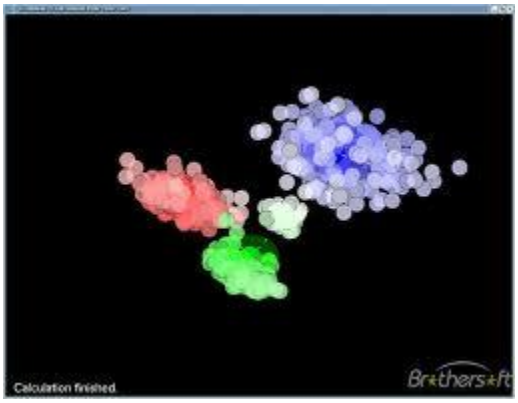
ΑΛΕΞΑΝΔΡΕΙΟ Τ.Ε.Ι. ΘΕΣΣΑΛΟΝΙΚΗΣ
ΣΧΟΛΗ ΤΕΧΝΟΛΟΓΙΚΩΝ ΕΦΑΡΜΟΓΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Πτυχιακή εργασία

«"ΕΦΑΡΜΟΓΗ ΤΗΣ ΤΕΧΝΙΚΗΣ ΣΥΣΤΑΔΟΠΟΙΗΣΗΣ ΔΕΔΟΜΕΝΩΝ ΓΙΑ ΤΗΝ ΕΠΙΤΑΧΥΝΣΗ ΤΗΣ ΑΝΑΖΗΤΗΣΗΣ ΟΜΟΙΟΤΗΤΑΣ ΑΠΟ ΜΕΓΑΛΕΣ ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ "»



Του φοιτητή

ΜΑΥΡΕΑ ΓΕΩΡΓΙΟΥ

Αρ. Μητρώου: 052856

Επιβλέπων καθηγητής

ΚΑΡΑΜΗΤΟΠΟΥΛΟΣ ΛΕΩΝΙΔΑΣ

Θεσσαλονίκη 2011

«Πτυχιακή εργασία του φοιτητή Μαυρέα Γεώργιου»

«Πτυχιακή εργασία του φοιτητή Μαυρέα Γεώργιου»

Μαυρέας Γεώργιος

Πτυχιακή Εργασία

που παρουσιάστηκε στο

Τμήμα Πληροφορικής

Νοέμβριος 2011

«Πτυχιακή εργασία του φοιτητή Μαυρέα Γεώργιου»

Περίληψη

Κύριος στόχος της εργασίας αυτής είναι η αναφορά βασικών μεθόδων επεξεργασίας δεδομένων, αναγνωρίσεις προτύπων, καθώς και αλγορίθμων συσταδοποίησης. Θα δοθεί ιδιαίτερη έμφαση σε αλγορίθμους κατάλληλους για κατηγοριοποίηση δεδομένων.

Abstract

The main objective of this paper is to report basic data-processing, pattern recognition and clustering algorithms. Give increased emphasis on algorithms suitable for data classification.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον καθηγητή κ. Λεωνίδα Καραμητόπουλο για την επίβλεψή του στην ολοκλήρωση αυτής της πτυχιακής εργασίας καθώς και για την ευκαιρία που μου δόθηκε να ασχοληθώ με ένα πολύ ενδιαφέρον αντικείμενο, από το οποίο αποκόμισα πολύτιμες εμπειρίες. Θα ήθελα να ευχαριστήσω προκαταβολικά τους καθηγητές κα. Χατζάρα και κος. Σαρηγιαννίδη για το χρόνο που διέθεσαν για την ανάγνωση της παρούσας Πτυχιακής εργασίας και για τις τυχόν παρατηρήσεις τους. Τέλος, θα ήθελα να ευχαριστήσω ιδιαίτερα την οικογένεια μου για την βοήθειά της και για την αμέριστη συμπαράστασή που μου έδειξαν καθ’ όλη την διάρκεια των σπουδών μου.

