που υπολογίζεται από τα δεδομένα. Στη συνέχεια, για κάθε δυνατή τιμή του χαρακτηριστικού δημιουργούνται οι αντίστοιχοι απόγονοι της ρίζας και τα δεδομένα μοιράζονται στους νέους κόμβους ανάλογα με την τιμή που έχουν για το χαρακτηριστικό που ελέγχεται στη ρίζα.

Η όλη διαδικασία επαναλαμβάνεται για κάθε νέο κόμβο. Η επιλογή όμως του κατάλληλου χαρακτηριστικού σε κάθε νέο κόμβο αποφασίζεται χρησιμοποιώντας

μόνο τα δεδομένα που ανήκουν σε αυτόν τον κόμβο. Η διαδικασία τερματίζει όταν οι κόμβοι γίνουν τερματικοί (ή φύλλα). Ένας κόμβος γίνεται τερματικός όταν:

- Όλα τα δεδομένα που ανήκουν σε αυτόν ανήκουν στην ίδια κατηγορία η οποία γίνεται και η τιμή του κόμβου. Ο κόμβος ονομάζεται *αμιγής κόμβος* (*pure node*).
- Σε κάποιο βάθος τελειώσουν τα χαρακτηριστικά προς έλεγχο. Η τιμή του κόμβου είναι η κατηγορία στην οποία ανήκει η πλειοψηφία των δεδομένων του κόμβου.

Το βασικότερο στάδιο του αλγορίθμου είναι η επιλογή της ανεξάρτητης μεταβλητής πάνω στην οποία θα συνεχιστεί η ανάπτυξη του δένδρου. Το σημείο αυτό απαιτεί ουσιαστικά τον ορισμό κάποιου μηχανισμού ο οποίος θα καθοδηγήσει την αναζήτηση προς το καλύτερο δένδρο (περιγραφή) μέσα στο σύνολο των δυνατών δένδρων.

Όσον αφορά τα δεδομένα εκπαίδευσης, ο ID3 δεν περιορίζει τον αριθμό των τιμών που μπορούν να πάρουν οι μεταβλητές, απαιτεί όμως οι τιμές τους να είναι διακριτές και όχι συνεχείς. Στη δεύτερη περίπτωση απαιτείται ο ορισμός διαστημάτων για τη μετατροπή των συνεχών τιμών σε διακριτές (διακριτοποίηση).

Έχουν προταθεί αρκετές παραλλαγές του αλγόριθμου ID3 και περιλαμβάνουν τεχνικές κλαδέματος πριν την ολοκλήρωση της κατασκευής του δένδρου, διαχείριση πεδίων χωρίς τιμή, χρήση διαφόρων κριτηρίων διαχωρισμού, αυτόματη διαχείριση συνεχόμενων αριθμητικών τιμών, κτλ. Ο αλγόριθμος C4.5 αποτελεί μία από τις περισσότερο διαδεδομένες βελτιώσεις του ID3 [Kotsiantis 2007], [Mitchell 1997], [Quinlan 1993], [Quinlan 1996].

Εντροπία και κέρδος πληροφορίας

Ένας από τους πιο διαδεδομένους μηχανισμούς διαχωρισμού είναι αυτός της εντροπίας της πληροφορίας (*information entropy*) ο οποίος επιλέγει εκείνη την ανεξάρτητη μεταβλητή που οδηγεί σε περισσότερο συμπαγές δένδρο. Η τιμή της εντροπίας της πληροφορίας δίνεται από τη σχέση:

$$E(S) = -p_+ \cdot \log_2(p_+) - p_- \cdot \log_2(p_-)$$

όπου S είναι το σύνολο των δεδομένων εκπαίδευσης στο στάδιο (κόμβο) του διαχωρισμού, $p+$ είναι το κλάσμα των θετικών παραδειγμάτων του S και $p-$ είναι το κλάσμα των αρνητικών παραδειγμάτων του S.

Γενικότερα, για c διαφορετικές κατηγορίες, η εντροπία ορίζεται από τη σχέση:

$$E(S) = -\sum_{i=1}^{c} p_i \cdot \log_2(p_i)$$

όπου pi το ποσοστό των παραδειγμάτων του S που ανήκουν στην κατηγορία i.

Η εντροπία της πληροφορίας μετρά ουσιαστικά την ανομοιογένεια που υπάρχει στο S αναφορικά με την υπό εξέταση εξαρτημένη μεταβλητή και έχει τις ρίζες της στη θεωρία πληροφοριών (information theory). Στην περίπτωση που έχουμε δυο κατηγορίες, η τιμή της είναι 0 αν όλα τα μέλη του S ανήκουν στην ίδια κατηγορία και 1 αν τα μισά μέλη ανήκουν στη μια και τα άλλα μισά στην άλλη κατηγορία. Σε όλους δε τους υπολογισμούς, θεωρούμε την ποσότητα $0 \cdot \log_2(0)$ ίση με μηδέν.

Στην πράξη, χρησιμοποιείται το κέρδος πληροφορίας (information gain), *Gain*(*S*,*A*) ή *G*(*S*,*A*) που αναπαριστά τη μείωση της εντροπίας του συνόλου εκπαίδευσης *S* αν επιλεγεί ως παράμετρος διαχωρισμού η μεταβλητή *A*. Όταν μειώνεται η πληροφοριακή εντροπία, αυξάνεται η πυκνότητα πληροφορίας και άρα η περιγραφή γίνεται περισσότερο συμπαγής. Το κέρδος πληροφορίας δίνεται από τη σχέση:

$$G(S,A) = E(S) - \sum_{u \in Values(A)} \frac{|S_u|}{|S|} \cdot E(S_u)$$

όπου $E(S)$ είναι η εντροπία πληροφορίας του υπό εξέταση κόμβου, A είναι η ανεξάρτητη μεταβλητή, με τιμές $Values(A)$, βάσει της οποίας επιχειρείται ο επόμενος διαχωρισμός, u είναι μία από τις δυνατές τιμές του A, S_u είναι το πλήθος των εγγραφών με $A=u$ και $E(S_u)$ η εντροπία πληροφορίας του υπό εξέταση κόμβου ως προς την τιμή $A=u$. Ουσιαστικά, ο δεύτερος όρος είναι η εντροπία των παραδειγμάτων μετά το διαχωρισμό τους σύμφωνα με την τιμή του χαρακτηριστικού Α και αποτελείται από το άθροισμα της εντροπίας για το κάθε σύνολο που προκύπτει μετά το διαχωρισμό.

2.6.2 Μάθηση κατά Περίπτωση

Σε αντίθεση με τις μεθόδους μηχανικής μάθησης που αναφέρθηκαν ως τώρα και οι οποίες κωδικοποιούν τα παραδείγματα εκπαίδευσης σε μια συμπαγή περιγραφή, στη μάθηση κατά περίπτωση (instance-based learning) τα δεδομένα εκπαίδευσης διατηρούνται αυτούσια. Όταν ένα τέτοιο σύστημα κληθεί να αποφασίσει για την κατηγορία μιας νέας περίπτωσης, εξετάζει εκείνη τη στιγμή τη σχέση της με τα ήδη αποθηκευμένα παραδείγματα. Δηλαδή η μέθοδος αυτή αναβάλλει τη μάθηση έως ότου εμφανιστεί μια νέα περίπτωση (στιγμιότυπο) και για το λόγο αυτό ονομάζεται αναβλητική μάθηση (lazy learning) σε αντίθεση με τις άλλες οι οποίες μπορεί να χαρακτηριστούν ως *έγκαιρες* μέθοδοι μάθησης (eager learners), αφού μαθαίνουν το μοντέλο από τα αποθηκευμένα παραδείγματα του συνόλου εκπαίδευσης, χωρίς να περιμένουν την άφιξη της νέας περίπτωσης.

Χαρακτηριστικός αλγόριθμος αυτής της κατηγορίας είναι ο αλγόριθμος των ***k-κοντινότερων γειτόνων* (*k-Nearest Neighbors*)**, στον οποίο γίνεται η παραδοχή ότι τα διάφορα παραδείγματα μπορεί να αναπαρασταθούν ως σημεία σε κάποιον *n*-διάστατο Ευκλείδειο χώρο *Rn* όπου *n* ο αριθμός των χαρακτηριστικών (ανεξάρτητων μεταβλητών). Κάθε νέα περίπτωση τοποθετείται στο χώρο αυτό ως νέο σημείο και η τιμή του προσδιορίζεται με βάση το χαρακτηρισμό των *k* γειτονικών σημείων. Οι κοντινότεροι γείτονες μιας περίπτωσης υπολογίζονται με βάση τη γνωστή από τη γεωμετρία Ευκλείδεια απόστασή τους.

Για παράδειγμα, η απόσταση μιας νέας περίπτωσης x' που περιγράφεται από το σύνολο χαρακτηριστικών $<a_1(x'), a_2(x'), a_3(x'), ..., a_n(x')>$, όπου $ar(x')$ είναι το r χαρακτηριστικό της και ενός αποθηκευμένου παραδείγματος x, που περιγράφεται από το σύνολο χαρακτηριστικών $<a_1(x), a_2(x), a_3(x), ..., a_n(x), y(x)>$, είναι το άθροισμα των Ευκλείδειων αποστάσεων όλων των χαρακτηριστικών των 2 σημείων. Δηλαδή:

$$d(x,x') = \sqrt{\sum_{r=1}^{n} (a_r(x) - a_r(x'))^2}$$

2.6.3 Μάθηση κατά Bayes

Η συλλογιστική κατά Bayes μπορεί να συνεισφέρει στο πρόβλημα της μηχανικής μάθησης γιατί παρέχει μια ποσοτική μεθοδολογία για την αξιολόγηση των διαφόρων ενδείξεων που υποστηρίζουν τις εναλλακτικές υποθέσεις, οι οποίες διερευνώνται κατά τη μάθηση. Αποτελεί τη θεωρητική βάση για αλγορίθμους μάθησης που διαχειρίζονται πιθανότητες αλλά ακόμη και σε περιπτώσεις που η υπολογιστική πολυπλοκότητα της μεθόδου καθιστά απαγορευτική τη χρήση της, μπορεί να χρησιμοποιηθεί ως κριτήριο για τον έλεγχο της απόδοσης άλλων αλγορίθμων που δε διαχειρίζονται πιθανότητες.

Στη *μάθηση κατά Bayes* (*Bayesian learning*) κάθε παράδειγμα εκπαίδευσης μπορεί σταδιακά να μειώσει ή να αυξήσει την πιθανότητα να είναι σωστή μια υπόθεση. Αυτό δίνει μεγάλη ευελιξία στους σχετικούς αλγορίθμους καθώς δεν απορρίπτουν αμέσως μια υπόθεση όταν προκύπτει ότι δεν είναι σε απόλυτη συμφωνία με τα παραδείγματα εκπαίδευσης. Επιπλέον, προϋπάρχουσα γνώση μπορεί να συνδυαστεί με τα δεδομένα εκπαίδευσης με τη μορφή αρχικών τιμών πιθανότητας για τις υπό εξέταση υποθέσεις.

Μια πρακτική δυσκολία στην εφαρμογή της μάθησης κατά Bayes είναι η απαίτηση για τη γνώση πολλών τιμών πιθανοτήτων. Όταν αυτές οι τιμές δεν είναι δυνατό να υπολογιστούν επακριβώς, υπολογίζονται κατ' εκτίμηση από παλαιότερες υποθέσεις, εμπειρική γνώση, κτλ. Η παραπάνω δυσκολία εφαρμογής έχει δώσει μεγάλη πρακτική αξία σε μια απλουστευμένη εκδοχή της μάθησης κατά Bayes, τον απλό ταξινομητή Bayes, στον οποίο γίνεται η παραδοχή ότι τα χαρακτηριστικά είναι ανεξάρτητα μεταξύ τους.

Απλός ταξινομητής Bayes

Ο απλός ταξινομητής Bayes (simple/naive Bayes classifier) είναι μια πρακτική μέθοδος μάθησης που στηρίζεται σε στατιστικά στοιχεία (κατανομές πιθανότητας). Η ποσότητα P που περιγράφει έναν απλό ταξινομητή Bayes για ένα σύνολο παραδειγμάτων εκφράζει την πιθανότητα να είναι c η τιμή της εξαρτημένης μεταβλητής C με βάση τις τιμές $x=(x_1, x_2, ..., x_n)$ των χαρακτηριστικών $X=(X_1, X_2, ..., X_n)$ και δίνεται από τη σχέση:

$$P(c \mid x) = P(c) \cdot \prod_i P(x_i \mid c)$$

όπου τα χαρακτηριστικά Xi θεωρούνται ανεξάρτητα μεταξύ τους. Ο υπολογισμός της παραπάνω ποσότητας για ένα σύνολο N παραδειγμάτων γίνεται με βάση τις σχέσεις:

- $P(c) = N(c) / N$,
- $P(x_i|c) = N(x_i,c) / N(c)$, για χαρακτηριστικό *Xi* με διακριτές τιμές,
- $P(x_i|c) = g(x_i, \mu_c, \sigma_c^2)$, για χαρακτηριστικό *Xi* με αριθμητικές τιμές,

όπου $N(c)$ είναι ο αριθμός των παραδειγμάτων που έχουν στην εξαρτημένη μεταβλητή την τιμή c, $N(x_i,c)$ είναι ο αριθμός των παραδειγμάτων που έχουν για το

χαρακτηριστικό X_i και την εξαρτημένη μεταβλητή, τιμές x_i και c αντίστοιχα, και $g(x_i, \mu_c,\sigma_c^2)$ είναι η συνάρτηση πυκνότητας πιθανότητας Gauss με μέσο όρο μ_c και διασπορά σ_c για το χαρακτηριστικό X_i.

Επειδή η ποσότητα $P(c|x)$ σύμφωνα με την παραπάνω σχέση υπολογίζεται με εκτίμηση από τα δεδομένα, γίνεται κανονικοποίηση αυτών των τιμών ώστε να δίνουν άθροισμα 1, δηλαδή:

$$\sum_c P(c|x)=1$$

Στην περίπτωση που τα παραδείγματα είναι οργανωμένα σε βάση δεδομένων ένα πλεονέκτημα της μεθόδου είναι ότι όλες οι ποσότητες μπορεί να υπολογιστούν με χρήση ερωτημάτων (queries) προς τη βάση δεδομένων, δηλαδή η μέθοδος είναι άμεσα υλοποιήσιμη σε οποιοδήποτε σύγχρονο σύστημα διαχείρισης βάσεων δεδομένων. Το βασικότερο μειονέκτημά της είναι ότι δεν μπορεί να εντοπίσει μοντέλα που βασίζονται σε αλληλεπίδραση δύο ή περισσοτέρων χαρακτηριστικών (πεδίων), διότι βασίζεται στην ακριβώς αντίθετη παραδοχή.

2.6.4 Άλλες Τεχνικές Μάθησης με Επίβλεψη

Εκτός από τις τεχνικές μηχανικής μάθησης με επίβλεψη που παρουσιάστηκαν στις προηγούμενες ενότητες, υπάρχουν και άλλες, οι σημαντικότερες από τις οποίες είναι η *παρεμβολή*, τα *νευρωνικά δίκτυα* και οι *μηχανές διανυσμάτων υποστήριξης*.

Παρεμβολή

Παρεμβολή ή *παλινδρόμηση* (*regression*) είναι η διαδικασία προσδιορισμού της σχέσης μιας μεταβλητής y (εξαρτημένη μεταβλητή ή έξοδος) με μια ή περισσότερες άλλες μεταβλητές x_1, x_2, ..., x_n (ανεξάρτητες μεταβλητές ή είσοδοι). Σκοπός της παρεμβολής είναι η πρόβλεψη της τιμής της εξόδου όταν είναι γνωστές οι είσοδοι.

Το πιο διαδεδομένο μοντέλο παρεμβολής είναι το *γραμμικό* (*linear regression*) που ονομάζεται έτσι επειδή η αναμενόμενη τιμή της εξόδου μοντελοποιείται με μία γραμμική συνάρτηση ή σταθμισμένο άθροισμα (weighted sum) των παραμέτρων εισόδου. Συνήθως, γράφεται ως:

$$y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + ... + \beta_n x_{nj} \qquad j = 1, 2, ..., m$$

όπου m είναι ο αριθμός των δεδομένων (παραδειγμάτων) εκπαίδευσης, ενώ το ζητούμενο είναι να υπολογιστούν οι συντελεστές β_i. Μια από τις πιο διαδεδομένες μεθόδους επίλυσης είναι η μέθοδος των *ελαχίστων τετραγώνων* (*least squares*) που ελαχιστοποιεί το σφάλμα μεταξύ της εκτιμώμενης συνάρτησης και των πραγματικών δεδομένων.

Στα μη γραμμικά μοντέλα παρεμβολής η αναμενόμενη τιμή εξόδου δεν είναι σταθμισμένο άθροισμα των παραμέτρων εισόδου αλλά συνδέεται με αυτά με πιο πολύπλοκο τρόπο, όπως για παράδειγμα:

$$y_j = \beta_0 x_{1j}^{\beta 1} \qquad j=1, 2, ..., m$$

Σε κάποιες περιπτώσεις, τα μη γραμμικά μοντέλα μπορεί να μετατραπούν σε γραμμικά με κατάλληλο μετασχηματισμό των μεταβλητών, ώστε τελικά να επιλυθούν με τη μέθοδο των ελαχίστων τετραγώνων.

Νευρωνικά δίκτυα

Τα *τεχνητά νευρωνικά δίκτυα* ή απλά νευρωνικά δίκτυα (neural networks) παρέχουν ένα πρακτικό (εύκολο) τρόπο για την εκμάθηση αριθμητικών και διανυσματικών συναρτήσεων ορισμένων σε συνεχή ή διακριτά μεγέθη. Χρησιμοποιούνται τόσο για παρεμβολή (γραμμική και μη γραμμική) όσο και για ταξινόμηση και έχουν το μεγάλο πλεονέκτημα της ανοχής που παρουσιάζουν σε δεδομένα εκπαίδευσης με θόρυβο, δηλαδή δεδομένα που περιστασιακά έχουν λανθασμένες τιμές (π.χ. λάθη καταχώρησης). Από την άλλη όμως αδυνατούν να εξηγήσουν ποιοτικά τη γνώση που μοντελοποιούν.

Υπάρχει μια ειδική κατηγορία νευρωνικών δικτύων, τα δίκτυα με ανατροφοδότηση τα οποία λόγω της ειδικής τοπολογίας τους έχουν τη δυνατότητα αυτο-οργάνωσης χωρίς εξωτερική καθοδήγηση και ανήκουν στην κατηγορία μάθησης χωρίς επίβλεψη.

Μηχανές διανυσμάτων υποστήριξης (ή εδραίων διανυσμάτων)

Οι *μηχανές διανυσμάτων υποστήριξης* ή *ΜΔΥ* (*Support Vector Machines, SVMs*) προτάθηκαν από τον Vladimir Vapnik και τους συνεργάτες του το 1992. Στηρίζονται στη *Θεωρία Στατιστικής Μάθησης* (*Statistical Learning Theory*) και στα νευρωνικά δίκτυα τύπου *Perceptron*.

Τα τελευταία χρόνια, οι ΜΔΥ έχουν εδραιωθεί ως μια από τις πιο διαδεδομένες μεθόδους (γραμμικής και μη) παρεμβολής και ταξινόμησης, αποτελώντας συνήθως την βέλτιστη επιλογή για εφαρμογές όπως η αναγνώριση γραφής (handwriting recognition), η ταξινόμηση κειμένων (text categorization) και η ταξινόμηση δεδομένων έκφρασης γονιδίων (gene expression data).

Στην περίπτωση της ταξινόμησης, οι ΜΔΥ προσπαθούν να βρουν μια υπερεπιφάνεια (hypersurface) που να διαχωρίζει στο χώρο των παραδειγμάτων τα αρνητικά από τα θετικά παραδείγματα. Η υπερεπιφάνεια αυτή επιλέγεται έτσι, ώστε να απέχει όσο το δυνατόν περισσότερο από τα κοντινότερα θετικά και αρνητικά παραδείγματα (maximum margin hypersurface). Έτσι, μια ΜΔΥ μπορεί και ταξινομεί περιπτώσεις που είναι παρόμοιες αλλά όχι πανομοιότυπες με κάποιο παράδειγμα εκπαίδευσης. Το αποτέλεσμα μιας ΜΔΥ είναι τελικά μια αριθμητική τιμή στο διάστημα [-1, +1] και όχι μια πιθανότητα όπως σε άλλους ταξινομητές.

Το βασικό πλεονέκτημα των ΜΔΥ έναντι των νευρωνικών δικτύων τύπου *Perceptron* είναι ότι μπορούν και παράγουν πιο σύνθετες υπερεπιφάνειες, ενσωματώνοντας μετασχηματισμούς και συνδυασμούς των αρχικών μεταβλητών ανάλογα με το πρόβλημα και ξεπερνώντας προβλήματα όπως τα τοπικά ελάχιστα και η διασπορά των λύσεων στο χώρο αναζήτησης. Για το σκοπό αυτό, χρησιμοποιούν έναν πεπερασμένο αριθμό υποσυνόλων του συνόλου εκπαίδευσης, που ονομάζονται *διανύσματα υποστήριξης* (*support vectors*) καθώς και *συναρτήσεις πυρήνα* (*kernel functions*), προκειμένου να μετασχηματίσουν τον αρχικό χώρο υποθέσεων ώστε να βρουν τη βέλτιστη μη γραμμική υπερεπιφάνεια που ελαχιστοποιεί το σφάλμα ταξινόμησης.