Μαυρέας Γιώργος

Θεσσαλονίκη, Νοέμβριος 2011

«Πτυχιακή εργασία του φοιτητή Μαυρέα Γεώργιου»

Περιεχόμενα

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

Πρόλογος

Στις μέρες μας είναι πλέον κοινά αποδεκτό ότι ζούμε σε μια κοινωνία της πληροφορίας. Είναι γεγονός ότι στη σημερινή εποχή παράγεται περισσότερη πληροφορία από πότε άλλοτε. Η ανάπτυξη της τεχνολογίας μας έδωσε την δυνατότητα να ξεπεράσουμε τα παλιά κλασσικά μέσα αποθήκευσης, όπως το χαρτί για παράδειγμα και να περάσουμε σε πιο εξελιγμένα ψηφιακά μέσα αποθήκευσης με πολλαπλάσια χωρητικότητα. Αυτό είχε ως αποτέλεσμα την δημιουργία τεράστιων πηγών από πληροφορίες. Σ' αυτή τη φάση, όμως, με την επίλυση του προβλήματος της αποθήκευσης των μέσων σε ψηφιακή μορφή, παράλληλα δημιουργήθηκε και ένα καινούριο πρόβλημα. Το πρόβλημα αυτό δεν ήταν άλλο από το πως θα μπορεί ο καθένας να άνακτα την επιθυμητή πληροφορία. Η ανάκτηση της επιθυμητής πληροφορίας, σε σύντομο χρονικό διάστημα αν όχι σε πραγματικό χρόνο, μέσα από το σύνολο της αποθηκευμένης πληροφορίας ήταν εξαιρετικά δύσκολη αν όχι αδύνατη διαδικασία. Αυτό είχε ως αποτέλεσμα την επινόηση μηχανισμών για την αυτόματη εύρεση της επιθυμητής πληροφορίας. Οι μηχανισμοί αυτοί στηρίζονται στο γεγονός ότι οι δυνατότητες για γρήγορη ανάκτηση της πληροφορίας εξαρτώνται σε πολύ μεγάλο βαθμό από τις μεθόδους που χρησιμοποιήθηκαν για την αποθήκευσή της. Η χρήση αποδοτικών μεθόδων αποθήκευσης της πληροφορίας επιταχύνει την ανάκτησή της, δεν λύνει όμως το πρόβλημα της εύρεσής της. Υπάρχουν δύο τομείς που ασχολούνται με την αποθήκευση και εύρεση πληροφορίας, ο τομέας των Βάσεων Δεδομένων και ο τομέας των Συστημάτων Ανάκτησης Εξόρυξης Πληροφορίας. Η Εξόρυξη πληροφοριών είναι ένα κομμάτι της διαδικασίας της «Ανακάλυψης Γνώσης από Βάσεις Δεδομένων».

Τα συστήματα διαχείρισης σχεσιακών βάσεων δεδομένων ακολουθούν ένα δομημένο τρόπο αποθήκευσης της πληροφορίας. Προσφέρουν ένα απλό μοντέλο αναπαράστασης των δεδομένων, κατανοητό στους χρήστες. Παρέχουν επίσης έναν απλό και αποδοτικό τρόπο έκφρασης σύνθετων και πολύπλοκων ερωτήσεων, ο οποίος είναι πολύ κοντά στον τρόπο αναπαράστασης των δεδομένων. Όλα αυτά έχουν συμβάλλει στην ευρεία αποδοχή και στη μεγάλη εμπορική επιτυχία αυτών των συστημάτων. Στα συστήματα ανάκτησης πληροφορίας και συγκεκριμένα στα συστήματα ανάκτησης κειμένου βασική οντότητα είναι το έγγραφο (document) το οποίο περιέχει μη δομημένη πληροφορία. Ένα τέτοιο σύστημα ανάκτησης πληροφορίας επεξεργάζεται, και γενικότερα εκτελεί το σύνολο των λειτουργιών του σε μια συλλογή (collection) από έγγραφα. Η απόκριση, στις αιτήσεις για ανάκτηση πληροφορίας που δέχεται, είναι ένα υποσύνολο του συνόλου των εγγράφων της συλλογής. Η απόφαση για την ανάκτηση ενός υποσυνόλου εγγράφων σε σχέση με μια ερώτηση βασίζεται στην “ομοιότητα” που υπάρχει μεταξύ των εγγράφων και της ερώτησης. Η ομοιότητα αυτή, υπολογίζεται συγκρίνοντας τις τιμές συγκεκριμένων χαρακτηριστικών, που αντιστοιχούν σε κάθε έγγραφο, με τις τιμές των χαρακτηριστικών που σχηματίζουν την ερώτηση. Σε κάθε έγγραφο που συμμετέχει στην απάντηση αντιστοιχεί διαφορετικός βαθμός ομοιότητας με την ερώτηση διότι τα έγγραφα της συλλογής δεν είναι πανομοιότυπα. Οι ερωτήσεις που εκτελούνται κατ’ αυτό τον τρόπο καλούνται ερωτήσεις ομοιότητας (similarity queries). *Αν και αυτοί οι δύο τομείς αναπτύσσονται ανεξάρτητα ο ένας από τον άλλο είναι φανερό ότι με τον κατάλληλο συνδυασμό τους θα μπορούσαν να προκύψουν συστήματα με πολύ ισχυρές δυνατότητες*.