2.7 Μάθηση Χωρίς Επίβλεψη

Στη μάθηση χωρίς επίβλεψη το σύστημα έχει στόχο να ανακαλύψει συσχετίσεις και ομάδες από τα δεδομένα, βασιζόμενο μόνο στις ιδιότητές τους. Σαν αποτέλεσμα προκύπτουν πρότυπα (περιγραφές), κάθε ένα από τα οποία περιγράφει ένα μέρος από τα δεδομένα. Παραδείγματα προτύπων πληροφόρησης είναι οι *κανόνες συσχέτισης* (*association rules*) και οι *ομάδες* (*clusters*), οι οποίες προκύπτουν από τη διαδικασία της *ομαδοποίησης* (*clustering*).

2.7.1 Κανόνες Συσχέτισης

Η ανακάλυψη ή εξόρυξη *κανόνων συσχέτισης* (*association rule mining*) εμφανίστηκε αρκετά αργότερα από τη μηχανική μάθηση και έχει περισσότερες επιρροές από την ερευνητική περιοχή των βάσεων δεδομένων. Προτάθηκε στις αρχές της δεκαετίας του '90 από τον Rakesh Agrawal ως τεχνική *ανάλυσης καλαθιού αγορών* (*market basket analysis*) όπου το ζητούμενο είναι η ανακάλυψη συσχετίσεων ανάμεσα στα αντικείμενα μιας βάσης δεδομένων.

Στο συγκεκριμένο πρόβλημα υπάρχει ένας μεγάλος αριθμός *αντικειμένων* (items), για παράδειγμα ψωμί, γάλα, κτλ. Οι πελάτες γεμίζουν τα καλάθια τους με κάποιο υποσύνολο αυτών των αντικειμένων και το ζητούμενο είναι να βρεθεί ποια από αυτά τα αντικείμενα αγοράζονται μαζί, χωρίς να ενδιαφέρει ποιος είναι ο αγοραστής.

Οι κανόνες συσχέτισης είναι προτάσεις της μορφής $\{X_1,...,X_n\} \rightarrow Y$, που σημαίνει ότι αν βρεθούν όλα τα $X_1, ..., X_n$ στο καλάθι (στην ανάλυση καλαθιού αγορών) τότε είναι πιθανό να βρεθεί και το Y. Για παράδειγμα, ένας τέτοιος κανόνας θα μπορούσε να λέει:

"*όποιος αγοράζει καφέ* (X_1) *και ζάχαρη* (X_2) *αγοράζει και αναψυκτικά* (Y)"

Η απλή αναφορά ενός τέτοιου κανόνα δεν έχει και μεγάλη αξία αν δε συνοδεύεται από κάποια ποσοτικά μεγέθη που μετρούν την ποιότητα των ευρεθέντων κανόνων συσχέτισης. Τέτοια μεγέθη είναι η *υποστήριξη* (*support*) και η *εμπιστοσύνη* (*confidence*) που ορίζονται ως εξής:

- *Υποστήριξη* (*support*) ή *κάλυψη* (*coverage*): εκφράζει την πιθανότητα να βρεθεί το καλάθι $\{X_1, ..., X_n, Y\}$ στη βάση δεδομένων και ισούται με το λόγο των εγγραφών που περιλαμβάνουν το $\{X_1,...,X_n, Y\}$ προς το σύνολο των εγγραφών.
- *Εμπιστοσύνη* (*confidence*) ή *ακρίβεια* (*accuracy*): εκφράζει την πιθανότητα να βρεθεί το Y σε ένα καλάθι που περιέχει τα $\{X_1, ..., X_n\}$ και ισούται με το λόγο των εγγραφών που περιλαμβάνουν το $\{X_1, ..., X_n, Y\}$ προς το σύνολο των εγγραφών που περιλαμβάνουν τα X_i.

Αν σε έναν κανόνα $X \rightarrow Y$, όπου X και Y σύνολα αντικειμένων, $S(X)$ είναι η υποστήριξη του X και $S(X \cup Y)$ είναι η υποστήριξη του συνόλου $\{X, Y\}$, η εμπιστοσύνη C του κανόνα σε σχέση με την υποστήριξη είναι:

$$C(X \rightarrow Y) = \frac{S(X \cup Y)}{S(X)}$$

Η μεγαλύτερη δυσκολία στην ανακάλυψη κανόνων συσχέτισης είναι ο μεγάλος αριθμός τέτοιων κανόνων που θεωρητικά υπάρχουν σε μία βάση δεδομένων και η επιλογή εκείνων που έχουν πρακτική αξία. Αυτό συνήθως γίνεται θέτοντας κάποιο κάτω όριο στις τιμές των μεγεθών *εμπιστοσύνη* και *υποστήριξη*.

Αλγόριθμοι εύρεσης κανόνων συσχέτισης

Για την ανακάλυψη κανόνων συσχέτισης χρησιμοποιείται η ιδιότητα της μονοτονίας (monotonicity property) ή αλλιώς ιδιότητα *a priori* σύμφωνα με την οποία: "*Αν ένα σύνολο αντικειμένων S είναι συχνό, τότε όλα τα υποσύνολα του S είναι επίσης συχνά*". Συχνό είναι ένα σύνολο αντικειμένων όταν εμφανίζεται σε ποσοστό των καλαθιών ίσο ή μεγαλύτερο από ένα όριο που συνήθως ορίζει ο χρήστης.

Σε έναν αλγόριθμο εύρεσης κανόνων συσχέτισης μας ενδιαφέρει κυρίως ο αριθμός των περασμάτων στα δεδομένα που απαιτείται κατά την εκτέλεσή του. Θεωρώντας ότι τα δεδομένα δε χωράνε στην κύρια μνήμη, το σημαντικότερο κόστος είναι ο αριθμός των προσπελάσεων στο δίσκο. Υπάρχουν δύο βασικές μέθοδοι (οικογένειες αλγορίθμων) για την ανακάλυψη κανόνων συσχέτισης.

Στην πρώτη μέθοδο, βρίσκονται αρχικά όλα τα συχνά αντικείμενα (σύνολα με μέγεθος 1), κατόπιν τα συχνά ζεύγη (σύνολα με μέγεθος 2), οι συχνές τριάδες, κτλ., μέχρι να βρεθούν τα μέγιστα συχνά σύνολα αντικειμένων (maximal frequent itemsets) δηλαδή τα σύνολα S εκείνα των οποίων κανένα υπερσύνολο δεν είναι συχνό. Η εύρεση συχνών ζευγών είναι σημαντικό στάδιο του αλγορίθμου γιατί σε πολλά σύνολα δεδομένων αποτελεί το δυσκολότερο κομμάτι, ενώ στη συνέχεια η ανακάλυψη τριάδων, τετράδων, κτλ., θέλει λιγότερο χρόνο. Οι αλγόριθμοι αυτού του είδους απαιτούν ένα πέρασμα στα δεδομένα για κάθε επίπεδο. Αντιπροσωπευτικό παράδειγμα είναι ο αλγόριθμος Apriori

Στη δεύτερη μέθοδο, βρίσκονται όλα τα μέγιστα συχνά σύνολα αντικειμένων, με ένα ή ελάχιστα περάσματα. Αυτές οι μέθοδοι έχουν συνήθως μεγάλες απαιτήσεις σε μνήμη λόγω του ότι βασίζονται σε πολύπλοκες (συνήθως δενδροειδείς) δομές δεδομένων για να αποθηκεύσουν πληροφορίες για τα δεδομένα. Αντιπροσωπευτικό παράδειγμα είναι ο αλγόριθμος FP-Growth.

Ο αλγόριθμος Apriori

Ο αλγόριθμος *Apriori* προτάθηκε από τον Rakesh Agrawal το 1994 και είναι ίσως ο κλασικότερος αλγόριθμος ανακάλυψης κανόνων συσχέτισης. Περιλαμβάνει δυο βασικά βήματα, τη *δημιουργία των συχνών συνόλων αντικειμένων* και τη *δημιουργία των κανόνων συσχέτισης* που περιγράφονται στη συνέχεια.

Η διαδικασία της δημιουργίας συχνών συνόλων αντικειμένων περιλαμβάνει δύο στάδια: αρχικά δημιουργείται ένα σύνολο υποψήφιων συχνών αντικειμένων C_i και στη συνέχεια, χρησιμοποιώντας το όριο υποστήριξης (support), δημιουργείται το σύνολο των συχνών συνόλων αντικειμένων L_i. Η διαδικασία επαναλαμβάνεται πραγματοποιώντας διαδοχικά περάσματα στα δεδομένα μέχρι να βρεθούν είτε τα συχνά σύνολα αντικειμένων ενός προκαθορισμένου επιπέδου ή τα *μέγιστα συχνά σύνολα αντικειμένων*. Το πρώτο στάδιο επιπλέον αποτελείται από ένα *βήμα*

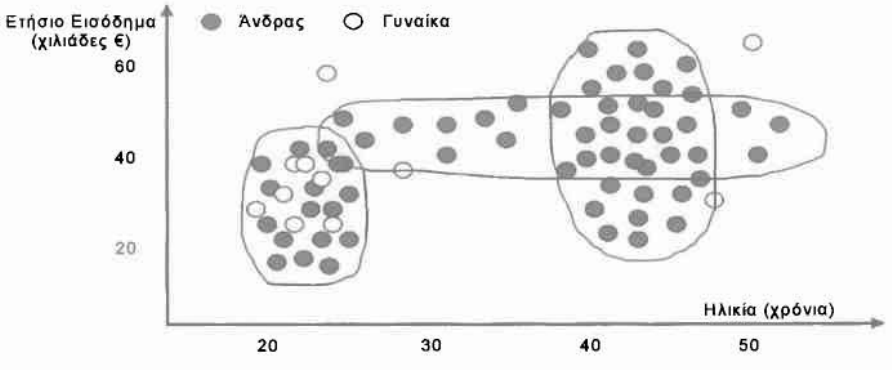
συνένωσης (*join step*) και ένα *βήμα κλαδέματος* (*prune step*) τα οποία συνήθως εκτελούνται στη μνήμη και έτσι δεν είναι ιδιαίτερα χρονοβόρα.

Για τη δημιουργία των κανόνων συσχέτισης ελέγχεται η εμπιστοσύνη (confidence) όλων των πιθανών κανόνων που προκύπτουν από τα μέγιστα συχνά σύνολα αντικειμένων και στο τέλος μένουν εκείνοι των οποίων εμπιστοσύνη ξεπερνά το όριο που τέθηκε από το χρήστη.

2.7.2 Ομάδες

Οι *ομάδες* (*clusters*) είναι πρότυπα πληροφόρησης που προκύπτουν με ομαδοποίηση (clustering) δηλαδή διαχωρισμό ενός συνόλου (συνήθως πολυδιάστατων) δεδομένων σε ομάδες έτσι ώστε σημεία που ανήκουν στην ίδια ομάδα να μοιάζουν όσο το δυνατόν περισσότερο και σημεία που ανήκουν σε διαφορετικές ομάδες να διαφέρουν όσο το δυνατόν περισσότερο.

Στο επόμενο σχήμα (Σχήμα 2.1) απεικονίζεται γραφικά μία υποθετική ομαδοποίηση σε δεδομένα αγοραστών σπορ αυτοκινήτων, με βάση την ηλικία (άξονας x), το ετήσιο εισόδημα (άξονας y) και το φύλλο (άνδρας-γυναίκα). Διακρίνονται τρεις ομάδες: "αγοραστές νεαρής ηλικίας ανεξαρτήτως φύλλου", "*άνδρες αγοραστές με υψηλό εισόδημα, όλων των ηλικιών μέχρι τα 53 χρόνια*" και "*άνδρες αγοραστές ηλικίας περίπου 44 ανεξαρτήτως εισοδήματος*".



Σχήμα 2.1: Υποθετική ομαδοποίηση αγοραστών σπορ αυτοκινήτων.

Η απόφαση για το πώς θα χρησιμοποιηθούν οι προκύπτουσες ομάδες λαμβάνεται από κοινού από τον ειδικό σε θέματα ανάλυσης δεδομένων και τον ειδικό του τομέα στον οποίο ανήκουν τα δεδομένα (για παράδειγμα καθορισμός διαφημιστικής προβολής).

Αλγόριθμοι ομαδοποίησης

Υπάρχουν τρεις γενικές κατηγορίες αλγορίθμων ομαδοποίησης:

- Οι αλγόριθμοι βασισμένοι σε *διαχωρισμούς* (*partition based*), που προσπαθούν να βρουν τον καλύτερο διαχωρισμό ενός συνόλου δεδομένων σε ένα συγκεκριμένο αριθμό ομάδων.
- Οι *ιεραρχικοί* (*hierarchical*) αλγόριθμοι, που προσπαθούν με ιεραρχικό τρόπο να ανακαλύψουν τον αριθμό και τη δομή των ομάδων.
- Οι *πιθανοκρατικοί* (*probabilistic*) αλγόριθμοι, που βασίζονται σε μοντέλα πιθανοτήτων.

Στη συνέχεια, παρουσιάζονται αλγόριθμοι των δύο πρώτων κατηγοριών.

Η ομαδοποίηση απαιτεί κάποιο μέτρο της ομοιότητας ή διαφοράς μεταξύ των δεδομένων. Συνήθως υπολογίζεται η "απόσταση" μεταξύ των δεδομένων. Έστω ένα σύνολο δεδομένων D, και δύο δεδομένα του, x, y που περιγράφονται από m χαρακτηριστικά $(x_1,x_2,\ldots,x_m)$, $(y_1,y_2,\ldots,y_m)$. Τυπικά μέτρα απόστασης αυτών των δύο δεδομένων είναι η απόσταση *Μανχάταν* και η *Ευκλείδεια* απόσταση:

$$d(x,y) = \sum_i |x_i - y_i|$$

Απόσταση *Μανχάταν*

$$d(x,y) = \sqrt{\sum_i (x_i - y_i)^2}$$

Ευκλείδεια απόσταση

Αν κάποια χαρακτηριστικά είναι διακριτά, τότε η απόσταση των τιμών τους θεωρείται 0 αν πρόκειται για την ίδια τιμή και 1 αν πρόκειται για διαφορετικές τιμές. Τα αριθμητικά χαρακτηριστικά θα πρέπει να ομογενοποιούνται ώστε η απόστασή τους να πέφτει μέσα στο διάστημα [0,1]. Αν *min* και *max* είναι η ελάχιστη και μέγιστη τιμή ενός αριθμητικού χαρακτηριστικού, και x_i είναι η τιμή του για κάποιο δεδομένο, τότε η διαδικασία της ομογενοποίησης μετατρέπει την τιμή στο διάστημα [0,1] με τον τύπο:

$$x_i = \frac{x_i - \min}{\max - \min}$$

Αλγόριθμοι βασισμένοι σε διαχωρισμούς

Ένας από τους πιο γνωστούς αλγόριθμους ομαδοποίησης αυτής της κατηγορίας είναι ο *αλγόριθμος των* ***K-μέσων*** **(*K-means*)** [MacQueen 1967]. Ο αριθμός K των ομάδων καθορίζεται πριν την εκτέλεση του αλγορίθμου. Ο αλγόριθμος ξεκινά διαλέγοντας K τυχαία σημεία από τα δεδομένα ως τα κέντρα των ομάδων. Έπειτα αναθέτει κάθε σημείο στην ομάδα της οποίας το κέντρο είναι πιο κοντά (μικρότερη απόσταση) σε αυτό το σημείο. Στη συνέχεια, υπολογίζει για κάθε ομάδα το μέσο όρο όλων των σημείων της (*μέσο διάνυσμα*) και ορίζει αυτό ως νέο κέντρο της. Τα δύο τελευταία βήματα επαναλαμβάνονται για ένα προκαθορισμένο αριθμό βημάτων ή μέχρι να μην υπάρχει αλλαγή στο διαχωρισμό των σημείων σε ομάδες.

Αλγόριθμοι ιεραρχικής ομαδοποίησης

Οι αλγόριθμοι ιεραρχικής ομαδοποίησης συνδυάζουν ομάδες σε μεγαλύτερες ομάδες ή διαιρούν μεγάλες ομάδες σε μικρότερες. Το αποτέλεσμα των αλγορίθμων αυτών είναι μια ιεραρχία από διαφορετικές ομαδοποιήσεις των δεδομένων στο ένα άκρο της οποίας βρίσκεται μια μόνο ομάδα με όλα τα δεδομένα, και στο άλλο τόσες ομάδες όσες και ο αριθμός των δεδομένων. Με βάση την κατεύθυνση ανάπτυξης της ιεραρχίας που ακολουθούν, οι ιεραρχικοί αλγόριθμοι ομαδοποίησης χωρίζονται στους αλγορίθμους *συγχώνευσης* (*agglomerative*) και στους αλγορίθμους *διαίρεσης* (*divisive*) [Jain and Dubes, 1988].

Οι αλγόριθμοι συγχώνευσης είναι οι πιο σημαντικοί και διαδεδομένοι από τους δύο. Βασίζονται σε μετρικές απόστασης ανάμεσα σε ομάδες. Δεδομένης μιας αρχικής ομαδοποίησης (για παράδειγμα, κάθε σημείο αποτελεί μια ομάδα), οι αλγόριθμοι

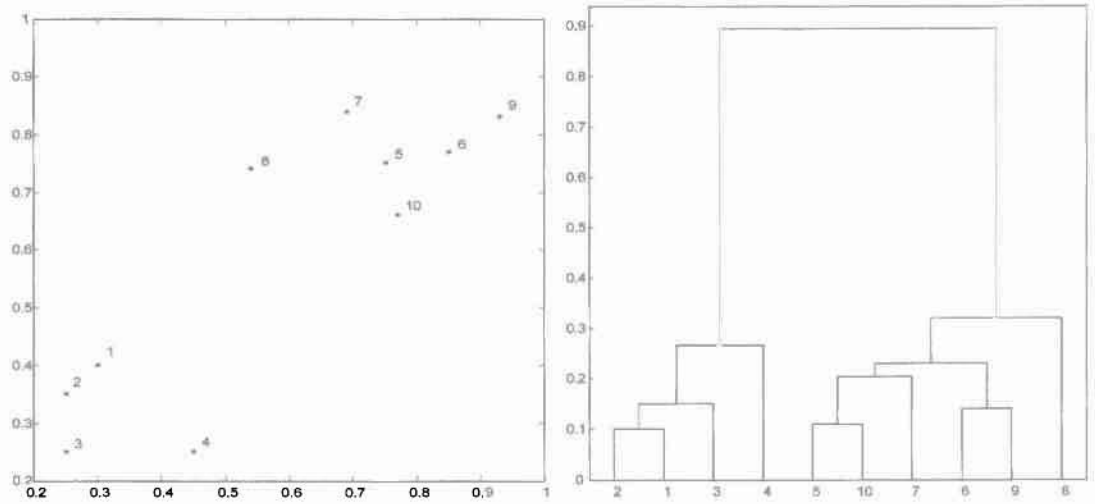
αυτοί βρίσκουν τις δύο πιο κοντινές ομάδες και τις συγχωνεύουν σε μία. Η διαδικασία συνεχίζεται μέχρις ότου προκύψει μία μόνο ομάδα. Βέβαια, θα πρέπει να προσδιοριστούν κάποιες λεπτομέρειες του αλγορίθμου και πιο συγκεκριμένα ο τρόπος υπολογισμού της απόστασης μεταξύ δύο σημείων και της απόστασης μεταξύ δύο ομάδων.

Για τη μεν πρώτη θα μπορούσε να χρησιμοποιηθεί η απόσταση *Μανχάταν* ή η *Ευκλείδεια* απόσταση. Για τη δεύτερη θα μπορούσε να υπολογισθεί η απόσταση μεταξύ των μέσων διανυσμάτων των σημείων που ανήκουν στην κάθε ομάδα ή η απόσταση μεταξύ των δύο πιο κοντινών ή πιο απομακρυσμένων σημείων των ομάδων.

Τέλος, θα πρέπει να εξεταστεί ποια ομαδοποίηση από όλη την ιεραρχία ομαδοποιήσεων είναι η καλύτερη. Μια επιλογή είναι αυτή για την οποία το "πλάτος" κάθε ομάδας είναι σημαντικά μικρότερο από την απόστασή της από τον κοντινότερο γείτονα. Εναλλακτικά, μπορεί κάποιος να μελετήσει περισσότερο ορισμένες ομαδοποιήσεις των δεδομένων, ανάλογα με την εφαρμογή και τον αριθμό των ομάδων που υποψιάζεται.

Η πολυπλοκότητα της ιεραρχικής ομαδοποίησης συγχώνευσης είναι στην καλύτερη περίπτωση $O(n^2)$, επειδή έχουμε n επαναλήψεις του βασικού βρόχου και σε κάθε επανάληψη k πρέπει να υπολογιστεί η μικρότερη απόσταση ανάμεσα σε n-k+1 ομάδες.

Οι ιεραρχίες που προκύπτουν από τους αλγορίθμους ιεραρχικής ομαδοποίησης μπορεί να απεικονιστούν με έναν πρακτικό και εύκολο τρόπο μέσω ενός γραφήματος δενδρικής μορφής, το οποίο ονομάζεται δενδρόγραμμα. Στο Σχήμα 2.2 αριστερά, φαίνονται κάποια σημεία στο δυσδιάστατο χώρο, ενώ δεξιά παρουσιάζεται το δενδρόγραμμα που προκύπτει μέσω ιεραρχικής ομαδοποίησης.



Σχήμα 2.2: Δενδρόγραμμα ιεραρχικής ομαδοποίησης.