Ένας τέτοιος συνδυασμός θα εκμεταλλεύεται το απλό μοντέλο αναπαράστασης των δεδομένων και την πρότυπη γλώσσα ερωτήσεων των συστημάτων διαχείρισης σχεσιακών βάσεων δεδομένων από τη μια και από την άλλη, τη δυνατότητα των συστημάτων ανάκτησης πληροφορίας για εκτέλεση πολύπλοκων ερωτήσεων που είναι πιο κοντά στη σημασιολογία που αποδίδει ο χρήστης στα δεδομένα. Μια από τις πιο σημαντικές διεργασίες στη διαδικασία εξόρυξης γνώσης είναι η συσταδοποίηση. Πρόκειται για μια μεθοδολογία ανακάλυψης συστάδων και κατανομών ή προτύπων που παρουσιάζουν ενδιαφέρον στα υπό μελέτη δεδομένα. Ως συστάδα ορίζεται μια συλλογή αντικειμένων από τα δεδομένα, με βάση τη μεταξύ τους ομοιότητα. Οι απαρχές της ανάλυσης κατά συστάδες εντοπίζονται από τη δεκαετία του 1960. Σήμερα, οι διαδικασίες της συσταδοποίησης θεωρούνται απαραίτητο συστατικό για πολλές εφαρμογές ετερόκλητων κλάδων, όπως η αναγνώριση προτύπων, η ανάλυση δεδομένων, η μεταποίηση εικόνας (*image processing*), η έρευνα αγοράς και άλλες. Η διαδικασία της συσταδοποίησης μπορεί να οδηγήσει σε διαφορετικές τμηματοποιήσεις ενός συνόλου δεδομένων, ανάλογα με το κριτήριο συσταδοποίησης που χρησιμοποιείται. Τα βασικά βήματα της διαδικασίας είναι η επιλογή των κατάλληλων γνωρισμάτων (*attributes*) στα οποία πρόκειται να εφαρμοστεί η συσταδοποίηση, η επιλογή ενός αλγορίθμου που οδηγεί στον καθορισμό ενός καλού σχήματος συσταδοποίησης, η επικύρωση των αποτελεσμάτων και εν τέλει η ερμηνεία τους. Ο αλγόριθμος που επιλέγεται καθορίζεται από το μέτρο εγγύτητας που προσδιορίζει πόσο «όμοια» είναι δύο αντικείμενα (τα χαρακτηριστικότερα μέτρα απόστασης / ομοιότητας δίνονται από τους Johnson and Wichern, 1998) και το κριτήριο συσταδοποίησης, το οποίο μπορεί να εκφραστεί μέσω μιας συνάρτησης ή κάποιου τύπου κανόνων, ή να ληφθεί υπόψη ο τύπος συστάδων που αναμένεται να εμφανιστούν στο σύνολο δεδομένων.

Στη μηχανική εκμάθηση και την αναγνώριση προτύπων, η ανάλυση συστάδων αναφέρεται συχνά ως μη εποπτευόμενη εκμάθηση. Εξάλλου, σύμφωνα με τον Marques de Sá (2001), η συσταδοποίηση κατατάσσεται στη μη επιβλεπόμενη ταξινόμηση όσον αφορά τις προσεγγίσεις της αναγνώρισης προτύπων. Αυτή είναι και η συνεισφορά της συσταδοποίησης στην αναγνώριση προτύπων. Χρησιμοποιώντας ένα μέτρο ομοιότητας, η συσταδοποίηση είναι σε θέση να οργανώσει τα δεδομένα -πρότυπα σε ενδιαφέρουσες ομάδες, χωρίς να έχει σχετική εκ των προτέρων πληροφορία, πράγμα το οποίο μπορεί να βοηθήσει μελλοντικά στην εφαρμογή ταξινόμησης σε νέα εισερχόμενα δεδομένα. Οι απαιτήσεις που έχουμε από έναν αλγόριθμο συσταδοποίησης είναι να έχει δυνατότητες κλιμάκωσης και να μπορεί να χειριστεί διαφορετικού τύπου μεταβλητές ή δεδομένα με θόρυβο ή μεγάλο πλήθος διαστάσεων ή μεταβλητών. Επίσης, ένας ιδανικός αλγόριθμος χρειάζεται να οδηγεί σε συστάδες αυθαίρετου σχήματος, να είναι κατανοητός και χρήσιμος, καθώς και να καλύπτει τους περιορισμούς που μπορεί να τίθενται στα πλαίσια εφαρμογής πραγματικών δεδομένων. Οι βασικότερες ομάδες στις οποίες διαχωρίζονται οι αλγόριθμοι με βάση τη μέθοδο συσταδοποίησης είναι η διαιρετική και η ιεραρχική συσταδοποίηση .

ΚΕΦΑΛΑΙΟ 1ο

<< Εισαγωγή στην αναζήτηση ομοιότητας μέσα από βάσεις δεδομένων >>

1. Εισαγωγή

Σ’ αυτήν την ενότητα θα αναφερθούμε σε κάποιες γενικές έννοιες καθώς και κάποιους ορισμούς όπως για παράδειγμα το data mining ή εξόρυξη δεδομένων όπως είναι η ελληνική απόδοση του ορισμού. Ακόμη γίνεται ένας διαχωρισμός μεταξύ του τι είναι η κατηγοριοποίηση (classification) και τι είναι η Συσταδοποίηση (clustering). Στη συνέχεια θα αναφερθούμε σε αλγόριθμους κατηγοριοποίησης και σε αλγόριθμους που είναι βασισμένοι στην απόσταση.

1.1 Τι είναι το data mining ;

Το data mining ή αλλιώς, εξόρυξη δεδομένων, είναι ένας όρος που ακούμε πολύ συχνά τον τελευταίο καιρό χωρίς όμως να γνωρίζουμε επακριβώς τι σημαίνει και που χρησιμοποιείται. Τι είναι λοιπόν το data mining;

Ο ορισμός που χρησιμοποιείται στην βιβλιογραφία για να αποδώσει το νόημα του data mining είναι ο ακόλουθος: "Η σύνθετη διαδικασία εξαγωγής συγκεκριμένης, προηγουμένως άγνωστης και δυνητικά ωφέλιμης, γνώσης από δεδομένα".

Πρακτικά μπορούμε να καταλάβουμε τι είναι το data mining μέσω ενός παραδείγματος. Ας υποθέσουμε λοιπόν πως έχουμε αγοράσει μια τεράστια έκταση χωραφιών για την οποία πληρώσαμε όλη μας την περιουσία. Ωστόσο μια μέρα έρχεται κάποιος και μας λέει πως υπάρχει κρυμμένος, κάπου στην έκταση που αγοράσαμε, ένας θησαυρός. Βέβαια εμείς δεν γνωρίζουμε ούτε που ακριβώς βρίσκεται ο θησαυρός, ούτε και πως θα τον εξάγουμε από το έδαφος. Κάπως έτσι νιώθουν και οι επιχειρηματίες όταν επενδύουν πολλά χρήματα σε μια βάση δεδομένων στην οποία αποθηκεύονται χρήσιμες πληροφορίες για την επιχείρηση τους, όταν προσπαθούν να εξάγουν χρήσιμα συμπεράσματα από έναν τεράστιο όγκο δεδομένων. Εδώ λοιπόν έρχεται το data mining για να βοηθήσει. Το data mining προσφέρει τα κατάλληλα εργαλεία έτσι ώστε να εξάγει όλη την χρήσιμη πληροφορία από τεράστιους όγκους δεδομένων. Πληροφορία, την οποία αν την εκμεταλλευτεί κάποιος έξυπνα και σωστά, μπορεί να αποκομίσει μεγάλο κέρδος.

1.2 Τεχνικές Data Mining

Το data mining συνήθως περιλαμβάνει 4 βασικές κατηγορίες από διαδικασίες:

•Clustering: είναι η διαδικασία του να ανακαλύπτουμε ομάδες και δομές από δεδομένα που είναι κατά κάποιο τρόπο ‘’όμοια’’ μεταξύ τους, χωρίς βέβαια να υπάρχουν από πριν γνωστές δομές στα δεδομένα.Το clustering θα αναλυθεί περαιτέρω παρακάτω.

•Classification: είναι η διαδικασία του να γενικεύουμε γνωστές δομές για να εφαρμόσουμε σε νέα δεδομένα. Με άλλα λόγια είναι η εκμάθηση μιας τεχνικής να προβλέπει την κλάση ενός στοιχείου επιλέγοντας από προκαθορισμένες τιμές.