ΚΕΦΑΛΑΙΟ 3

Ασαφής Ανακάλυψη Γνώσης

3.1 Εισαγωγή

Η διαδικασία που ακολουθεί την αποθήκευση πληροφορίας, είναι η ανάλυση και επεξεργασία των δεδομένων, με στόχο την εξαγωγή χρήσιμων συμπερασμάτων. Οι περισσότερες μεθοδολογίες που χρησιμοποιούνται, θεωρούν πως τα δεδομένα είναι καθορισμένα με μεγάλη σαφήνεια και προέρχονται από ακριβείς μετρήσεις. Ωστόσο στον πραγματικό κόσμο και λόγο, άφθονες είναι οι ασαφείς και ανακριβείς (imprecise) έννοιες. Για παράδειγμα οι εκφράσεις «Ο Γιώργος είναι ψηλός» ή «Σήμερα έχει πολλή ζέστη» περιέχουν τις έννοιες «ψηλός» και «πολλή ζέστη» που αν και δεν είναι σαφώς καθορισμένες περιέχουν πληροφορία. Τέτοιες προτάσεις δεν μπορούν να προσδιοριστούν με μεγαλύτερη ακρίβεια χωρίς να χαθεί κάποιο τμήμα της σημασίας τους.

Η ασάφεια είναι η έννοια που σχετίζεται με την ποσοτικοποίηση μιας ποιοτικής πληροφορίας. Είναι ένα εγγενές χαρακτηριστικό της γλώσσας που οφείλεται τόσο στις έννοιες που χρησιμοποιούμε, όσο και στην προσωπική αντίληψη του καθένα για τους λεκτικούς προσδιορισμούς ποσοτικών μεγεθών.

Η Ασαφής Λογική είναι ένα υπερσύνολο της Boolean (Δίτιμης) λογικής η οποία έχει επεκταθεί ώστε να μπορεί να χειριστεί τιμές αληθείας μεταξύ του «απολύτως αληθούς» και του «απολύτως ψευδούς». Σύμφωνα με τη Boolean λογική ένα αντικείμενο είτε ανήκει σε ένα σύνολο είτε όχι, και οι πιθανές τιμές μιας μεταβλητής είναι ανάλογα 1 και 0. Ωστόσο στις περισσότερες πραγματικές περιπτώσεις είναι δύσκολος ο καθορισμός των ορίων των συνόλων και η κατάταξη των δεδομένων που βρίσκονται γύρω από τα όρια αυτά, ώστε να μην επηρεάζεται η ποιοτική πληροφορία που αυτά μεταφέρουν. Αυτή τη δυσκολία, η Ασαφής Λογική την αντιμετωπίζει χρησιμοποιώντας ενδιάμεσες τιμές στο διάστημα [0,1] που εκφράζουν σε ποιο άκρο του συνόλου πλησιάζει το κάθε αντικείμενο, και κατά πόσο, δηλαδή σε ποιο βαθμό ανήκει στο σύνολο [Zadeh, L. A. et al. 1996].

Η Ασαφής Λογική έχει τις ρίζες της στη Θεωρία των Ασαφών Συνόλων (Fuzzy Set Theory), η οποία προτάθηκε από τον Lofti Zadeh τη δεκαετία του '60. Η βασική ιδέα αυτής της θεωρίας είναι ότι η διαδικασία της μετατροπής διακριτών μεγεθών σε ασαφή (fuzzification) επιτρέπει τη γενίκευση μιας διακριτής (distinct) θεωρίας σε συνεχόμενη (continuous).

3.2 Βασικές Έννοιες της Ασαφούς Λογικής

Θεωρώντας ένα σύνολο στοιχείων X, ένα *ασαφές σύνολο* (fuzzy set) A ορισμένο στο X, είναι το σύνολο των διατεταγμένων ζευγών $(x, u_A(x))$, όπου $x \in X$ και $u_A(x) \in [0,1]$.

Η συνάρτηση $u_A(x)$ ονομάζεται *συνάρτηση συγγένειας* (membership function), ενώ η τιμή $u_A(x)$ λέγεται *βαθμός αληθείας* (degree of truth) και εκφράζει το βαθμό της συγγένειας του x στο A.

Η συνάρτηση συγγένειας είναι αντίστοιχη με αυτή που χρησιμοποιούμε στην κλασική λογική για να τοποθετήσουμε ένα στοιχείο σε ένα σύνολο. Ωστόσο εκεί $u_A(x) \in \{0,1\}$, δηλαδή είτε το x ανήκει στο A ($u_A(x)=1$), είτε δεν ανήκει στο A ($u_A(x)=0$). Συνεπώς η ασαφής θεωρία συνόλων μεταπίπτει στην αντίστοιχη κλασική όταν οι δυνατές τιμές της συνάρτησης συγγένειας περιορίζονται στο 0 και 1.

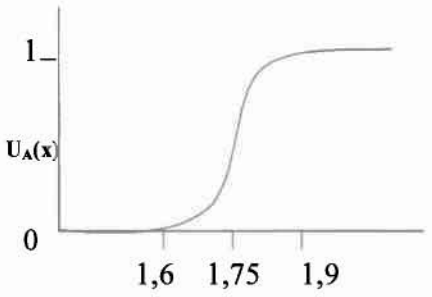
Παρατηρούμε πως υπάρχει μεγάλη ομοιότητα με την έκφραση πιθανότητας, όπου επίσης οι τιμές είναι στο διάστημα [0,1] και το 0 δηλώνει Ψέμα (μη ύπαρξη) και το 1 Αλήθεια (ύπαρξη). Η διαφορά βρίσκεται στη σημασιολογία με την οποία γίνεται η ερμηνεία. Η πιθανότητα, εκφράζει το μέτρο στο οποίο γνωρίζουμε αν ένα αντικείμενο ανήκει ή όχι σε ένα σύνολο, ενώ σε ένα ασαφές σύνολο η συνάρτηση συγγένειας δίνει το πόσο περισσότερο ή λιγότερο ανήκει το αντικείμενο στο σύνολο αυτό.

Στην πράξη η συνάρτηση συγγενείας μπορεί να προέρχεται από:

- Υποκειμενικές εκτιμήσεις
- Συχνότητες εμφανίσεων και πιθανότητες
- Φυσικές μετρήσεις
- Διαδικασίες μάθησης και προσαρμογής.

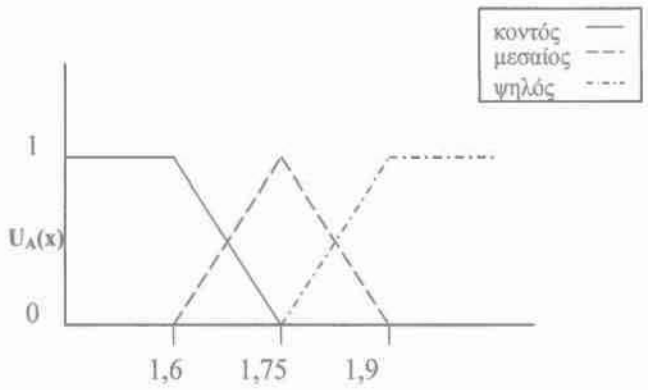
3.2.1 Αναπαράσταση Ασαφών Συνόλων

Ένας τρόπος αναπαράστασης ασαφών συνόλων είναι μέσω της συνάρτησης συγγένειάς τους. Παρακάτω (Σχήμα 3.1) φαίνεται μια συνάρτηση συγγένειας που προσδιορίζει πότε κάποιος θεωρείται ψηλός. Ένα άτομο x με ύψος>1.95 θεωρείται ψηλό με $u_A(x) \approx 1$, κάποιος με ύψος 1.75 θεωρείται ψηλός με $u_A(x) \approx 0.5$ και κάποιος με ύψος<1.60 θεωρείται ψηλός με $u_A(x) \sim 0$.



Σχήμα 3.1: Συνάρτηση Συγγένειας για την έννοια «ψηλός».

Ένας άλλος τρόπος αναπαράστασης που χρησιμοποιείται σε πολλές εφαρμογές είναι η Τμηματικώς Γραμμική Απεικόνιση της συνάρτησης συγγένειας. Για παράδειγμα η συνάρτηση συγγένειας για τα σύνολα : «κοντός», «μεσαίος», «ψηλός». Παρατηρούμε πως υπάρχει αλληλοεπικάλυψη για κάποιες τιμές του ύψους, αυτό όμως είναι εγγενές χαρακτηριστικό της θεωρίας των ασαφών συνόλων και πάνω σε αυτό στηρίζεται η ασαφής συλλογιστική.



Σχήμα 3.2: Τμηματικώς Γραμμική Συνάρτηση Συγγένειας για τα ασαφή σύνολα «κοντός», «μεσαίος», «ψηλός».

Τα ασαφή σύνολα μπορούν να αναπαρασταθούν και σαν ένα σύνολο ζευγών της μορφής $u_A(x)/x$ όπου x το στοιχείο του συνόλου και $u_A(x)$ ο βαθμός συγγένειάς του. Σε αυτήν την περίπτωση το παραπάνω σύνολο «Ψηλός» παριστάνεται ως:

A=(0/1.5, 0/1.6, 0.5/1.75, 1/1.9, 1/2.1).

Επίσης η αναπαράσταση των ζευγών μπορεί να έχει και τη μορφή $(x, u_A(x))$. Δηλαδή:

A=((1.7, 0), (1.75, 0), (1.80, 0.33), (1.85, 0.66), (1.90, 1), (1.95, 1)) .

Η αναπαράσταση με ζεύγη τιμών περιγράφει διακριτές τιμές (μόνο τα συγκεκριμένα ύψη), ενώ η αναλυτική έκφραση της u(x) δίνει συνεχείς τιμές (όλα τα ύψη). Η επιλογή γίνεται ανάλογα με τις απαιτήσεις της εφαρμογής.

3.2.2 Ιδιότητες Ασαφών Συνόλων

Έστω δύο ασαφή σύνολα A και B, ορισμένα στο S. Τα σύνολα αυτά θεωρούνται ***ίσα*** αν οι συναρτήσεις συγγένειάς τους είναι ίσες σε όλο το πεδίο ορισμού τους, $u_A(x)= u_B(x)$ για$\forall\ x\in S$.

Κενό ασαφές σύνολο,$\varnothing$ είναι αυτό με συνάρτηση συγγένειας 0.

Το ***συμπληρωματικό*** (complement) ενός ασαφούς συνόλου A είναι το A΄ με $u'_A(x)=1- u_A(x)$, και ισοδυναμεί με την άρνηση (NOT) στην ασαφή λογική.

Το A είναι ***υποσύνολο*** του B ($A\subseteq B$), αν $u_A(x) \leq u_B(x)\ \forall x\in S$.

Η ***ένωση*** δύο ασαφών συνόλων A και B, είναι ένα νέο ασαφές σύνολο $A\cup B$ ορισμένο επίσης στο S, για το οποίο ισχύει:

$A\cup B$: $u_{A\cup B}(x) = \vee (u_A(x), u_B(x)) = \max(u_A(x), u_B(x))\ \ \forall\ x\in S$

Η ένωση δύο ασαφών συνόλων σχετίζεται με τη διάζευξη (OR) της ασαφούς λογικής.

Ανάλογα, η ***τομή*** δύο ασαφών συνόλων A και B ορισμένων στο ίδιο σύνολο S, είναι ένα νέο ασαφές σύνολο $A \cap B$ ορισμένο επίσης στο S για το οποίο ισχύει:

$A \cap B$: $u_{A \cap B}(x) = \wedge (u_A(x), u_B(x)) = \min(u_A(x), u_B(x)) \quad \forall x \in S$

Η τομή δύο ασαφών συνόλων σχετίζεται με τη σύζευξη (AND) της ασαφούς λογικής.

Οι περισσότερες ιδιότητες της κλασικής θεωρίας συνόλων, ισχύουν και για τα ασαφή σύνολα.

Αντιμεταθετική Ιδιότητα : $A \cap B = B \cap A$ και $A \cup B = B \cup A$

Προσεταιριστική Ιδιότητα : $(A \cup B) \cup C = A \cup (B \cup C)$ και

$$(A \cap B) \cap C = A \cap (B \cap C)$$

Επιμεριστική Ιδιότητα : $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ και

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$

Ωστόσο υπάρχουν και ιδιότητες που ισχύουν στα ασαφή σύνολα και όχι στην κλασική θεωρία συνόλων.

Ο ***νόμος της αντίφασης*** (law of contradiction) : $A \cap \bar{A} \neq \varnothing$

Ο ***νόμος του αποκλειόμενου μέσου*** (law of the excluded middle): $A \cup \bar{A} \neq S$

3.2.3 Ασαφείς Μεταβλητές

Αν το πεδίο τιμών μιας μεταβλητής καλυφθεί από διάφορα ασαφή σύνολα με την ανάλογη σημασιολογία, τότε προκύπτει η ασαφής μεταβλητή που παίρνει τιμές από τα σύνολα αυτά, και μας επιτρέπει πλέον να κάνουμε υπολογισμούς με λέξεις και όχι με αριθμούς. Για παράδειγμα, τα ασαφή σύνολα {κοντός, μεσαίος, ψηλός} που απεικονίστηκαν παραπάνω θα μπορούσαν να αποτελούν το πεδίο τιμών μιας ασαφούς μεταβλητής "Ύψος". Αυτός ο τρόπος αναπαράστασης των μεταβλητών ταιριάζει ιδιαίτερα στις πραγματικές εφαρμογές όπου χρησιμοποιούνται μεγέθη ασαφούς φύσης, αλλά και μη ακριβείς ή υποκειμενικές τιμές.

Οι ασαφείς μεταβλητές όπως το "Ύψος" χαρακτηρίζονται και ως λεκτικές (linguistic) μεταβλητές, ενώ οι τιμές "κοντός", "μεσαίος", "ψηλός" ως πρωταρχικές λεκτικές τιμές. Από ένα μικρό αρχικό αριθμό πρωταρχικών λεκτικών τιμών, μπορούμε να πάρουμε ένα πολύ μεγαλύτερο αριθμό σύνθετων λεκτικών τιμών με χρήση λεκτικών τελεστών όπως AND, OR, NOT, VERY κτλ. Αυτοί είναι οι αντίστοιχοι των κλασικών τελεστών της Boolean λογικής, οι οποίοι έχουν επεκταθεί στα ασαφή σύνολα. Αυτή η οικογένεια τελεστών προκύπτει από μία ή περισσότερες τιμές του ασαφούς συνόλου, χωρίς να απαιτεί τον συνδυασμό όλων των τιμών του ασαφούς συνόλου όπως συμβαίνει στους ασαφείς αριθμούς. Για παράδειγμα οι παρακάτω λεκτικοί τελεστές επιδρούν στη συνάρτηση συγγένειας ως εξής:

A AND B: $u_{A \text{ AND } B}(x) = \min\{u_A(x), u_B(x)\}$	VERY A : $u_{CON(A)}(x) = [u_A(x)]^2$
A OR B : $u_{A \text{ OR } B}(x) = \max\{u_A(x), u_B(x)\}$	PLUS A : $u_{PLUS(A)}(x) = [u_A(x)]^{1.25}$
NOT A : $u_{NOT\ A}(x) = 1 - u_A(x)$	MINUS A: $u_{MINUS(A)}(x) = [u_A(x)]^{0.75}$

3.2.4 Ασαφείς Κανόνες

Μία πρόταση λέγεται ασαφής όταν θέτει μία τιμή σε μια ασαφή μεταβλητή. Π.χ. στην ασαφή πρόταση "Το κέρδος της εταιρείας είναι χαμηλό", η ασαφής μεταβλητή *κέρδος* παίρνει την τιμή *χαμηλό* που είναι ένα ασαφές σύνολο. Ένας ασαφής κανόνας (fuzzy rule) είναι μια υπό συνθήκη έκφραση που συσχετίζει δύο ή περισσότερες ασαφείς προτάσεις:

IF age(x) IS young THEN risk(x) IS high

Αυτός είναι ένας κανόνας που περιγράφει τον παράγοντα του ρίσκου για μια ασφαλιστική εταιρία αυτοκινήτων. Ένας βασικός τύπος ασαφών κανόνων που χρησιμοποιείται ευρύτερα έχει τη μορφή:

IF x_1 IS A_1 AND ... AND x_n IS A_n THEN y IS B

όπου στο τμήμα του συμπεράσματος παίρνει τιμή μία λεκτική μεταβλητή ενώ στο τμήμα της υπόθεσης συμμετέχουν περισσότερες από μία λεκτικές μεταβλητές.

Η αναλυτική περιγραφή ενός ασαφούς κανόνα if/then είναι μια ασαφής σχέση R(x,y) που ονομάζεται σχέση συνεπαγωγής (implication relation). Προκύπτει με συνδυασμό των συναρτήσεων συγγένειας των ασαφών συνόλων x και y.

$$R(x,y) \equiv u(x,y) = \varphi(u_A(x),u_B(y))$$

Ο τελεστής φ εκφράζει τον τρόπο με τον οποίο πρέπει να συνδυαστούν οι συναρτήσεις συγγένειας του if και του then τμήματος σε έναν ασαφή κανόνα, ώστε να προκύψει η αναλυτική του έκφραση, και λέγεται τελεστής συνεπαγωγής (implication operator). Η επιλογή του φ εξαρτάται από την εκάστοτε εφαρμογή. Μερικοί ασαφείς τελεστές συνεπαγωγής είναι:

Ονομασία Τελεστή	Αναλυτική έκφραση του $\varphi[u_A(x),u_B(y)]$
φ_m : Zadeh Max-Min	$(u_A(x) \wedge u_B(y)) \vee (1- u_A(x))$
φ_c : Mandani Min	$u_A(x) \wedge u_B(y)$
φ_p :Larsen Product	$u_A(x) * u_B(y)$
φ_a : Arithmetic	$1 \wedge (1-u_A(x)+ u_B(y))$
φ_b : Boolean	$(1-u_A(x)) \vee u_B(y)$
Όπου $\wedge \equiv$ min και $\vee \equiv$ max	

Χρησιμοποιώντας ασαφείς κανόνες τα προβλήματα που μπορούν να εκφραστούν είναι δύο ειδών.

- Να είναι γνωστή η τιμή A της ασαφούς μεταβλητής x και να πρέπει να υπολογιστεί η τιμή B, της ασαφούς μεταβλητής y (συλλογιστική διαδικασία Generalized Modus Ponens (GMP)).
- Να είναι γνωστή η τιμή B της ασαφούς μεταβλητής y και να πρέπει να υπολογιστεί η τιμή A της ασαφούς μεταβλητής x (συλλογιστική διαδικασία Generalized Modus Tollens (GMT)).

Η περιγραφή ενός προβλήματος με ασαφείς μεταβλητές, ασαφείς τιμές και ασαφείς κανόνες ονομάζεται ασαφής λεκτική περιγραφή (fuzzy linguistic description).

3.2.5 Εφαρμογή της ασαφούς λογικής

Η εφαρμογή της ασαφούς λογικής σε ένα σύστημα περιλαμβάνει συγκεκριμένα βήματα και ενέργειες.

Το πρώτο βήμα αφορά τα δεδομένα εισόδου, τα οποία πρέπει να μετατραπούν σε ασαφή (*ασαφοποίηση- fuzzification*). Αυτό γίνεται μέσω των συναρτήσεων συγγένειας, οι οποίες έχουν καθοριστεί με τη βοήθεια κάποιου ειδικού στον τομέα.

Στη συνέχεια, οι ασαφείς κανόνες οι οποίοι ικανοποιούνται από τα δεδομένα εισόδου, ενεργοποιούνται για να δώσουν τις τιμές αλήθειας για τις παραμέτρους εξόδου (*εξαγωγή συμπερασμάτων - inference*).

Ακολουθεί η φάση της *σύνθεσης (combination)*. Εδώ οι συναρτήσεις συγγένειας των μεταβλητών εξόδου, που έχουν προκύψει από τους κανόνες που ενεργοποιήθηκαν, συνδυάζονται σε μία συνάρτηση συγγένειας, η οποία και μας δίνει το τελικό αποτέλεσμα.

Πολλές φορές, ανάλογα με τη εφαρμογή, υπάρχει ένα στάδιο ακόμα, στο οποίο γίνεται η *αποσαφήνιση* ή αποασαφοποίηση (*defuzzification*) του αποτελέσματος στην έξοδο. Μετατρέπεται δηλαδή σε ακριβή τιμή (crisp), που μπορεί να εκφραστεί αριθμητικά.

3.3 Εξαγωγή Ασαφών Μοντέλων από Δεδομένα

Η θεωρία ασαφών συστημάτων προσφέρει, τελευταία, ένα σημαντικό εργαλείο για την ανάπτυξη μοντέλων τα οποία είναι εύκολα ερμηνεύσιμα από τους ειδικούς, ενώ παράλληλα επιτυγχάνει αποτελεσματική μοντελοποίηση της ασάφειας και της αοριστίας η οποία είναι εγγενής στην φύση. Οι ασαφείς ταξινομητές αναφέρονται συνήθως ως ελαστικοί ταξινομητές με μεταβλητούς βαθμούς συμμετοχής στις κλάσεις, σε αντιδιαστολή με τους κλασικούς ("σκληρούς")ταξινομητές οι οποίοι παρέχουν βαθμούς συμμετοχής σε μία μόνο κλάση. Η σημασία τους στην ταξινόμηση ειδών είναι μεγάλη, δεδομένης της ύπαρξης μεικτών δεδομένων με αμφίβολα χαρακτηριστικά.

Οι αλγόριθμοι ασαφούς ομαδοποίησης επεκτείνουν τις αρχές των κλασικών τεχνικών ομαδοποίησης, εκμεταλλευόμενοι τα ελκυστικά χαρακτηριστικά της θεωρίας των ασαφών συνόλων. Κατά τα τελευταία χρόνια, η έρευνα έχει εστιαστεί στην ανάπτυξη αλγορίθμων επιβλεπόμενης ασαφούς ομαδοποίησης όπου χρησιμοποιούνται γενετικοί αλγόριθμοι για τη βελτιστοποίηση των παραμέτρων. Συγκεκριμένα, οι γενετικοί αλγόριθμοι χρησιμοποιούνται για την τοποθέτηση των ομάδων στον χώρο των χαρακτηριστικών, θεωρώντας διαφορετικές μεθοδολογίες ασαφούς ομαδοποίησης, όπως ο fuzzy c-means. Γενετικοί αλγόριθμοι πολύ-παραγοντικής βελτιστοποίησης έχουν επίσης προταθεί με στόχο τη διαδοχική βελτιστοποίηση πολλαπλών κριτηρίων αξιολόγησης, σε μια προσπάθεια να αντιμετωπισθεί το πρόβλημα του διαμερισμού του χώρου χαρακτηριστικών.

Στην περιοχή των ασαφών συστημάτων τα συστήματα βασισμένα σε ασαφείς κανόνες (fuzzy rule based systems, FRBS) παρέχουν την πιο διαισθητική παράσταση γνώσης. Συγκεκριμένα, αποτελούνται από ασαφείς κανόνες οι οποίοι προσομοιάζουν σε μεγάλο βαθμό στις λογικές προτάσεις και τον μηχανισμό απόφασης και εξαγωγής συμπερασμάτων του ανθρώπου. Τα συστήματα FRBS έχουν τύχει εκτεταμένης εφαρμογής σε μια μεγάλη ποικιλία εφαρμογών, δεδομένου ότι προσδίδουν στους αναλυτές σημαντική ποιοτική πληροφορία του υπό εξέταση συστήματος, τα συστήματα ταξινόμησης βασισμένα στους ασαφείς κανόνες ταξινόμησης (Fuzzy Rule-Based Classification Systems, FRBCS).

Στη συνέχεια θα παρουσιαστούν ενδεικτικοί αλγόριθμοι ασαφούς μάθησης και συγκεκριμένα ο αλγόριθμος ασαφούς ταξινόμησης Fuzzy KNN και ο αλγόριθμος ασαφούς ομαδοποίησης Fuzzy C-means.

3.3.1 Ασαφής αλγόριθμος των k-κοντινότερων γειτόνων (Fuzzy k-Nearest Neighbors-Fuzzy KNN)

Το πρόβλημα της ταξινόμησης N οντοτήτων σε M κατηγορίες μπορεί να διατυπωθεί ως $\Omega=\{w_1,w_2,\ldots,w_M\}$, όπου τα w_i δείχνουν την κατηγορία i_{th}. Οι διαθέσιμες πληροφορίες υποτίθεται ότι συστάθηκαν σε ένα σύνολο δεδομένων εκπαίδευσης $\Pi=\{(\mathbf{X}_1,c_1),\ldots,(\mathbf{X}_N,c_N)\}$ από N πρότυπα $\mathbf{X}_i$ $(i=1,2,\ldots,N)$ και τις αντίστοιχες ετικέτες κατηγορίας τους $c_i(i=1,2,\ldots,N)$ που παίρνει τις τιμές από το Ω. Ο κανόνας KNN [Cover and Hart. 1967] είναι καλά γνωστός στην περιοχή της αναγνώρισης προτύπων. Σύμφωνα με αυτόν τον κανόνα, ένα αταξινόμητο δεδομένο (περίπτωση) $\mathbf{X}$ κατατάσσεται στην κατηγορία που αντιπροσωπεύεται από μια πλειοψηφία των κοντινότερών του γειτόνων K στο Π. Εάν δύο κατηγορίες αντιπροσωπεύονται από τον ίδιο αριθμό γειτόνων, τότε το δείγμα κατατάσσεται σε μια από τις δύο κατηγορίες τυχαία. Εντούτοις, περιπτώσεις όπως αυτή συμβαίνουν σπάνια. Αυτός ο κανόνας σήμερα συνήθως καλείται «κανόνας K-NN ψηφοφορίας». Ο KNN είναι δημοφιλής στην κοινότητα αναγνώρισης προτύπων και αυτό οφείλεται κυρίως στην καλή εκτέλεσή του και στην απλή του χρήση. Από την εμφάνιση του KNN έχουν προταθεί μερικές παραλλαγές προκειμένου να βελτιωθεί η απόδοσή του.

Το 1985 ο Keller πρότεινε μια νέα προσέγγιση με το συνδυασμό της ασαφούς θεωρίας με τον αλγόριθμο KNN, και την ονόμασε ως «ασαφής αλγόριθμος ταξινομητών KNN» [Keller et al., 1985]. Σύμφωνα με την προσέγγισή του, παρά τις μεμονωμένες κατηγορίες όπως στον KNN, τα δείγματα σχετίζονται με τις διαφορετικές κατηγορίες σύμφωνα με την ακόλουθη σχέση:

$$\mu_i(\mathbf{X}) = \frac{\sum_{j=1}^{K} \mu_i(\mathbf{X}_j) d(\mathbf{X},\mathbf{X}_j)^{-2/p-1}}{\sum_{j=1}^{K} d(\mathbf{X},\mathbf{X}_j)^{-2/p-1}}, \tag{2}$$

όπου το p, $p\in(1,\infty)$, είναι ο ασαφής συντελεστής ο οποίος επιλέγεται από τον χρήστη, K είναι ο αριθμός κοντινότερων γειτόνων, $\mu_i(\mathbf{X})$ δείχνει τη συγγένεια του δείγματος X με την κατηγορία i, και το $d(\mathbf{X},\mathbf{X}_j)$ είναι η απόσταση μεταξύ του δείγματος δοκιμής X και των κοντινότερων δειγμάτων του $\mathbf{X}_j$.

Διάφορες μετρικές μπορούν να επιλεγούν για το $d(\mathbf{X},\mathbf{X}_j)$, όπως η Ευκλείδεια απόσταση, η απόσταση Hamming, και η απόσταση Mahalanobis, μεταξύ άλλων αποστάσεων. Αφού υπολογίσει όλες τις τιμές συγγένειας για ένα δείγμα ερώτησης, το κατατάσσει στην κατηγορία με την οποία έχει την υψηλότερη τιμή συγγένειας.

3.3.2 Ασαφής αλγόριθμος ομαδοποίησης (Fuzzy Clustering-FCM)

Η ασαφής ομαδοποίηση είναι μια τεχνική στην οποία ένα σύνολο δεδομένων ομαδοποιείται σε n ομάδες με κάθε δεδομένο να ανήκει σε κάθε ομάδα με κάποιο συγκεκριμένο βαθμό. Για παράδειγμα, ένα ορισμένο δεδομένο που βρίσκεται κοντά στο κέντρο μιας ομάδας θα έχει έναν υψηλό βαθμό να ανήκει ή να έχει συγγένεια σε εκείνη την ομάδα και ένα άλλο δεδομένο που βρίσκεται μακριά από το κέντρο μιας ομάδας θα έχει έναν χαμηλό βαθμό να ανήκει ή να έχει συγγένεια σε εκείνη την ομάδα. Δυο γνωστοί αλγόριθμοι είναι ο Ασαφής Αλγόριθμος των K-μέσων και η Αφαιρετική Ομαδοποίηση.

Fuzzy C-Means Clustering (FCM)

Ο ασαφής αλγόριθμος των K-μέσων (Fuzzy c-mean ή fcm) υποστηρίζεται στο Matlab από τη συνάρτηση fcm που περιλαμβάνεται στο Fuzzy Logic Toolbox™. Η συνάρτηση ξεκινά με μια αρχική εικασία για τα κέντρα των ομάδων, τα οποία προορίζονται να χαρακτηρίσουν τη μέση θέση κάθε ομάδας. Η αρχική εικασία για αυτά τα κέντρα ομάδων είναι πιο πιθανό να είναι ανακριβής. Έπειτα, η fcm ορίζει σε κάθε σημείο στοιχείων έναν βαθμό συγγένειας για κάθε ομάδα. Με το να ενημερώνει διαδοχικά τα κέντρα ομάδων και τους βαθμούς συγγένειας για κάθε σημείο στοιχείων, η fcm διαδοχικά κινεί τα κέντρα ομάδων προς τη σωστή θέση μέσα σε ένα σύνολο δεδομένων. Αυτή η επανάληψη είναι βασισμένη στην ελαχιστοποίηση της απόστασης ενός οποιουδήποτε δεδομένου από τα κέντρα των ομάδων που σταθμίζεται από το βαθμό συγγένειας εκείνου του δεδομένου.

Αφαιρετική Ομαδοποίηση (Subtractive Clustering)

Η αφαιρετική ομαδοποίηση, [Chiu 1994], είναι ένας γρήγορος, αλγόριθμος ενός-περάσματος για τον αριθμό ομάδων και των κέντρων ομάδων σε ένα σύνολο δεδομένων, όταν δεν έχουμε μια σαφή ιδέα του πόσες ομάδες πρέπει να υπάρξουν για ένα δεδομένο σύνολο στοιχείων. Οι εκτιμήσεις ομάδων που λαμβάνονται από τη συνάρτηση subclust μπορούν να χρησιμοποιηθούν για να μονογράψουν τις επαναληπτικές βασισμένες σε βελτιστοποίηση μεθόδους ομαδοποίησης (fcm) και τις πρότυπες μεθόδους προσδιορισμού (όπως τα anfis). Η συνάρτηση subclust βρίσκει τις ομάδες με τη χρήση της αφαιρετικής μεθόδου ομαδοποίησης.

Η συνάρτηση genfis2 χτίζει επάνω στη συνάρτηση subclust για να παρέχει μια γρήγορη μέθοδο ενός-περάσματος για να πάρει τα στοιχεία εισόδου/εξόδου εκπαίδευσης και να παραγάγει ένα Sugeno-type ασαφές σύστημα συμπεράσματος που διαμορφώνει τη συμπεριφορά των στοιχείων.

3.4 Το εργαλείο Matlab

Το *MATLAB®* είναι μια υψηλής απόδοσης γλώσσα η οποία συνοδεύεται από ένα εύχρηστο προγραμματιστικό περιβάλλον για προγραμματισμό, τεχνικούς υπολογισμούς και οπτικοποίηση (*visualization*) των δεδομένων. Η βασική έννοια στο περιβάλλον MATLAB® είναι ο πίνακας (*matrix*). Μάλιστα το όνομα MATLAB® προέρχεται από τα αρχικά ***Matrix Laboratory***.

Οι πίνακες (*matrices*), τα διανύσματα (*vectors*) που είναι μονοδιάστατοι πίνακες και οι μεταβλητές, δηλώνονται και δημιουργούνται αυτόματα με τη χρήση τους χωρίς περιττές δηλώσεις που αφορούν τον τύπο, το μέγεθος και τις διαστάσεις τους. Το γεγονός αυτό παρέχει μεγάλη ευχρηστία και ταχύτητα στον προγραμματισμό, ιδιαίτερα στην επίλυση προβλημάτων που μορφοποιούνται υπό μορφή πινάκων. Επίσης, είναι εφοδιασμένο με πληθώρα συναρτήσεων π.χ. εύρεση ορίζουσας, υπολογισμό αντίστροφου, διαχείριση υποπινάκων χωρίς να είναι απαραίτητη η δημιουργία πολύπλοκων συναρτήσεων, δέσμευση και αποδέσμευση μνήμης (γίνεται αυτόματα). Επιπλέον, παρέχει ισχυρές δυνατότητες οπτικοποίησης των αποτελεσμάτων με ταχύτητα και ευκολία.

Το μειονέκτημα που έχει είναι ότι οι εντολές που δίνουμε δεν μπορούν να δημιουργήσουν αυτόνομη εφαρμογή (*standalone executable*), αλλά χρειάζονται το περιβάλλον για να εκτελεστούν. Το *MATLAB®* δηλαδή είναι ένας αλληλεπιδραστικός διερμηνευτής εντολών (*Interactive command interpreter*) και με αυτή τη φιλοσοφία πρέπει να χρησιμοποιείται. Βέβαια πρόσφατες εκδόσεις του MATLAB® έχουν ενσωματώσει το MATLAB Application Program Interface (*M-API*) που παρέχει τη δυνατότητα να κληθούν οι συναρτήσεις του από γλώσσες προγραμματισμού όπως η *C* και η *FORTRAN*, υπό μορφή κλήσεων DLL (Dynamic Link Library).

Εκτός από τη μεγάλη ευκολία που μας παρέχει σχετικά με επεξεργασία και απεικόνιση δεδομένων το MATLAB® είναι εφοδιασμένο και με ένα σύνολο από εργαλειοθήκες (toolboxes) που είναι ολοκληρωμένα περιβάλλοντα και βιβλιοθήκες που μας δίνουν τεράστια ευκολία σχεδιασμού και ανάλυσης σε εξειδικευμένα πεδία της επιστήμης. Στην εργασία αυτή χρησιμοποιήσαμε την εργαλειοθήκη ασαφούς λογικής (fuzzy toolbox) που μας δίνει τη δυνατότητα σχεδιασμού ασαφών συστημάτων. Η εργαλειοθήκη της ασαφούς λογικής ενεργοποιείται με την εντολή:

```
>> fuzzy
```

Το MATLAB® ανταποκρίνεται ενεργοποιώντας τον συντάκτη συστημάτων ασαφούς συμπερασμού (FIS - Fuzzy Inference System Editor), ο οποίος μας επιτρέπει να σχεδιάσουμε πλήρως ένα ασαφές σύστημα. Ο προκαθορισμένος τύπος ασαφούς συστήματος είναι ο τύπος mamdani, μπορούμε όμως να καθορίσουμε και ασαφές σύστημα τύπου TSK, από το menu\file\new fis.

Κεφάλαιο 4

Ανακάλυψη Γνώσης σε Ιατρικά Δεδομένα

4.1 Εισαγωγή

Ο τομέας της Υγείας, σε παγκόσμιο επίπεδο, έρχεται αντιμέτωπος με πλήθος οργανωτικών, διαρθρωτικών και λειτουργικών αλλαγών που προέρχονται από την απαίτηση για παροχή ποιοτικότερων υπηρεσιών προς τους πολίτες αλλά και την αποδοτικότερη λειτουργία του όλου συστήματος. Η εισαγωγή της πληροφορικής στον τομέα αυτό, αν και ακόμη βρίσκεται σε σχετικά αρχικό στάδιο, έχει συμβάλει σημαντικά στην συλλογή δεδομένων ασθενών που αφορούν σε αποτελέσματα ιατρικών εξετάσεων, διαγνώσεις, φαρμακευτική αγωγή, ιατρικό ιστορικό, τα οποία όμως δεν προσφέρουν ουσιαστική πρόσθετη γνώση και πληροφορία.

Σε οργανισμούς παροχής υπηρεσιών ιατρικής περίθαλψης, συνήθης είναι και η τάση συλλογής δεδομένων που αφορούν στην εύρυθμη λειτουργία τους, τα οποία μάλιστα προέρχονται από διαφορετικές πηγές: χρεώσεις νοσηλειών και ιατρικών εξετάσεων, δαπάνες προμήθειας νοσοκομειακού υλικού, δαπάνες μισθοδοσίας προσωπικού που αν αξιοποιηθούν σωστά προσφέρουν γνώση στρατηγικής σημασίας.

Η ικανότητα ανακάλυψης γνώσης στις παραπάνω περιπτώσεις είναι επιτακτική ανάγκη και παρουσιάζει αυξανόμενη ζήτηση. Σε πολλούς οργανισμούς κατάλληλα συστήματα, προσφέρουν πρόσθετη γνώση σε ειδικούς, η οποία τις περισσότερες φορές δεν είναι προφανής και η οποία μπορεί να αποτελέσει χρήσιμο εργαλείο στα χέρια ειδικών.

Στην παρούσα εργασία, για την πειραματική εφαρμογή των αλγορίθμων ανακάλυψης γνώσης και σύγκριση των αποτελεσμάτων, επιλέχθηκαν δύο σύνολα δεδομένων από το UCI machine learning repository (αποθετήριο του University of California Irvine), το "Thyroid Data set" και το "Pima Indians Diabetes Data set". Ο λόγος που επιλέχθηκαν είναι ότι χρησιμοποιούνται πολύ συχνά από αλγορίθμους μηχανικής μάθησης για έλεγχο της απόδοσης τους.

Όσον αφορά τους αλγορίθμους επιλέχθηκαν ως πιο αντιπροσωπευτικοί, 2 αλγόριθμοι ταξινόμησης (ο αλγόριθμος ταξινόμησης δένδρων και ο αλγόριθμος των K-κοντινότερων γειτόνων k-nn), ένας αλγόριθμος ομαδοποίησης (των K-μέσων) και 3 αλγόριθμοι ασαφούς μάθησης, εκ των οποίων 2 ταξινόμησης (των K-κοντινότερων γειτόνων και οι ασαφείς κανόνες) και ένας αλγόριθμος ασαφούς ομαδοποίησης (ο F-cmean).

4.2 Περιγραφή Δεδομένων

Στη συνέχεια περιγράφονται τα στοιχεία των δυο συνόλων δεδομένων που χρησιμοποιήθηκαν για την εφαρμογή των αλγορίθμων δηλαδή το "Thyroid Data set" και το "Pima Indians Diabetes Data set.

4.2.1 Thyroid Data Set

Ο θυρεοειδής αδένας είναι ο μεγαλύτερος ενδοκρινής αδένας του ανθρώπινου σώματος που μοιάζει όπως μια πεταλούδα. Εντοπίζεται στην πρόσθια περιοχή της τραχείας, έχει βάρος περίπου 20 γραμμάρια και αποτελείται από 2 λοβούς (δεξιό και αριστερό), οι οποίοι συνδέονται μεταξύ τους με τον ισθμό. Η βασική λειτουργία του είναι να ελέγχει το μεταβολισμό του σώματος. Για το σκοπό αυτό παράγει τις ορμόνες, T4 και T3, οι οποίες λένε στα κύτταρα του σώματος πόση ενέργεια να χρησιμοποιήσουν.

Είναι ένα από τα σημαντικότερα όργανα στο σώμα δεδομένου ότι η λειτουργία του επιδρά σε κάθε ουσιαστικό όργανο στο σώμα. Ορισμένες διαταραχές του θυρεοειδή είναι πολύ σοβαρές, όπως η θύελλα θυρεοειδή (ένα επεισόδιο σοβαρού υπερθυρεοειδισμού) και το κώμα myxedema (το τελικό στάδιο μη θεραπευμένου υποθυρεοειδισμού) τα οποία μπορεί να οδηγήσουν στο θάνατο σε έναν σημαντικό αριθμό περιπτώσεων [Zhang & Berardi, 1998].

Ένας θυρεοειδής που λειτουργεί σωστά θα διατηρήσει το σωστό ποσό ορμονών που απαιτούνται για να κρατήσουν τη λειτουργία του μεταβολισμού του σώματος σε ένα ικανοποιητικό ποσοστό. Δεδομένου ότι οι ορμόνες χρησιμοποιούνται (καταναλίσκονται), ο θυρεοειδής τις αντικαθιστά. Η ποσότητα ορμονών του θυρεοειδή στην κυκλοφορία του αίματος επιτηρείται και ελέγχεται από το βλεννογόνο αδένα. Όταν ο βλεννογόνος αδένας, που βρίσκεται στο κέντρο του κρανίου κάτω από τον εγκέφαλο, διαπιστώσει είτε μια έλλειψη είτε ένα υψηλό επίπεδο ορμονών θυρεοειδή, αυτό θα ρυθμίσει την ορμόνη του (TSH) και θα την στείλει στο θυρεοειδή για να του πει τι να κάνει. Όταν ο θυρεοειδής παράγει πάρα πολλή ορμόνη, το σώμα χρησιμοποιεί ενέργεια πιο γρήγορα από ότι θα έπρεπε. Αυτός ο όρος καλείται *υπερθυρεοειδισμός*. Όταν ο θυρεοειδής δεν παράγει αρκετή ορμόνη, το σώμα χρησιμοποιεί ενέργεια πιο αργά από ότι θα έπρεπε. Αυτός ο όρος καλείται *υποθυρεοειδισμός*. Υπάρχουν πολλοί διαφορετικοί λόγοι για τους οποίους καθεμία από αυτές τις συνθήκες θα μπορούσαν να δημιουργηθούν.

Αυτήν την περίοδο, περίπου 20 εκατομμύρια Αμερικανοί έχουν κάποια μορφή ασθένειας θυρεοειδή. Άτομα όλων των ηλικιών και των φυλών μπορεί να έχουν την ασθένεια θυρεοειδή. Εντούτοις, οι γυναίκες είναι πέντε έως οκτώ φορές πιθανότερο από τους άνδρες να έχουν προβλήματα. Η ασθένεια θυρεοειδή μπορεί να είναι δύσκολο να διαγνωστεί επειδή τα συμπτώματα εύκολα συγχέονται με άλλες ασθένειες. Όταν ανιχνεύεται νωρίς, η θεραπεία μπορεί να ελέγξει τη διαταραχή ακόμη και πριν από την έναρξη των συμπτωμάτων. Ευτυχώς, υπάρχει μια δοκιμή, αποκαλούμενη δοκιμή θυρεοειδοτρόπους ορμόνης (TSH), που μπορεί να εντοπίσει διαταραχές του θυρεοειδούς, ακόμη και πριν από την έναρξη των συμπτωμάτων.

Το σύνολο δεδομένων *thyroid* περιλαμβάνει 215 παραδείγματα που διακρίνονται σε (3) κατηγορίες (κλάσεις) τις, Κανονική, Υπερθυρεοειδισμός, και Υποθυρεοειδισμός, οι οποίες κατανέμονται ως εξής:

Κατηγορία 1: (normal) 150 δεδομένα

Κατηγορία 2: (hyper) 35 δεδομένα

Κατηγορία 3: (hypo) 30 δεδομένα

Κάθε παράδειγμα περιλαμβάνει 5 χαρακτηριστικά που προκύπτουν από εργαστηριακούς ελέγχους και χρησιμοποιούνται για να προβλέψουμε σε ποιά κατηγορία ανήκει ένας ασθενής του θυρεοειδούς. Τα χαρακτηριστικά παίρνουν συνεχείς αριθμητικές τιμές και είναι τα ακόλουθα:

1: Ιδιότητα κατηγορίας (1 = normal, 2 = hyper, 3 = hypo)

2: T3-resin δοκιμή λήψης (ένα ποσοστό)

3: Συνολικός ορός θυροξίνης (thyroxin) όπως μετριέται με την ισοτοπική μέθοδο μετατοπίσεων

4: Συνολικό ορός τριιωδοθυρονίνης (triiodothyronine) όπως μετριέται από τη δοκιμή radioimmuno

5: θυρεοειδοτρόπος ορμόνη (basal thyroid-stimulating hormone (TSH)) όπως μετριέται από τη δοκιμή radioimmuno

6: Μέγιστη απόλυτη διαφορά της αξίας TSH μετά από την έγχυση 200 γραμμαρίων θυρεοτροπίνης-απελευθερωτικής ορμόνης (thyrotropin-releasing hormone) σε σύγκριση με τη βασική αξία

Τα παραδείγματα έχουν την ακόλουθη μορφή:

α/α	Κατηγορία	T3	Thyroxin	Triiod.	TSH	Diff.
1	1	107	10.1	2.2	0.9	2.7
2	1	113	9.9	3.1	2.0	5.9
3	2	134	16.4	4.8	0.6	0.1
4	2	110	3.5	0.6	1.7	1.4
5	3	108	3.5	0.6	1.7	1.4
6	3	120	3.0	2.5	1.2	4.5
....	...	...	....	....	....	....
215	....	....	....	....	....	....

4.2.2 Pima Indians Diabetes Data set

Ο **σακχαρώδης διαβήτης** είναι μεταβολική ασθένεια η οποία χαρακτηρίζεται από αύξηση της συγκέντρωσης του σακχάρου στο αίμα (υπεργλυκαιμία) και διαταραχή του μεταβολισμού της γλυκόζης, είτε ως αποτέλεσμα ελαττωμένης έκκρισης ινσουλίνης είτε λόγω ελάττωσης της ευαισθησίας των κυττάρων του σώματος στην

ινσουλίνη. Οι κύριοι τύποι σακχαρώδους διαβήτη είναι ο διαβήτης τύπου 1, ο διαβήτης τύπου 2 και ο διαβήτης της κύησης. Ο σακχαρώδης διαβήτης έχει χρόνια πορεία και μπορεί να προκαλέσει μια σειρά σοβαρών επιπλοκών όπως καρδιαγγειακή νόσο, χρόνια νεφρική ανεπάρκεια, βλάβες του αμφιβληστροειδούς, βλάβες των νεύρων, στυτική δυσλειτουργία, κ.ά. Πρωτεύοντα ρόλο στη θεραπεία του σακχαρώδους διαβήτη παίζει η χορήγηση ινσουλίνης.

Η έρευνα NIDDK που πραγματοποιείται για τους Ινδιάνους Pima (ο πληθυσμός ζει κοντά στο Phoenix, Arizona, ΗΠΑ) για τα προηγούμενα 30 έτη έχει βοηθήσει τους επιστήμονες να αποδείξουν ότι η παχυσαρκία είναι ένας σημαντικός παράγοντας κινδύνου στην ανάπτυξη του διαβήτη. Το 50% των ενηλίκων των Ινδιάνων Pima έχουν διαβήτη και 95% των ατόμων με διαβήτη είναι υπέρβαροι.

Οι μελέτες αυτές, που πραγματοποιούνται με τη βοήθεια των Ινδιάνων Pima, έχουν δείξει ότι, πριν από την απόκτηση βάρους, τα υπέρβαρα άτομα έχουν μεταβολικό ρυθμό βραδύτερο σε σχέση με άτομα του ιδίου βάρους. Οι επιστήμονες θεωρούν ότι αυτός ο βραδύτερος ρυθμός μεταβολισμού, σε συνδυασμό με την υψηλής περιεκτικότητας σε λίπος διατροφή και μια γενετική τάση να διατηρούν το λίπος, μπορεί να προκαλεί το επιδημικό υπερβολικό βάρος που βλέπουμε στους Ινδιάνους Pima .

Μαζί με τη γενετική σύνθεση, η διατροφή είναι ένας βασικός παράγοντας για τον υγιή τρόπο ζωής. Η επιρροή των παραδοσιακών συγκομιδών στο μεταβολισμό των Ινδιάνων Pima είναι υπό μελέτη για να προσδιορίσει τον τρόπο να αποτρέψει την εμφάνιση του διαβήτη και της παχυσαρκίας.

Οι επιστήμονες χρησιμοποιούν τη θεωρία «οικονόμων γονιδίων» που προτάθηκε το 1962 από το γενετιστή James Neel για να βοηθήσει να εξηγήσουν γιατί πολλοί Ινδιάνοι Pima είναι υπέρβαροι. Η θεωρία του Neel είναι βασισμένη στο γεγονός ότι για χιλιάδες έτη πληθυσμοί που στηρίχθηκαν στην γεωργία, το κυνήγι και την αλιεία για τα τρόφιμα, όπως οι Ινδιάνοι Pima, έζησαν τις εναλλασσόμενες περιόδους αφθονίας αγαθών και του λιμού. Ο Neel είπε ότι για να προσαρμοστούν σε αυτές τις ακραίες αλλαγές στις θερμιδικές ανάγκες, αυτοί οι άνθρωποι ανέπτυξαν ένα οικονόμο γονίδιο που τους επέτρεψε να αποθηκεύουν το λίπος κατά τη διάρκεια των χρόνων της αφθονίας έτσι ώστε δεν θα λιμοκτονούσαν κατά τη διάρκεια του λιμού.

Αυτό το γονίδιο ήταν χρήσιμο εφ' όσον υπήρξαν περίοδοι λιμού. Αλλά μόλις υιοθέτησαν αυτοί οι πληθυσμοί το χαρακτηριστικό δυτικό τρόπο ζωής, με λιγότερη σωματική δραστηριότητα, μια πλούσια σε λίπη διατροφή, και την πρόσβαση σε έναν σταθερό ανεφοδιασμό των θερμίδων, αυτό το γονίδιο άρχισε να λειτουργεί σε βάρος τους, συνεχίζοντας να αποθηκεύουν θερμίδες στο πλαίσιο της προετοιμασίας για την καταπολέμηση του λιμού. Οι επιστήμονες πιστεύουν ότι το οικονόμο γονίδιο που κάποτε προστάτευε τους ανθρώπους από το λιμό μπορεί να συμβάλει στη διατήρηση των ανθυγιεινών ποσοτήτων λίπους τους.

Η διαγνωστική, δυαδικώς-εκτιμημένη μεταβλητή που ερευνάται είναι, εάν ο ασθενής παρουσιάζει συμπτώματα του διαβήτη, σύμφωνα με τα κριτήρια της Παγκόσμιας

Οργάνωσης Υγείας, δηλαδή, εάν η τιμή σακχάρου 2 ώρες μετά από φόρτιση γλυκόζης πλάσματος ήταν τουλάχιστον 200 mg / dl σε οποιαδήποτε έρευνα ή εξέταση ή αν διαπιστώθηκε κατά τη διάρκεια της συνήθους ιατρικής περίθαλψης.

Διάφοροι περιορισμοί είχαν τεθεί σχετικά με την επιλογή αυτών των περιπτώσεων από μια μεγαλύτερη βάση δεδομένων. Ειδικότερα, όλοι οι ασθενείς εδώ είναι θηλυκά τουλάχιστον 21 ετών της κληρονομικότητας των Ινδιάνων Pima .

Το σύνολο δεδομένων diabetes περιλαμβάνει 768 παραδείγματα που διακρίνονται σε (2) κατηγορίες (κλάσεις), Αρνητικοί και Θετικοί, οι οποίες κατανέμονται ως εξής:

Κατηγορία 0: (tested negative) 500 δεδομένα

Κατηγορία 1: (tested positive) 268 δεδομένα

Κάθε παράδειγμα περιλαμβάνει 8 χαρακτηριστικά που προκύπτουν από εργαστηριακούς ελέγχους και χρησιμοποιούνται στην πρόβλεψη του διαβήτη. Τα χαρακτηριστικά παίρνουν συνεχείς αριθμητικές τιμές και είναι τα ακόλουθα:

1: Αριθμός φορών που έμεινε έγκυος

2: Συγκέντρωση της γλυκόζης του πλάσματος μετά από 2 ώρες, σε μια από του στόματος δοκιμασία ανοχής γλυκόζης

3: Διαστολική πίεση αίματος (mm Hg)

4: Πάχος της δερματοπτυχής τρικέφαλου μυ (mm)

5: 2-ωρος ορός ινσουλίνης (mu U/ml)

6: Δείκτης Μάζας Σώματος (βάρος σε kg/(ύψος σε m)^2)

7: Γενεαλογική λειτουργία διαβήτη

8: Ηλικία (σε έτη)

9: Μεταβλητή κατηγορίας (0 ή 1)

Τα παραδείγματα έχουν την ακόλουθη μορφή:

α/α	Κατηγορία	Preg	Plas	Pres	Skin	Insu	Mass	Pedi	Age
1	1	6	148	72	35	0	33.6	0.627	50
2	0	1	85	66	29	0	26.6	0.351	31
3	1	8	183	64	0	0	23.3	0.672	32
4	0	1	89	66	23	94	28.1	0.167	21
5	1	0	137	40	35	168	43.1	2.288	33
6	0	5	116	74	0	0	25.6	0.201	30
....	...	...	....	....	....	....	....	....	....
768	...	...	....	....	....	....	....	....	....

4.3 Ο Αλγόριθμος ταξινόμησης δένδρων (classification tree)

Ο αλγόριθμος δημιουργεί δένδρα ταξινόμησης/απόφασης (classification/decision trees) σύμφωνα με όσα περιγράφηκαν στην παράγραφο 2.6.1 και υλοποιείται από την συνάρτηση classregtree του Matlab.

Σύνταξη

t = classregtree(X,y)

t = classregtree(X,y,param1,val1,param2,val2)

Περιγραφή

Η t = classregtree(X,y) δημιουργεί ένα δέντρο απόφασης (t) για την πρόβλεψη των y με βάση τις στήλες του X ο οποίος είναι ένας n x m πίνακας. Εάν το y είναι ένα διάνυσμα τιμών, η classregtree εκτελεί *παρεμβολή*. Εάν το y έχει κατηγορικές τιμές, η classregtree εκτελεί *ταξινόμηση*. Και στις 2 περιπτώσεις, το t είναι ένα δυαδικό δέντρο όπου κάθε κόμβος χωρίζεται με βάση τις τιμές μιας στήλης του X. Μη αποδεκτές τιμές στο X ή στο y λαμβάνονται ως ελλιπείς τιμές και δεν χρησιμοποιούνται στην ταξινόμηση/παρεμβολή.

Αξιολόγηση

Για την αξιολόγηση χρησιμοποιήθηκε η μέθοδος της 1θαπλής επαναληπτικής διασταύρωσης (10-fold cross-validation) κατά την οποία το σύνολο δεδομένων χωρίζεται σε 10 υποσύνολα ίσου μεγέθους από τα οποία τα 9 χρησιμοποιούνται για την δημιουργία του μοντέλου (του δένδρου) και το τελευταίο υποσύνολο χρησιμοποιείται για την δοκιμή (αξιολόγηση) του μοντέλου

Σύνταξη

c = cvpartition(y_targets,'k',10);

fun = @(xT,yT,xt,yt)(sum(~strcmp(yt,classify(xt,xT,yT))));

error_rate = sum(crossval(fun, y_patterns,y,'partition',c))/sum(c.TestSize)

Περιγραφή

Η διαδικασία της διασταύρωσης υλοποιείται με την συνάρτηση **crossval** που απαιτεί ως παράμετρο την συνάρτηση **strcmp** και **cvpartition** οι οποίες διαχωρίζουν τα δεδομένα y σε K(=10) διαφορετικά υποσύνολα με βάση την κλάση στην οποία ανήκουν (y_targets) ώστε να υπάρχει ομοιόμορφη κατανομή των κλάσεων στα υποσύνολα. Το μέσο σφάλμα από την επαναληπτική εκτέλεση της διαδικασίας αποθηκεύεται στην μεταβλητή **error_rate**.

4.3.1 Εφαρμογή του αλγορίθμου στο σύνολο δεδομένων Thyroid

% Επιλογή του συνόλου δεδομένων Thyroid

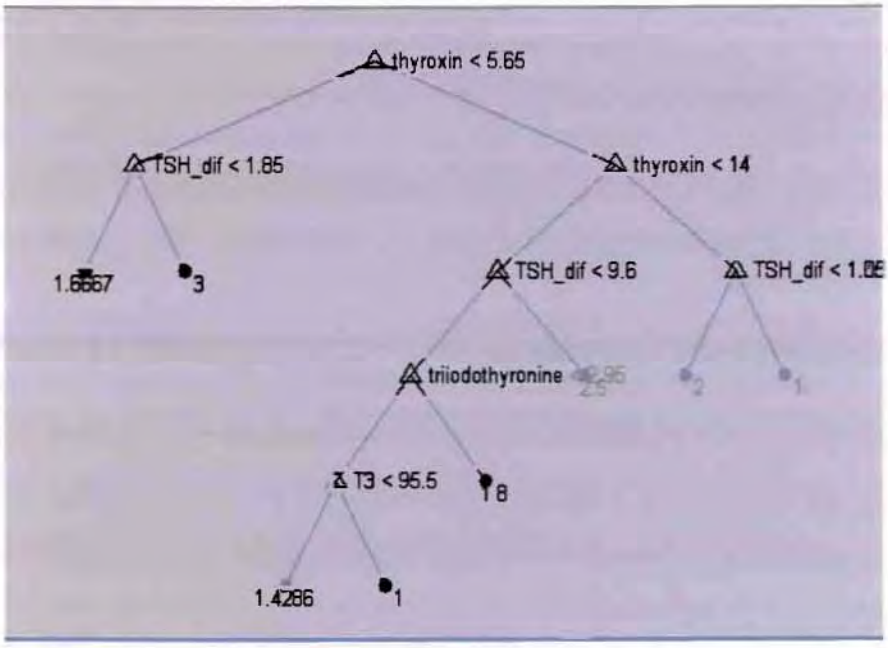
load thyroid;

% Κατασκευή του δένδρου ταξινόμησης

t = classregtree(thyroid_patterns,thyroid_targets,'names',{'T3' 'thyroxin' 'triiodothyronine' 'TSH' 'TSH_dif'})

% Απεικόνιση του δένδρου

view(t)



Thyroid

Αξιολόγηση:

% Cross-Validation

y = thyroid_targets;

c = cvpartition(y,'k',10);

fun = @(xT,yT,xt,yt)(sum(~strcmp(yt,classify(xt,xT,yT))));

error_rate = sum(crossval(fun,thyroid_patterns,y,'partition',c))/sum(c.TestSize)

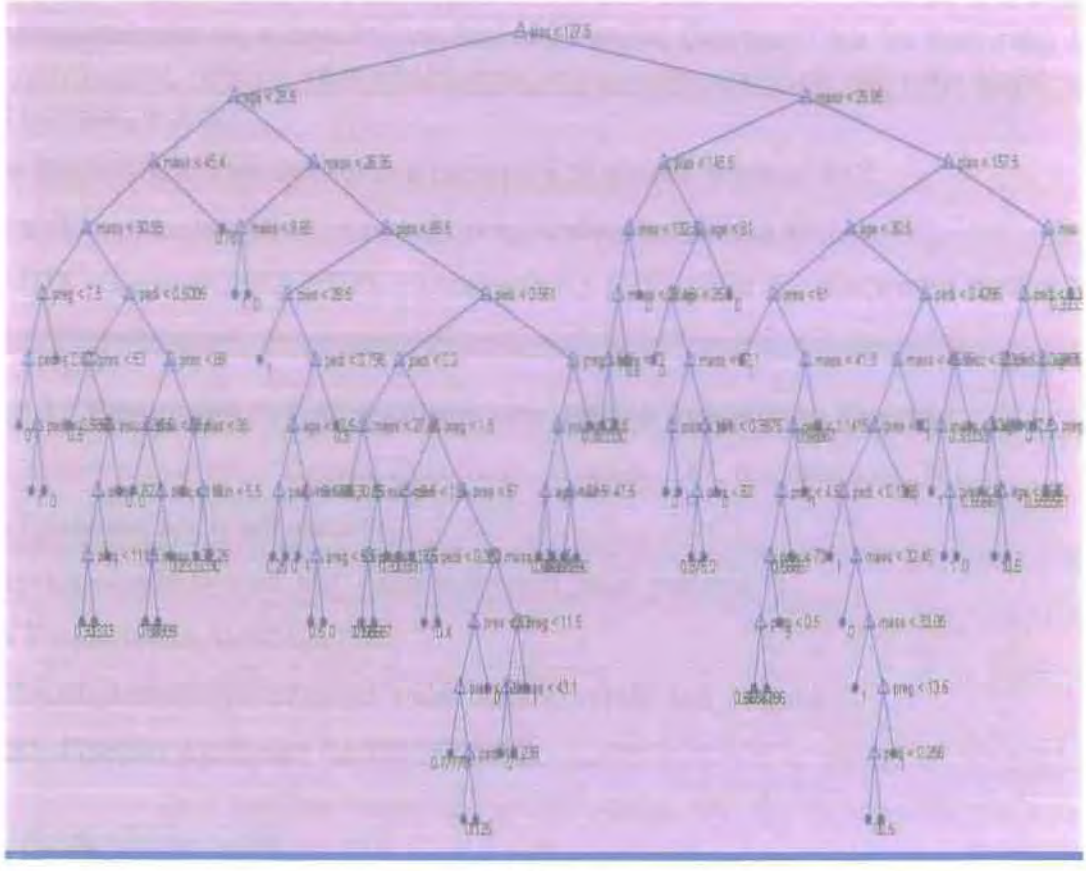
Αποτέλεσμα:

error_rate = 0.0465

Ακρίβεια= 1-error_rate= 0.9535 (95.35%)

4.3.2 Εφαρμογή του αλγορίθμου στο σύνολο δεδομένων Diabetes

```
% Επιλογή του συνόλου δεδομένων Diabetes
load diabetes;
% Κατασκευή του δένδρου ταξινόμησης
t = classregtree(diabetes_patterns,diabetes_targets,'names',{'preg' 'plas' 'pres' 'skin'
'insu' 'mass' 'pedi' 'age'})
% Απεικόνιση του δένδρου
view(t)
```



Diabetes

Αξιολόγηση:

```
% Cross-Validation
y = diabetes_targets;
c = cvpartition(y,'k',10);
fun = @(xT,yT,xt,yt)(sum(~strcmp(yt,classify(xt,xT,yT))));
error_rate = sum(crossval(fun,diabetes_patterns,y,'partition',c))/sum(c.TestSize)
```

Αποτέλεσμα:

error_rate = 0.0130

Ακρίβεια= 1-error_rate =0.9870 (98.7%)

4.4 Ο αλγόριθμος των k-κοντινότερων γειτόνων (k-Nearest Neighbors- KNN)

Ο γραμμικός K-NN [Yi Cao, 2008] είναι η απλούστερη προσέγγιση του KNN. Η αναζήτηση βασίζεται στον υπολογισμό όλων των αποστάσεων μεταξύ του ερωτήματος και των δεδομένων.

Σύνταξη:

IDX = knnsearch (Q, R, K)

Περιγραφή

Ο IDX = knnsearch (Q, R, K) ψάχνει το σύνολο δεδομένων R (πίνακας n x d που αντιπροσωπεύει τα n σημεία σε ένα d -διάστατο διάστημα) για να βρει τους K-κοντινότερους γείτονες κάθε ερωτήματος που αντιπροσωπεύεται από κάθε σειρά του Q (πίνακας m x d).

Τα αποτελέσματα αποθηκεύονται στον (m x d) πίνακα δεικτών, IDX.

Ο IDX = knnsearch (Q, R) παίρνει την προκαθορισμένη τιμή K=1.

Ο IDX = knnsearch (Q) ή IDX = knnsearch (Q, [], K) κάνει την αναζήτηση για R = Q.

4.4.1 Εφαρμογή του αλγορίθμου στο σύνολο δεδομένων Thyroid

% ==========Thyroid Data (train instances: 187, test instances: 28)

% Εκτέλεση του k-NN για k=1

idx=knnsearch(thyroid_test_patterns,thyroid_train_patterns,1);

% Υπολογισμός Ακρίβειας

calculateAccuracy(idx,thyroid_train_targets,thyroid_test_targets)

Αποτέλεσμα: Ακρίβεια= 0.9286 (92.86%)

% Εκτέλεση του k-NN για k=3

idx=knnsearch(thyroid_test_patterns,thyroid_train_patterns,3);

% Υπολογισμός Ακρίβειας

calculateAccuracy(idx,thyroid_train_targets,thyroid_test_targets)

Αποτέλεσμα: Ακρίβεια=0.8571 (85.71%)

% Εκτέλεση του k-NN για k=5

idx=knnsearch(thyroid_test_patterns,thyroid_train_patterns,5);

% Υπολογισμός Ακρίβειας

calculateAccuracy(idx,thyroid_train_targets,thyroid_test_targets)

Αποτέλεσμα: Ακρίβεια=0.9286 (92.86%)

% Εκτέλεση του k-NN για k=7

idx=knnsearch(thyroid_test_patterns,thyroid_train_patterns,7);

% Υπολογισμός Ακρίβειας

calculateAccuracy(idx,thyroid_train_targets,thyroid_test_targets)

Αποτέλεσμα: Ακρίβεια=0.9286 (92.86%)

% Εκτέλεση του k-NN για k=9

idx=knnsearch(thyroid_test_patterns,thyroid_train_patterns,9);

% Υπολογισμός Ακρίβειας

calculateAccuracy(idx,thyroid_train_targets,thyroid_test_targets)

Αποτέλεσμα: Ακρίβεια=0.9286 (92.86%)

4.4.2 Εφαρμογή του αλγορίθμου στο σύνολο δεδομένων Diabetes

% ==========Diabetes Data (train instances: 513, test instances: 255)

% Εκτέλεση του k-NN για k=1

idx=knnsearch(diabetes_test_patterns,diabetes_train_patterns,1);

% Υπολογισμός Ακρίβειας

calculateAccuracy(idx,diabetes_train_targets,diabetes_test_targets)

Αποτέλεσμα: Ακρίβεια=0.6353 (63.53%)

% Εκτέλεση του k-NN για k=3

idx=knnsearch(diabetes_test_patterns,diabetes_train_patterns,3);

% Υπολογισμός Ακρίβειας

calculateAccuracy(idx,diabetes_train_targets,diabetes_test_targets)

Αποτέλεσμα: Ακρίβεια=0.6941 (69.41%)

% Εκτέλεση του k-NN για k=5

idx=knnsearch(diabetes_test_patterns,diabetes_train_patterns,5);

% Υπολογισμός Ακρίβειας

calculateAccuracy(idx,diabetes_train_targets,diabetes_test_targets)

Αποτέλεσμα: Ακρίβεια=0.7176 (71.76%)

% Εκτέλεση του k-NN για k=7

idx=knnsearch(diabetes_test_patterns,diabetes_train_patterns,7);

% Υπολογισμός Ακρίβειας

calculateAccuracy(idx,diabetes_train_targets,diabetes_test_targets)

Αποτέλεσμα: Ακρίβεια=0.7529 (75.29%)

% Εκτέλεση του k-NN για k=9

idx=knnsearch(diabetes_test_patterns,diabetes_train_patterns,9);

% Υπολογισμός Ακρίβειας

calculateAccuracy(idx,diabetes_train_targets,diabetes_test_targets)

Αποτέλεσμα: Ακρίβεια=0.7529 (75.29%)

4.5 Ο Αλγόριθμος ομαδοποίησης των Κ-μέσων (k-means)

Η ομαδοποίηση των Κ-μέσων είναι μία μέθοδος διαχωρισμού. Η συνάρτηση kmeans (του Matlab) χωρίζει τα δεδομένα σε Κ αμοιβαία αποκλειόμενες ομάδες και επιστρέφει τον δείκτη της ομάδας στην οποία ανήκει κάθε δεδομένο.

Σύνταξη:

IDX = kmeans(X,k)

Περιγραφή

Ο k-means χωρίζει τα σημεία στον πίνακα X, n x ρ στοιχείων, σε Κ ομάδες. Αυτός ο επαναληπτικός διαχωρισμός ελαχιστοποιεί το άθροισμα, σε όλες τις ομάδες των αθροισμάτων των αποστάσεων κάθε σημείου από το μέσο διάνυσμα κάθε ομάδας. Οι σειρές του X αντιστοιχούν στα σημεία, οι στήλες αντιστοιχούν στις μεταβλητές. Ο k-means επιστρέφει ένα n x 1 διάνυσμα IDX που περιέχει τους δείκτες ομάδων κάθε σημείου. Εξ ορισμού, ο k-means χρησιμοποιεί Ευκλείδειες αποστάσεις.

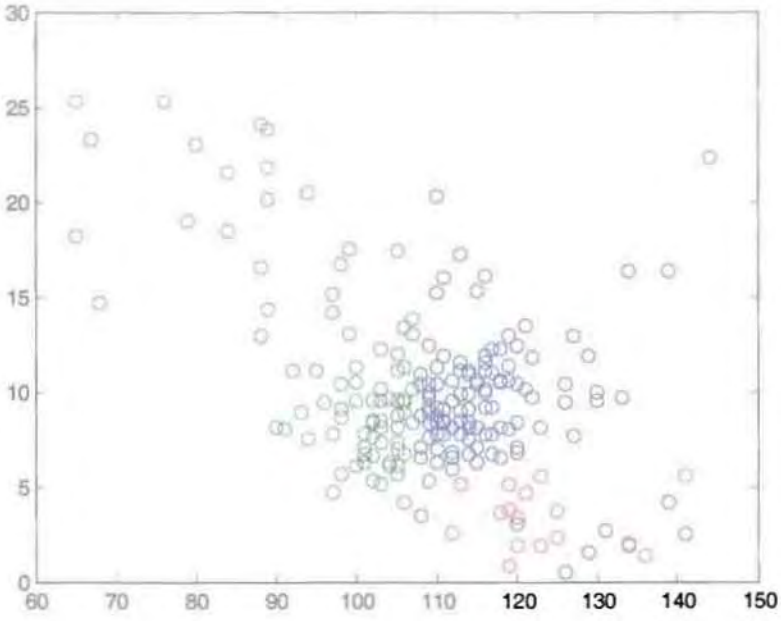
4.5.1 Εφαρμογή του αλγορίθμου στο σύνολο δεδομένων Thyroid

% Καλεί τον k-means για 3 ομάδες (αφού γνωρίζουμε ότι τα δεδομένα ανήκουν σε 3 κλάσεις)

IDX = kmeans(thyroid_patterns, 3)

% Απεικόνιση των αποτελεσμάτων

plot(thyroid_patterns(idx==1,1),thyroid_patterns(idx==1,2),'ro',thyroid_patterns(idx==2,1),thyroid_patterns(idx==2,2),'go',thyroid_patterns(idx==3,1),thyroid_patterns(idx==3,2),'bo')



Διάγραμμα των δεδομένων x1-x2

% Αξιολόγηση: Βρίσκουμε τις διαφορές για όλες τις περιπτώσεις αντιστοίχισης ομάδων με labels. Δηλαδή βλέπουμε αν ο K-means (στο διάνυσμα IDX) έχει δώσει στα δεδομένα που ανήκουν στην ίδια κλάση, το ίδιο label.

```
DIF1=IDX-thyroid_targets;
DIF2=IDX-thyroid_targets2;
DIF3=IDX-thyroid_targets3;
DIF4=IDX-thyroid_targets4;
DIF5=IDX-thyroid_targets5;
DIF6=IDX-thyroid_targets6;
```

% Παίρνουμε το μεγαλύτερο ποσοστό από όλες τις περιπτώσεις, δηλαδή τις μεγαλύτερες ομάδες των δεδομένων που ανήκουν στην ίδια κλάση.

```
Acc=max([sum(DIF1==0)./size(DIF1,1),
    sum(DIF2==0)./size(DIF2,1),
    sum(DIF3==0)./size(DIF3,1),
    sum(DIF4==0)./size(DIF4,1),
    sum(DIF5==0)./size(DIF5,1),
    sum(DIF6==0)./size(DIF6,1)])
```

% Αποτέλεσμα

Acc=0.6512 (65,12%)

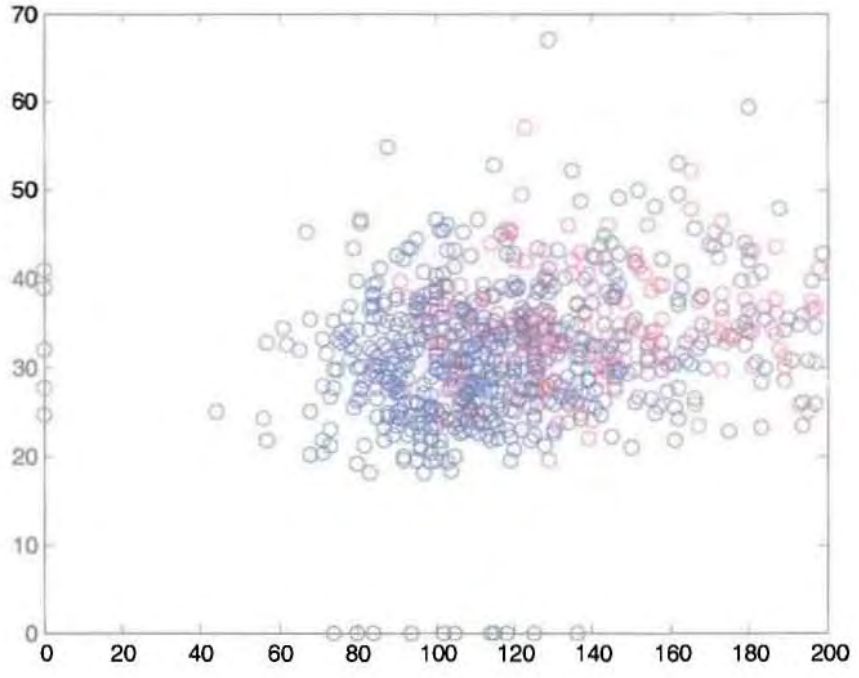
4.5.2 Εφαρμογή του αλγορίθμου στο σύνολο δεδομένων Diabetes

% Καλεί τον k-means για 2 ομάδες (αφού γνωρίζουμε ότι τα δεδομένα ανήκουν σε 2 κλάσεις)

IDX = kmeans(diabetes_patterns, 2)

% Απεικόνιση των αποτελεσμάτων

plot(diabetes_patterns(idx==1,2),diabetes_patterns(idx==1,6),'bo',diabetes_patterns(idx==2,2),diabetes_patterns(idx==2,6),'ro')



Διάγραμμα των δεδομένων x2-x6

% Αξιολόγηση: Υπολογίζουμε τη διαφορά των IDs της ομαδοποίησης με τα labels των targets (δηλαδή τις κλάσεις των δεδομένων) για να υπολογίσουμε την ακρίβεια της ομαδοποίησης (Το +1 το βάζουμε για υπολογιστικούς λόγους γιατί οι κλάσεις παίρνουν τιμές 0 και 1).

DIF=IDX-(diabetes_targets+1);

% Υπολογίζουμε την ακρίβεια της ομαδοποίησης (ποσοστό λανθασμένης ομαδοποίησης)

tempAcc=sum(abs(DIF))./size(DIF,1);

% Λαμβάνουμε υπόψη ότι μπορεί τα labels 1 να αντιστοιχίζονται (σωστά) στην ομάδα 2 αντί της ομάδας 1.

Acc=max(tempAcc,1-tempAcc)

% Αποτέλεσμα
Acc=0.6602 (66,02%)

4.6 Ο ασαφής αλγόριθμος των k-κοντινότερων γειτόνων (Fuzzy k-Nearest Neighbors-Fuzzy KNN)

Ο ταξινομητής Fuzzy KNN [Mertayak 2008] εφαρμόζεται σε δεδομένα των οποίων η κλάση στην οποία ανήκουν είναι ασαφής, δηλαδή τα δεδομένα ανήκουν σε όλες τις κλάσεις με έναν βαθμό συγγένειας.

Σύνταξη:

[y,predict_class] = f_knn(tr,tr_memberships,te,k)

Όπου:
tr: παραδείγματα εκπαίδευσης
tr_memberships: τιμές συγγένειας των δεδομένων (παραδειγμάτων) εκπαίδευσης
te: δεδομένα ελέγχου
k: Διάνυσμα των τιμών του αριθμού των γειτόνων k (περισσότερες από μια τιμές είναι δυνατές)
y: συναρτήσεις συγγένειας των δεδομένων ελέγχου, δηλαδή αυτών που θέλουμε να προβλέψουμε την τιμή.
predict_class: οι πιο πιθανές κλάσεις για τα δεδομένα ελέγχου (δηλαδή αυτές με την μεγαλύτερη τιμή συγγένειας)

Στην παρούσα πτυχιακή υλοποιήθηκαν δύο μέθοδοι για τον υπολογισμό των τιμών συγγένειας κάθε παραδείγματος με κάθε κλάση. Στην πρώτη υλοποίηση ο υπολογισμός έγινε βάσει της απόστασης κάθε παραδείγματος από τις μέσες τιμές των παραδειγμάτων κάθε κλάσης, ενώ στη δεύτερη βάσει του αλγορίθμου fcm που περιλαμβάνεται στο MATLAB για τον υπολογισμό των τιμών συγγένειας.

Ο αλγόριθμος εφαρμόστηκε μόνο στο σύνολο δεδομένων thyroid γιατί στο σύνολο δεδομένων diabetes η απόδοση ήταν εξαιρετικά χαμηλή γεγονός που αποδίδεται στη μεγάλη επικάλυψη μεταξύ των δεδομένων που ανήκουν στις 2 κλάσεις.

Α' Υλοποίηση

Το σύνολο δεδομένων thyroid διαιρέθηκε σε δύο σύνολα εκπαίδευσης και ελέγχου όπως απαιτεί ο fuzzy kNN. Ο διαχωρισμός έγινε με τρόπο που να διατηρείται η αναλογία των κλάσεων του αρχικού συνόλου και στα δύο νέα σύνολα. Το σύνολο train περιλαμβάνει 187 παραδείγματα και το σύνολο test 28 παραδείγματα. Στην προσέγγιση αυτή αρχικά υπολογίστηκαν οι μέσες τιμές των παραδειγμάτων (διανυσμάτων) κάθε κλάσης του συνόλου εκπαίδευσης.

Τα μέσα διανύσματα που προέκυψαν χρησιμοποιήθηκαν στη συνέχεια ως κέντρα των κλάσεων για τον υπολογισμό των τιμών συγγένειας κάθε παραδείγματος με κάθε κλάση (187x3). Ο υπολογισμός έγινε βάσει της Ευκλείδειας απόστασης, δηλαδή όσο

πιο κοντά βρίσκεται ένα παράδειγμα σε ένα κέντρο μιας κλάσης τόσο μεγαλύτερη είναι η τιμή συγγένειας για την αντίστοιχη κλάση.

Οι τιμές συγγένειας μπορούν να κυμαίνονται από 0 έως 1, ενώ το άθροισμα των τιμών συγγένειας ενός παραδείγματος για όλες τις κλάσεις είναι 1. Στη συνέχεια ο πίνακας συγγένειας που προέκυψε (thyroid_train_membership) χρησιμοποιήθηκε ως είσοδος στον αλγόριθμο f_knn. Το σύνολο train χρησιμοποιήθηκε για εκπαίδευση και το σύνολο test για αξιολόγηση.

```
% Υπολογισμός μέσων τιμών ανά κλάση του συνόλου εκπαίδευσης
thyroid_train_means=[mean(thyroid_train_patterns(thyroid_train_targets==1,:));
                     mean(thyroid_train_patterns(thyroid_train_targets==2,:));
                     mean(thyroid_train_patterns(thyroid_train_targets==3,:))];

% Υπολογισμός των συναρτήσεων συγγένειας (memberships)
thyroid_train_membership=zeros(187,3);
for k=1:187;
  s=0;
  for i=1:3;
    ssum=0;
    for j=1:5
      ssum=ssum+(thyroid_train_patterns(k,j)-thyroid_train_means(i,j))^2;
    end
    thyroid_train_membership(k,i)=1./(ssum.^0.5);
    s=s+thyroid_train_membership(k,i);
  end
  for i=1:3;
    thyroid_train_membership(k,i)=thyroid_train_membership(k,i)./s;
  end
end

% Εφαρμογή του fknn
k=1;    % αριθμός γειτόνων
[y,predict_class]=f_knn(thyroid_train_patterns,thyroid_train_membership,thyroid_tes
t_patterns,k);
% Υπολογισμός ακρίβειας, δηλαδή ποσοστό των σωστών (μεγαλύτερο βαθμό
συγγένειας) κλάσεων.
accuracy = sum(predict_class-thyroid_test_targets==0)/size(thyroid_test_targets,1)
```

Αξιολόγηση:

Εκτελέσαμε τον αλγόριθμο για διάφορες τιμές του K και προέκυψαν τα ακόλουθα αποτελέσματα:

k=1 : accuracy=0.8571
k=3 : accuracy=0.8571
k=5 : accuracy=0.8929
k=7 : accuracy=0.8929
k=9 : accuracy=0.8929

Β' Υλοποίηση

Στην προσέγγιση αυτή αρχικά υπολογίστηκαν τα κέντρα των κλάσεων και στη συνέχεια ο πίνακας συγγένειας (thyroid_train_membership) από τον αλγόριθμο fcm (Fuzzy c-means clustering). Οι τιμές συγγένειας και σε αυτήν την περίπτωση μπορούν να κυμαίνονται από 0 έως 1, ενώ το άθροισμα των τιμών συγγένειας ενός παραδείγματος για όλες τις κλάσεις είναι 1.

Από τη έξοδο του fcm λήφθηκε ο πίνακας συγγένειας (3x187) ο οποίος στη συνέχεια αναστράφηκε για να έρθει στην κατάλληλη μορφή (187x3). Ο πίνακας αυτός χρησιμοποιήθηκε ως είσοδος στον αλγόριθμο f_knn, όπου και πάλι το σύνολο train χρησιμοποιήθηκε για εκπαίδευση και το σύνολο test για αξιολόγηση.

```
% Υπολογισμός των συναρτήσεων συγγένειας (memberships)
[center, thyroid_train_membership, obj_fcn] = fcm(thyroid_train_patterns, 3, 1.5);

% Αναστροφή του πίνακα των memberships
thyroid_train_membership=thyroid_train_membership';

% Εφαρμογή του fknn
k=1;    % αριθμός γειτόνων
[y,predict_class]=f_knn(thyroid_train_patterns,thyroid_train_membership,thyroid_tes
t_patterns,k);
accuracy = sum(predict_class-thyroid_test_targets==0)/size(thyroid_test_targets,1)
```

Αξιολόγηση:

Εκτελέσαμε τον αλγόριθμο για διάφορες τιμές του K (1,3,5,7) στο σύνολο δεδομένων thyroid και προέκυψε σε όλες τις περιπτώσεις accuracy=0.8571.

4.7 Ο Αλγόριθμος ασαφούς ομαδοποίησης K-μέσων (Fuzzy c-means clustering)

Ο αλγόριθμος ομαδοποιεί τα δεδομένα έτσι ώστε κάθε δεδομένο να ανήκει σε κάθε ομάδα με κάποιο συγκεκριμένο βαθμό (συγγένειας). Στο Matlab ο αλγόριθμος υλοποιείται από τη συνάρτηση fcm.

Σύνταξη:

[center,U,obj_fcn] = fcm(data,cluster_n)

Περιγραφή:

Οι παράμετροι εισόδου είναι:

- data: Τα δεδομένα προς ομαδοποίηση
- cluster_n: Αριθμός των ομάδων (μεγαλύτερος του 1)

Η έξοδος της συνάρτησης είναι:

- center: πίνακας με τα τελικά κέντρα των ομάδων, όπου κάθε στήλη περιέχει τις συντεταγμένες των κέντρων.
- U: Τελικός πίνακας συναρτήσεων συγγένειας.
- obj_fcn: τιμές της αντικειμενικής συνάρτησης κατά τη διάρκεια των επαναλήψεων. Όταν η τιμή της μεταξύ 2 συνεχόμενων επαναλήψεων είναι μικρότερη από μια τιμή (η προκαθορισμένη είναι 1e-5) τότε η διαδικασία της ομαδοποίησης τερματίζει.

4.7.1 Εφαρμογή στο σύνολο δεδομένων Thyroid

Φορτώνουμε το σύνολο δεδομένων thyroid_noclass.dat όπου από το αρχικό σετ δεδομένων έχουμε αφαιρέσει τη γνωστή κλάση και εκτελούμε τον αλγόριθμο fcm ζητώντας την εύρεση τριών (3) κέντρων, δεδομένου ότι τόσες είναι οι γνωστές κλάσεις/ομάδες.

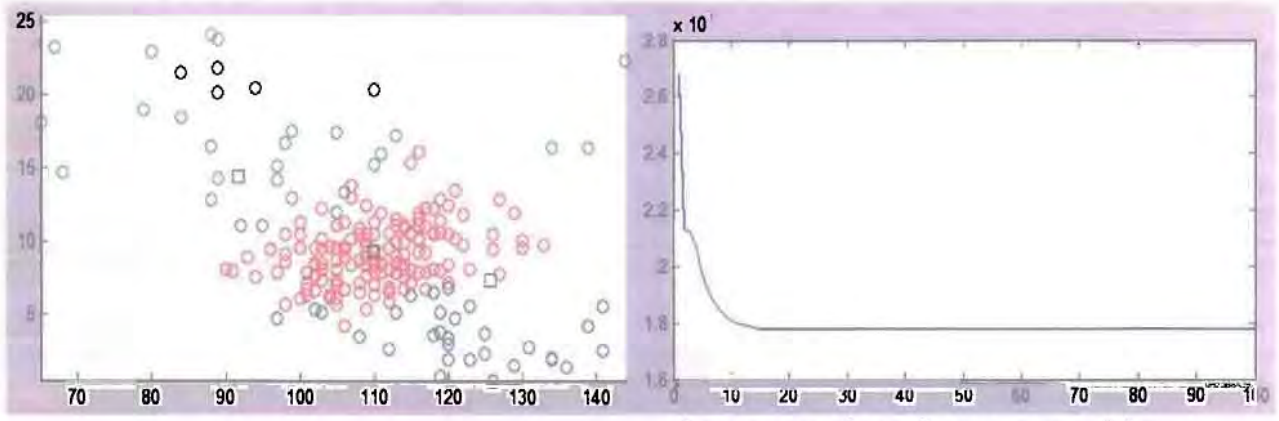
```
Data=load('thyroid_noclass.dat');
[Center, U, ObjFunction] = fcm(Data,3);
```

Τα κέντρα που υπολογίζονται είναι:

	x1	x2	x3	x4	x5
c1	109.9	9.3	1.7	1.6	2.8
c2	91.8	14.4	3.5	1.6	1.5
c3	125.7	7.3	1.7	6.6	8.3

Όπου x1-x5 είναι τα χαρακτηριστικά του συνόλου δεδομένων όπως αυτά περιγράφονται στην παράγραφο 4.2.1 (δηλ. x1=T3 κτλ.).

Στο διάγραμμα που ακολουθεί απεικονίζονται οι θέσεις των κέντρων σε διάγραμμα x1-x2 ως προς τα δεδομένα (αριστερά). Σύμφωνα με το αρχικό σύνολο δεδομένων, τα



κόκκινα σημεία ανήκουν στην κλάση 1, τα πράσινα στην 2 και τα μπλε στην 3. Στο διάγραμμα (δεξιά) βλέπουμε την εξέλιξη της τιμής της παραμέτρου ObjFunction (objective function) της συνάρτησης fcm.

Μετά την εκτέλεση της fcm ο πίνακας U περιέχει τις τιμές συγγένειας των δεδομένων στις 3 ομάδες που υπολογίστηκαν. Αν επιχειρήσουμε να αποασαφοποιήσουμε την κατηγορία στην οποία ανήκει κάθε παράδειγμα με βάση τη μέγιστη τιμή συγγένειας από τις 3 που του αναθέτει ο fcm τότε προκύπτει ότι σε σχέση με τις αρχικές γνωστές τιμές, το ποσοστό επιτυχίας είναι 80%.

4.7.2 Εφαρμογή στο σύνολο δεδομένων Diabetes

Φορτώνουμε το σύνολο δεδομένων diabetes_noclass.dat όπου από το αρχικό σετ δεδομένων έχουμε αφαιρέσει τη γνωστή κλάση και εκτελούμε τον αλγόριθμο fcm ζητώντας την εύρεση δυο (2) κέντρων, δεδομένου ότι τόσες είναι οι γνωστές κλάσεις/ομάδες.

```
Data=load('diabetes_noclass.dat');
[Center, U, ObjFunction] = fcm(Data,2);
```

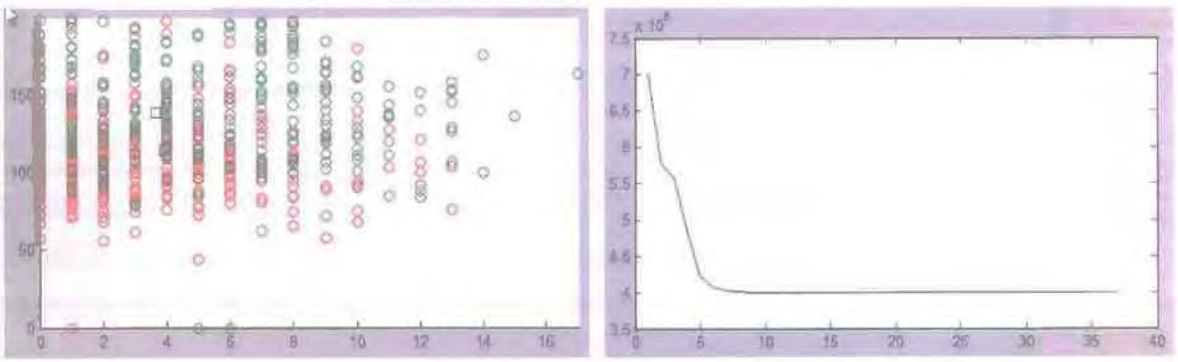
Τα κέντρα που υπολογίζονται είναι:

	x1	x2	x3	x4	x5	x6	x7	x8
c1	3.7	138.6	72.4	31.1	227.1	34.8	0.6	33.2
c2	3.9	114.3	68.3	16.0	23.5	30.9	0.4	33.4

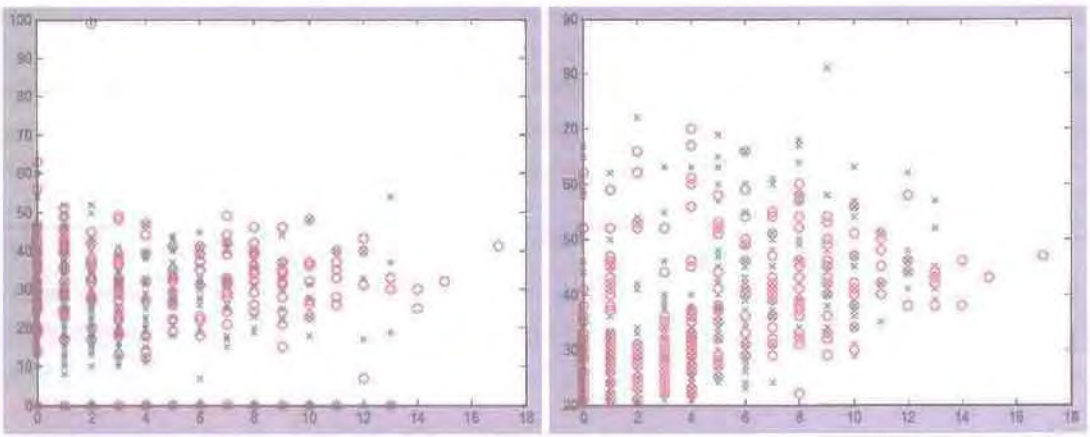
Όπου x1-x8 είναι τα χαρακτηριστικά του συνόλου δεδομένων όπως αυτά περιγράφονται στην παράγραφο 4.2.2 (δηλ. x1=Preg κτλ.).

Στο διάγραμμα που ακολουθεί απεικονίζονται οι θέσεις των κέντρων σε διάγραμμα x1-x2 ως προς τα δεδομένα (αριστερά). Σύμφωνα με το αρχικό σετ δεδομένων, τα κόκκινα σημεία ανήκουν στην κλάση 1, τα πράσινα στην 2. Στο διάγραμμα δεξιά

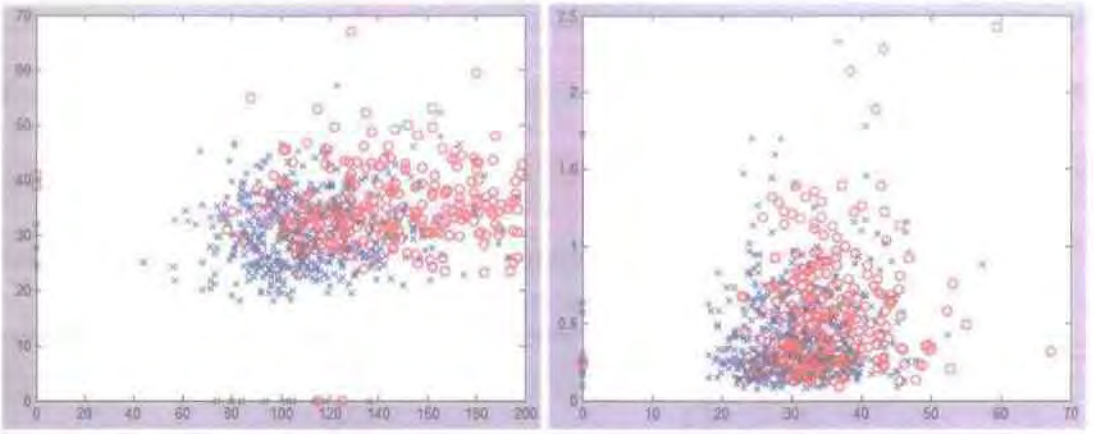
βλέπουμε την εξέλιξη της τιμής της παραμέτρου ObjFunction (objective function) της συνάρτησης fcm.



Μετά την εκτέλεση της fcm ο πίνακας U περιέχει τις τιμές συγγένειας των δεδομένων στις 2 ομάδες που υπολογίστηκαν. Αν επιχειρήσουμε να αποασαφοποιήσουμε την κατηγορία στην οποία ανήκει κάθε παράδειγμα με βάση τη μέγιστη τιμή συγγένειας από τις 2 που του αναθέτει ο fcm τότε προκύπτει ότι σε σχέση με τις αρχικές γνωστές τιμές, το ποσοστό επιτυχίας είναι 34.2%. Το χαμηλό ποσοστό οφείλεται καθαρά στο ότι οι ομάδες δεν είναι διακριτές (δηλαδή επικαλύπτονται) στο χώρο των ανεξάρτητων παραμέτρων x1..x8. Ενδεικτικά παρατίθενται τα διαγράμματα x1-x4, x1-x8, x2-x6, x6-x7 με χρωματική απεικόνιση των ομάδων.



Διαγράμματα x1-x4 και x1-x8 στο diabetes.dat



Διαγράμματα x2-x6 και x6-x7 στο diabetes.dat

4.8 Ασαφείς Κανόνες

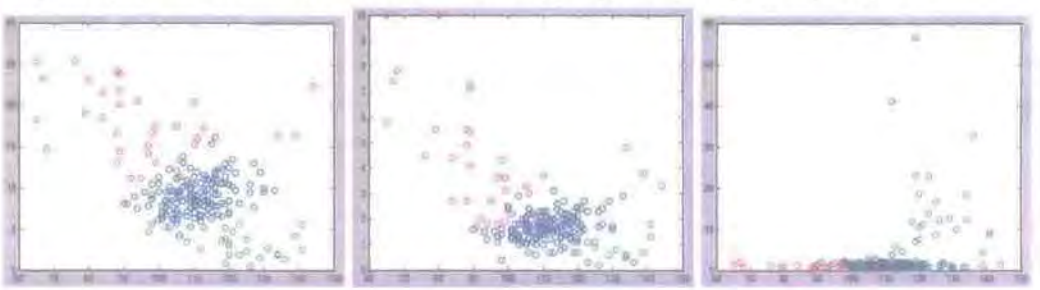
Η ασαφής εξαγωγή συμπερασμάτων είναι μια διαδικασία απεικόνισης μεταξύ δεδομένων εισόδου/εξόδου (δηλαδή των ανεξάρτητων και της εξαρτημένης μεταβλητής) χρησιμοποιώντας ασαφή λογική. Η απεικόνιση που παράγεται παρέχει έναν μηχανισμό λήψης απόφασης (εύρεση κλάσης ή αριθμητικής τιμής). Η διαδικασία περιλαμβάνει ασαφοποίηση των δεδομένων εισόδου (δηλαδή ορισμό συναρτήσεων συγγένειας), συνδυασμό των δεδομένων εισόδου με λογικούς τελεστές (π.χ. AND) και ορισμό/παραγωγή IF-THEN κανόνων οι οποίοι τελικά συνιστούν ένα Fuzzy Inference System (FIS).

Το Matlab περιλαμβάνει βιβλιοθήκη συναρτήσεων που υποστηρίζει τα παραπάνω καθώς και 2 διεπαφές χρήστη για την ευκολότερη χρήση της. Η πρώτη διεπαφή (εντολή fuzzy) επιτρέπει την πλήρη υλοποίηση ενός FIS χειροκίνητα ενώ η δεύτερη (εντολή anfisedit) επιτρέπει την αυτόματη παραγωγή ενός FIS από δεδομένα.

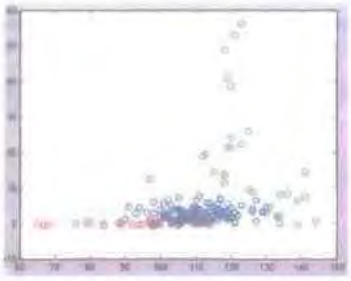
4.8.1 Εφαρμογή στο σύνολο δεδομένων Thyroid

Όπως προαναφέρθηκε το σύνολο δεδομένων thyroid.dat περιλαμβάνει 215 εγγραφές. Κάθε εγγραφή έχει 5 παραμέτρους εισόδου (x_1-x_5) και 1 εξόδου (τελευταία στήλη). Η στήλη εξόδου μπορεί να πάρει 3 τιμές (1, 2, 3) που είναι και οι 3 κλάσεις.

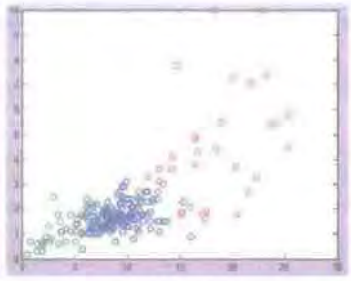
Χρησιμοποιώντας τις δυνατότητες γραφικής απεικόνισης του matlab (εντολή plot) διερευνήθηκαν τα δεδομένα ανά δύο σε σχέση με την παράμετρο εξόδου.



$x_1 - x_2$ $x_1 - x_3$ $x_1 - x_4$

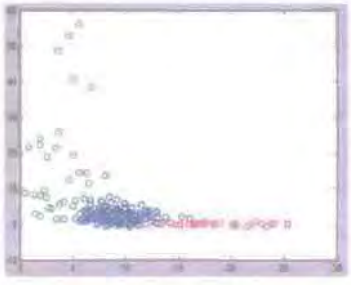


$x_1 - x_5$

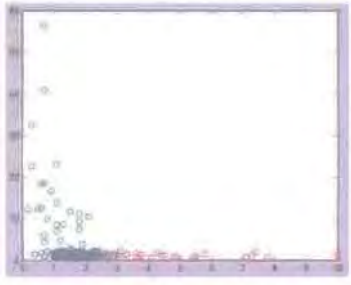


$x_2 - x_3$

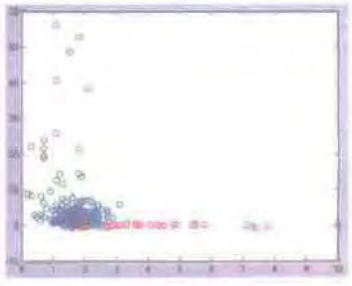
$x_2 - x_4$



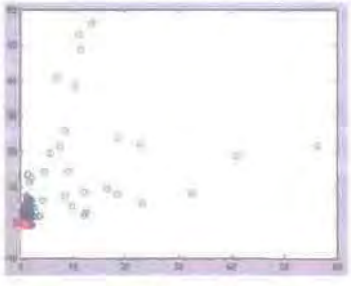
$x_2 - x_5$



$x_3 - x_4$



$x_3 - x_5$



$x_4 - x_5$

Σε διερεύνηση των δεδομένων φάνηκε ότι όντως υπάρχουν 3 κλάσεις "καλά" εντοπισμένες στο χώρο των 5 διαστάσεων (5 input) όπως φαίνεται και από τα γραφήματα στα οποία απεικονίζονται οι κλάσεις (red=1, green=2, blue=3) σε γράφημα 2 εκ των 5 input. Επίσης φαίνεται ότι γενικά υπάρχουν 3 ζώνες τιμών για input, με κάθε μία να αφορά σε μία κλάση. Αυτό κάνει τη διαδικασία ασαφοποίησης των input σχετικά εύκολη.

Στη συνέχεια τα δεδομένα χωρίστηκαν με τυχαίο τρόπο σε δεδομένα εκπαίδευσης (training set) (thyroid_train.dat, 187 records) και δεδομένα ελέγχου (testing set) (thyroid_test.dat, 28 records) φροντίζοντας μόνο να κρατηθούν οι αναλογίες των κλάσεων του αρχικού σετ δεδομένων. Τα πρώτα χρησιμοποιήθηκαν για την κατασκευή ενός FIS και τα δεύτερα για την αξιολόγηση του.

Σε κάθε διάσταση input, τα δεδομένα εκπαίδευσης χωρίστηκαν σε 3 ομάδες, βάσει της τιμής του output. Η ασαφοποίηση έγινε χρησιμοποιώντας τη Γκαουσιανή συνάρτηση ως συνάρτηση συγγένειας στα διάφορα fuzzy sets που προέκυψαν. Η συνάρτηση αυτή έχει τη μορφή:

$$\mu(x) = e^{\frac{(x-m)^2}{\sigma^2}}$$

Σε κάθε ομάδα και για κάθε διάσταση input, υπολογίστηκαν η μέση τιμή m και η τυπική απόκλιση σ των τιμών των input. Οι τιμές δίνονται στους πίνακες που ακολουθούν:

Πίνακας 1: Τιμές μέσης τιμής m της ασαφοποίησης.

m	x1	x2	x3	x4	x5
class 1	110.4	9.1	1.7	1.3	2.4
class 2	***94.3***	17.8	4.2	1.0	0.0
class 3	120.8	3.7	1.1	11.6	17.3

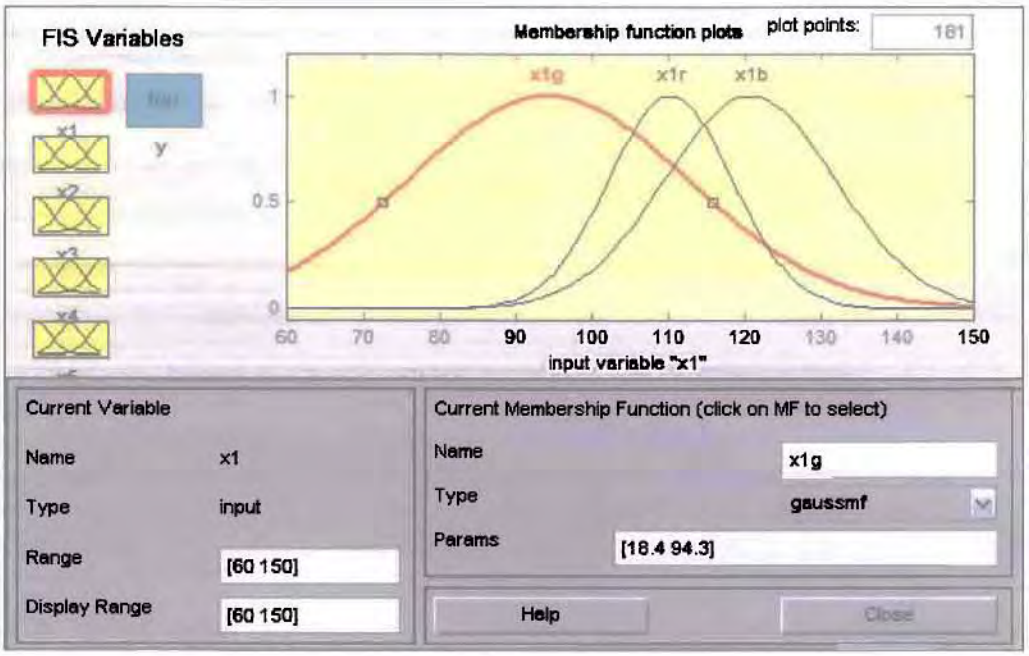
Πίνακας 2: Τιμές τυπικής απόκλισης σ της ασαφοποίησης.

σ	x1	x2	x3	x4	x5
class 1	8.1	2.0	0.5	0.5	1.7
class 2	***18.4***	4.1	2.4	0.4	0.3
class 3	11.2	1.8	0.6	12.5	14.7

Παράδειγμα αναλυτικού υπολογισμού:

Έστω η πρώτη παράμετρος x_1. Χωρίζουμε τα δεδομένα thyroid_train.dat σε 3 ομάδες βάσει της τιμής της κλάσης (output). Οι πληθυσμοί των ομάδων είναι: 130 εγγραφές της κλάσης 1, 31 εγγραφές της κλάσης 2 και 26 εγγραφές της κλάσης 3 (αυτός ο διαχωρισμός θα χρησιμοποιηθεί σε όλα τα input, επειδή όπως ειπώθηκε, οι κλάσεις είναι καλά προσδιορισμένες στο χώρο των input).

Η ασαφοποίηση των x_1 φαίνεται εποπτικά στο επόμενο γράφημα που παράγει ο FIS editor του Matlab, κατά την κωδικοποίηση σε αυτό (για την κατασκευή του FIS) των ασαφών συνόλων που προέκυψαν.



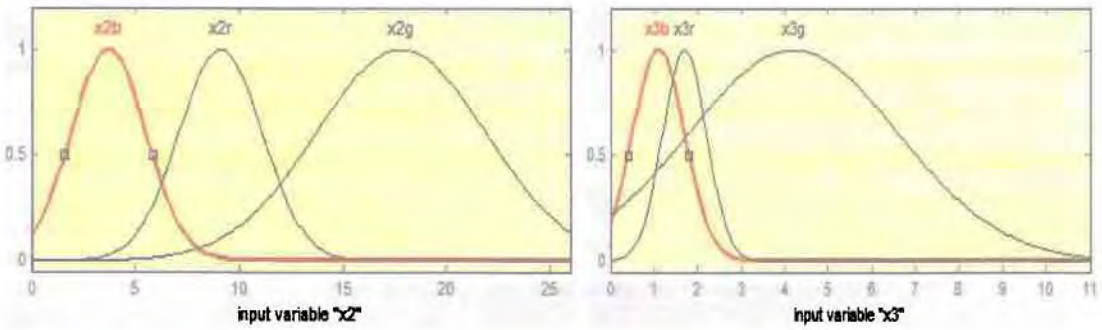
Εικόνα 3: Οι συναρτήσεις συμμετοχής (συγγένειας) για τις ασαφείς τιμές της x_1.

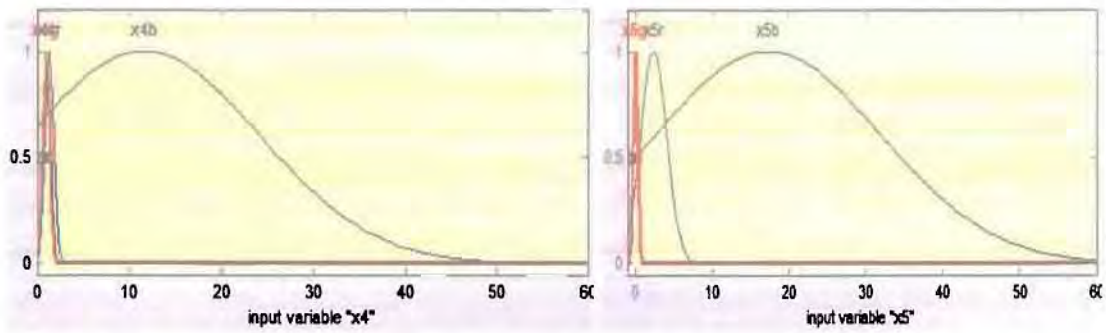
Ειδικότερα απεικονίζονται οι 3 συναρτήσεις συγγένειας των τριών ασαφών συνόλων που ορίσαμε και τις οποίες θα μπορούσαμε λεκτικά να ονομάσουμε:

- (μέσες προς) μικρές τιμές x_1 (καμπύλη x1g)
- μέσες τιμές x_1 (καμπύλη x1r)
- (μέσες προς) μεγάλες τιμές x_1 (καμπύλη x1b)

Στην Εικόνα 3 στο πεδίο Params φαίνονται οι τιμές σ=18.4 και m=94.3 που υπολογίστηκαν πριν (Πίνακες 1 & 2).

Με παρόμοια διαδικασία ασαφοποιούνται (και κωδικοποιούνται) και οι υπόλοιπες διαστάσεις input. Οι σχετικές γραφικές παραστάσεις δίνονται στη συνέχεια. Τα b, r και g που φαίνονται στα ονόματα των καμπύλων μπήκαν με βάση την κλάση y με την οποία σχετίζονται καθώς οι κλάσεις, ονομάστηκαν red (y=1), green (y=2), blue (y=3).





Εικόνα 4: Οι συναρτήσεις συγγένειας για τις ασαφείς τιμές των x_2, x_3, x_4 και x_5

Στη συνέχεια κωδικοποιήθηκαν οι κανόνες στον FIS editor. Επειδή κάθε διάσταση input ασαφοποιήθηκε με 3 λεκτικές τιμές (π.χ. η διάσταση x_1 παίρνει τιμές x_{1r}, x_{1g} και x_{1b}) και επειδή οι 3 κλάσεις είναι καλά οριοθετημένες στο χώρο των input, οι κανόνες που προκύπτουν είναι προφανείς. Για παράδειγμα, τα δεδομένα που έχουν κλάση class1R έχουν input $X_1=x_{1r}$, $X_2=x_{2r}$, $X_3=x_{3r}$, $X_4=x_{4r}$, $X_5=x_{5r}$. Άρα προκύπτει ο κανόνας:

if (X_1 is x_{1r}) **and** (X_2 is x_{2r}) **and** (X_3 is x_{3r}) **and** (X_4 is x_{4r}) **and** (X_5 is x_{5r})
then (Y is class1R)

Ομοίως οι άλλοι 2 οι κανόνες που προκύπτουν είναι οι ακόλουθοι:

if (X_1 is x_{1g}) **and** (X_2 is x_{2g}) **and** (X_3 is x_{3g}) **and** (X_4 is x_{4g}) **and** (X_5 is x_{5g})
then (Y is class1G)

if (X_1 is x_{1b}) **and** (X_2 is x_{2b}) **and** (X_3 is x_{3b}) **and** (X_4 is x_{4b}) **and** (X_5 is x_{5b})
then (Y is class1B)

Για την αξιολόγηση του FIS που δημιουργήθηκε χρησιμοποιήθηκε το σχετικό παράθυρο του FIS editor. Στο πεδίο input δόθηκαν πεντάδες [x_1, x_2, x_3, x_4, x_5] του αρχείου thyroid_test.dat και καταγράφηκε η τιμή που υπολογίζει το FIS. Συγκρίνοντας τις υπολογιζόμενες με τις γνωστές τιμές για την κλάση στην οποία ανήκουν οι εγγραφές υπολογίστηκε ποσοστό επιτυχούς πρόβλεψης 82.17%.

Η μέθοδος της αφαιρετικής ομαδοποίησης (subtractive clustering) δεν έδωσε καλά αποτελέσματα για τα κέντρα των κλάσεων. Ως επακόλουθο, η αυτόματη δημιουργία FIS μέσω anfisedit επίσης δεν έδωσε καλά αποτελέσματα. Χρησιμοποιώντας grid partitioning στο anfisedit, προέκυψε ένα FIS με 243 κανόνες, αποτέλεσμα του απλού τρόπου με τον οποίο φτιάχνει τα clusters η μέθοδος grid partitioning. Με τόσους πολλούς κανόνες σε τέτοιου μεγέθους dataset οδηγηθήκαμε στο συμπέρασμα ότι προέκυψαν φαινόμενα υπερμοντελοποίησης (τόσοι κανόνες όσα και τα δεδομένα).

4.8.2 Εφαρμογή στο σύνολο δεδομένων Diabetes

Διερευνώντας τη δυνατότητα κατασκευής FIS για το συγκεκριμένο σύνολο δεδομένων προέκυψε ότι ήταν πρακτικά αδύνατο να οριστούν ασαφείς τιμές για τα δεδομένα εισόδου. Αυτό οφείλεται στον μεγάλο βαθμό επικάλυψης μεταξύ των ομάδων που οδήγησε σε χαμηλή απόδοση τη συνάρτηση fcm αλλά και σε πολύ μεγάλο πλήθος ομάδων στις μεθόδους ομαδοποίησης που χρησιμοποιεί το anfisedit. Ως αποτέλεσμα προέκυπτε πολύπλοκη ασαφοποίηση που οδηγούσε σε τεράστιο πλήθος κανόνων με αποτέλεσμα τη δυσλειτουργία του Matlab.

Κεφάλαιο 5

Συμπεράσματα και Μελλοντική Εργασία

Σκοπός της εργασίας αυτής ήταν η εφαρμογή της διαδικασίας Ανακάλυψης Γνώσης (KDD) σε ιατρικά δεδομένα. Για τον σκοπό αυτό εφαρμόστηκαν αλγόριθμοι μηχανικής μάθησης και ανακάλυψης γνώσης ασαφούς και μη, λογικής. Από τα διαθέσιμα εργαλεία, χρησιμοποιήθηκε το εργαλείο Matlab που προσφέρει πολλές δυνατότητες και χρησιμοποιείται ευρέως τόσο από την επιστημονική όσο και την επιχειρηματική κοινότητα. Μετά την υλοποίηση ακολούθησε η αξιολόγηση και η σύγκριση της απόδοσης των μεθόδων και της ποιότητας των συμπερασμάτων που προέκυψαν.

Στην παρούσα εργασία, χρησιμοποιήθηκαν δύο σύνολα δεδομένων από το UCI machine learning repository (αποθετήριο του University of California Irvine), το "Thyroid Data set" και το "Pima Indians Diabetes Data set".

Όσον αφορά στους αλγορίθμους επιλέχθηκαν 2 αλγόριθμοι ταξινόμησης (ο αλγόριθμος ταξινόμησης δένδρων και ο αλγόριθμος των K-κοντινότερων γειτόνων k-nn), ένας αλγόριθμος ομαδοποίησης (των κ-μέσων) και 3 αλγόριθμοι ασαφούς μάθησης, εκ των οποίων 2 ταξινόμησης (των K-κοντινότερων γειτόνων και οι ασαφείς κανόνες) και ένας αλγόριθμος ασαφούς ομαδοποίησης (ο F-cmean).

Συμπεράσματα

Η συγκριτική απόδοση των αλγορίθμων στα δύο σύνολα δεδομένων (thyroid, diabetes) που χρησιμοποιήθηκαν στην παρούσα πτυχιακή εργασία απεικονίζεται στον ακόλουθο πίνακα:

	Thyroid	Diabetes
Δένδρο Απόφασης	95	98
KNN	92	75
K-means	65,15	66,02
Fuzzy KNN	89 και 85	-
Fuzzy k-means	80	34,02
Fuzzy Rules	82	-

Όπως φαίνεται, τα δένδρα απόφασης πέτυχαν καλύτερη ακρίβεια πρόβλεψης (95% και 98% αντίστοιχα) από τον αλγόριθμο των k κοντινότερων γειτόνων (92% και 75% οι καλύτερες επιδόσεις αντίστοιχα). Αυτό οφείλεται στην αδυναμία του τελευταίου αλγορίθμου να δώσει διαφορετική βαρύτητα σε κάθε χαρακτηριστικό (feature) του συνόλου δεδομένων. Έτσι οδηγείται σε μειωμένη απόδοση όταν υπάρχουν χαρακτηριστικά που είναι λιγότερο σχετικά με το πρόβλημα ταξινόμησης που αντιμετωπίζεται. Αντίθετα, ο αλγόριθμος κατασκευής δένδρων απόφασης χρησιμοποιώντας το κέρδος πληροφορίας (information gain) είναι σε θέση να αντιμετωπίσει καλύτερα το παραπάνω πρόβλημα, είτε χρησιμοποιώντας τα λιγότερο σχετικά χαρακτηριστικά σε χαμηλότερους κόμβους του δέντρου είτε απορρίπτοντάς τα τελείως.

Επιπλέον η απόδοση του ασαφούς αλγορίθμου k κοντινότερων γειτόνων (fuzzy k-NN) έναντι του απλού είναι συγκρίσιμη (89% και 85% στις 2 υλοποιήσεις αντίστοιχα) στο σύνολο thyroid. Αντίστοιχη συμπεριφορά, στο σύνολο thyroid, παρουσιάζεται και στην ομαδοποίηση, όπου ο ασαφής αλγόριθμος ομαδοποίησης fcm, οδηγεί σε καλύτερους διαχωρισμούς (επιτυχία 80%) έναντι του απλού k-means (επιτυχία 65,12%). Αντίθετα όμως, στο σύνολο δεδομένων diabetes οι ασαφείς αλγόριθμοι δεν μπόρεσαν να εφαρμοστούν (είχαν εξαιρετικά χαμηλή απόδοση, ο fcm είχε απόδοση 34%) επειδή δεν υπήρχε καλός χωρικός διαχωρισμός των κλάσεων ως προς τα δεδομένα εισόδου και δεν μπόρεσε να γίνει καλή ασαφοποίηση τους.

Γενικά όταν οι διαφορετικές κλάσεις αλληλεπικαλύπτονται σε κάποιο βαθμό τότε είναι προτιμότεροι γιατί συλλαμβάνουν καλύτερα αυτόν τον μη αυστηρό ορισμό των κλάσεων και αποδίδουν καλύτερα. Αυτές οι περιπτώσεις είναι οι πλέον συνηθισμένες ειδικά στον χώρο της ιατρικής. Για παράδειγμα, κάποιος δεν είναι απόλυτα ασθενής ή απόλυτα υγιής από μια ασθένεια (π.χ. του θυρεοειδή), αλλά έχει είτε λιγότερο είτε περισσότερο αυτήν την ασθένεια.

Όσον αφορά την μοντελοποίηση με ασαφείς κανόνες, και αυτή είναι εφικτή όταν υπάρχει καλός χωρικός διαχωρισμός των κλάσεων ως προς τα δεδομένα εισόδου γιατί επιτρέπει την καλή ασαφοποίηση τους. Για αυτό και δεν μπόρεσε να εφαρμοστεί στο σύνολο diabetes. Το μοντέλο που παράγεται (3 κανόνες στο thyroid data set) παρέχει την πιο διαισθητική παράσταση γνώσης. Οι κανόνες προσομοιάζουν σε μεγάλο βαθμό στις λογικές προτάσεις και τον μηχανισμό απόφασης και εξαγωγής συμπερασμάτων του ανθρώπου.

Η απόδοση (ακρίβεια) του μοντέλου (στο thyroid) είναι σε ικανοποιητικά επίπεδα (82%) και συγκρίσιμη με τις άλλες μεθόδους ταξινόμησης.

Επίλογος

Η διαδικασία της ανακάλυψης γνώσης από βάσεις δεδομένων αποτελεί ένα χρήσιμο και ενδιαφέρον επιστημονικό πεδίο, συνεχώς εξελισσόμενο καθώς αυξάνεται σημαντικά το πλήθος και η ποιότητα των υλοποιούμενων αλγορίθμων και των εργαλείων ανακάλυψης γνώσης.

Ο ευαίσθητος τομέας της υγείας αποτελεί σημαντικό πεδίο εφαρμογής της διαδικασίας αυτής. Το γεγονός αυτό ενισχύεται ακόμη περισσότερο από το ολοένα αυξανόμενο πλήθος και την ποιότητα των ιατρικών δεδομένων που υπάρχουν συγκεντρωμένα, χειρόγραφα και λιγότερο ηλεκτρονικά, αλλά και από το γεγονός της μεταβλητής φύσης της ιατρικής πράξης που προσαρμόζεται στην μοναδική περίπτωση του κάθε ασθενή. Περιοριστική είναι όμως η ιδιαιτερότητα του τομέα της υγείας, καθώς πρόκειται για αυστηρά ευαίσθητα προσωπικά δεδομένα που χρήζουν ιδιαίτερης αντιμετώπισης. Μάλιστα, όπως συμβαίνει σε όλες τις περιπτώσεις ανακάλυψης γνώσης, η συμβολή του ειδικού του τομέα είναι απαραίτητη για την αξιολόγηση των συμπερασμάτων που προκύπτουν.

Μελλοντικές Εφαρμογές

Οι τεράστιες δυνατότητες που προσφέρονται από την εφαρμογή των αλγορίθμων εξόρυξης γνώσης μπορούν να φανούν χρήσιμες στον τομέα της υγείας για την παροχή ποιοτικότερων υπηρεσιών περίθαλψης στους ασθενείς και αποδοτικότερη διοικητικό-οικονομική διαχείριση των παρόχων των υπηρεσιών αυτών.

Τα συμπεράσματα που προέκυψαν από την εργασία αναφέρονται σε πραγματικά ιατρικά δεδομένα ασθενών που πάσχουν από σακχαρώδη διαβήτη και λαμβάνουν ή όχι ινσουλίνη ως μέσο θεραπείας και μπορούν να εφαρμοστούν στην καθημερινή ιατρική πράξη για τη γρήγορη κατηγοριοποίηση των ασθενών ώστε να προσφέρεται πιο άμεση και σωστή περίθαλψη. Παράλληλα, χρειάζεται και η συλλογή περισσότερων χαρακτηριστικών ή και περιπτώσεων για εξαγωγή νέων συμπερασμάτων καθώς σημαντικό εξακολουθεί να είναι το πρόβλημα της διάθεσης επαρκών δεδομένων προς ανάλυση.

Μια σημαντική επέκταση στη διαδικασία αυτή θα ήταν οι συμμετοχή των ίδιων των ασθενών που πάσχουν από τη νόσο και λαμβάνουν ή όχι ινσουλίνη, στην συλλογή των αντίστοιχων δεδομένων μέσω των δυνατοτήτων που προσφέρει το διαδίκτυο, τόσο για τη διατήρηση πληρέστερου ιστορικού τους, αλλά και την καλύτερη απόδοση των αλγορίθμων και τη μελέτη στοιχείων που δεν λήφθηκαν υπόψη στην ανάλυση λόγω μη επαρκών δεδομένων.

Βιβλιογραφία

[Berka et al, 2009] Petr Berka, Jan Rauch, Djamel Abdelkader Zighed, Data Mining and Medical Knowledge Management: Cases and Applications, Medical Information Science Reference; 2009.

[Berry and Gordon, 2004] Michael J. A. Berry and Gordon S. Linoff , Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management, Wiley Computer Publishing; 2004.

[Chen and Lonardi, 2009] Jake Y. Chen, Stefano Lonardi (Eds), Biological Data Mining, Chapman & Hall/CRC; 2009.

[Chiu 1994] Chiu, S., "Fuzzy Model Identification Based on Cluster Estimation," Journal of Intelligent & Fuzzy Systems, Vol. 2, No. 3, 1994.

[Connolly and Begg, 2004] Thomas M. Connolly and Carolyn E. Begg, DataBase Systems: A Practical Approach to Design, Implementation and Management (4th Edition), Addison Wesley, 2004.

[Cover and Hart, 1967] T.M. Cover and P.E. Hart, Nearest neighbour pattern classification, *IEEE Trans. Inform. Theory* **IT-13** (1967), pp. 21–27.

[Dunham, 2002], Margaret H. Dunham, Data Mining: Introductory and Advanced Topics, Prentice Hall; 2002 (και μετάφραση από τις Εκδόσεις Νέων Τεχνολογιών, 2004).

[Ghahramani 2004] Zoubin Ghahramani, Unsupervised Learning, Advanced Lectures on Machine Learning, LNAI 3176, Springer-Verlag, 2004.

[Giannotti and Pedreschi, 2008] Giannotti Fosca and Pedreschi Dino (Eds.), Mobility, Data Mining and Privacy, Springer, 2008.

[Jain and Dubes, 1988] Anil K. Jain and Richard C. Dubes, Algorithms for Clustering Data, Prentice Hall 1988.

[Keller et al., 1985] J.M. Keller, M.R. Gray and J.A. Givens, A fuzzy k-nearest neighbours algorithm, *IEEE Trans. Syst. Man Cybern.* **15**, pp. 580–585, 1985.

[Kohavi and Provost, 2001] Ronny Kohavi and Foster Provost, (Eds), Applications of Data Mining to Electronic Commerce, Springer, 2001.

[Kotsiantis 2007] S. B. Kotsiantis, Supervised Machine Learning: A Review of ClassificationTechniques, Informatica 31, 249-268, 2007.

[MacQueen 1967] MacQueen J. B., Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1:281-297, 1967.

[Maimon and Rokach 2005] Oded Maimon and Lior Rokach (Eds), Data Mining and Knowledge Discovery Handbook, Springer; 2005.

[Mattison 1997] Rob Mattison, Data Warehousing and Data Mining for Telecommunications, Artech House Publishers, 1997.

[Mertayak 2008] Cuneyt Mertayak Fuzzy k-NN, Matlab central, http://www.mathworks.com/matlabcentral/fileexchange/21326, 2008.

[Michalski 1983] Michalski, R.S., A Theory and Methodology of Inductive Learning, in Machine Learning-An Artificial Intelligence Approach, Vol. I, Los Altos, CA:Morgan Kaufman, 83-134, 1983.

[Michalski et.al 1986] R.S. Michalski, S. Amarel, D.B. Lenat, D. Michie, and P. Winston. Machine learning: Challenges of the eighties. In R.S. Michalski, J.G. Carbonell, and T.M. Mitchell, (Eds), Machine Learning: An Al Approach, volume 2, pages 27-42. Morgan Kaufmann Publishers, 1986.

[Mitchell 1997] Thomas Mitchell, Machine Learning, McGraw Hill; 1997.

[Quinlan 1993] Quinlan, J. R., C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, 1993.

[Quinlan 1996] Quinlan J. R., Improved use of continuous attributes in c4.5, Journal of Artificial Intelligence Research, 4:77-90, 1996.

[Vlahavas et al. 2006] Vlahavas I., Kefalas P., Bassiliades N., Kokkoras F., Sakellariou I., "Artificial Intelligence"- 3rd Edition, (in Greek), Giourdas Publications, ISBN 960-387-432-0, 2006.

[Wang et al, 2004] Jason T. L. Wang, Mohammed J. Zaki, Hannu T.T. Toivonen, Dennis E. Shasha (Eds), Data Mining in Bioinformatics, Springer; 2004.

[Witten and Frank, 2005] Ian H. Witten and Eibe Frank, Data Mining: Practical Machine Learning Tools and Techniques, 2nd Edition, Morgan Kaufmann, 2005.

[Wu and Kumar, 2009] Xindong Wu and Vipin Kumar (Eds), The Top Ten Algorithms in Data Mining, Chapman & Hall/CRC; 2009.

[Xie et.al 2009] Tao Xie, Suresh Thummalapenta, David lo, Chao Liu, Data Mining for software engineering, Computer, Vol.42, Num8, pp.55-62, 2009.

[Yi Cao, 2008] Efficient K-Nearest Neighbor, Matlab central, http://www.mathworks.com/matlabcentral/fileexchange/19345, 2008.

[Zadeh, L. A. et al. 1996] Zadeh, L. A. et al., Fuzzy Sets, Fuzzy Logic, Fuzzy Systems, World Scientific Press, 1996.

[Zhang and Berardi, 1998] G. Zhang and L.V. Berardi, An investigation of neural networks in thyroid function diagnosis, Health Care Management Science (1998), pp. 29–37, 1998.