Για παράδειγμα ένα πρόγραμμα ηλεκτρονικού ταχυδρομείου προσπαθεί να κατηγοριοποιήσει τα εισερχόμενα μηνύματα σε κανονικά ή spam. Οι πιο γνωστές τεχνικές κατηγοριοποίησης περιλαμβάνουν τα decision trees, nearest neighbor, naive Bayesian classification και τα νευρωνικά δίκτυα (neural networks).

•Regression Analysis: περιλαμβάνει τεχνικές για μοντελοποίηση και ανάλυση μεταβλητών. Με άλλα λόγια regression είναι η διαδικασία του να βρεις μια συνάρτηση η οποία να μοντελοποιεί τα δεδομένα με το μικρότερο δυνατό σφάλμα. Χρησιμοποιείται συνήθως για στατιστικούς σκοπούς.

•Association Rule Learning: με αυτήν την διαδικασία ψάχνουμε για σχέσεις μεταξύ των μεταβλητών. Για παράδειγμα ένα σούπερ μάρκετ μπορεί να συγκεντρώσει δεδομένα για τις αγοραστικές συνήθειες των καταναλωτών. Χρησιμοποιώντας τεχνικές association rule learning, το σούπερ μάρκετ μπορεί να καθορίσει ποια προϊόντα αγοράζονται συχνότερα μαζί και να χρησιμοποιήσει αυτές τις πληροφορίες προς όφελος του. Το association rule learning αναφέρεται μερικές φορές και ως market basket analysis.

1.3. Εφαρμογές του Data Mining

Που ακριβώς όμως χρησιμοποιείται το διάσημο data mining; Η χρήση του data mining επεκτείνεται συνεχώς όπου υπάρχει η ανάγκη εξόρυξης χρήσιμης γνώσης από τεράστια σύνολα δεδομένων. Εφαρμογές της τεχνολογίας αυτής μπορούμε να εντοπίσουμε σε εμπορικούς τομείς όπως για παράδειγμα σε αλυσίδες σουπερ-μάρκετ, τράπεζες, στα μέσα μαζικής ενημέρωσης και στην διαφήμιση. Γενικότερα με το data mining μπορούμε να προβλέψουμε συμπεριφορές και να εντοπίσουμε τάσεις και μοτίβα, οπότε όπως φαίνετε είναι μια πολύ χρήσιμη τεχνική για όλους τους παραπάνω τομείς.

Ακόμη και στον επιστημονικό τομέα το data mining χρησιμοποιείται ευρέως. Έτσι μπορούμε να δούμε την χρήση του, εκτός φυσικά από τον τομέα της πληροφορικής, στην ιατρική, στην βιολογία αλλά και στις τηλεπικοινωνίες.

1.4 Κατηγοριοποίηση (classification)

Η ***κατηγοριοποίηση (classification)*** είναι η πιο γνωστή και πιο δημοφιλής τεχνική ***εξόρυξης γνώσης (data mining)***. Πολλές εταιρίες του ιδιωτικού και του δημόσιου τομέα χρησιμοποιούν σε καθημερινή βάση συστήματα κατηγοριοποίησης. Παραδείγματα τέτοιου είδους συστημάτων είναι τα συστήματα αναγνώρισης προτύπων, συστήματα ιατρικών διαγνώσεων, συστήματα έγκρισης δανείων και πιστωτικών καρτών, συστήματα ανίχνευσης λαθών σε βιομηχανικές εφαρμογές, συστήματα κατηγοριοποίησης των τάσεων στην οικονομία κ.α. Για παράδειγμα όταν κάποιος προβλέπει μια ηλικία, στην ουσία επιλύει ένα πρόβλημα κατηγοριοποίησης. Ένα άλλο, πιο καλά ορισμένο, παράδειγμα παρουσιάζεται παρακάτω:

Παράδειγμα 1.1: Οι δάσκαλοι κατηγοριοποιούν τους μαθητές ως A,B,C,D ή F με βάση τους βαθμούς τους. Χρησιμοποιώντας απλά όρια (60, 70, 80, 90) μπορούμε να έχουμε τον παρακάτω διαχωρισμό των μαθητών σε κλάσεις:

90 ≤ βαθμός A,

80 ≤ βαθμός < 90 B,

70 ≤ βαθμός < 80 C,

60 ≤ βαθμός < 70 D,

Βαθμός < 60 F

Όλες οι προσεγγίσεις στην εκτέλεση της κατηγοριοποίησης προϋποθέτουν γνώση των δεδομένων. Συνήθως χρησιμοποιούμε ένα σύνολο εκπαίδευσης για να καθορίσει τις συγκεκριμένες παραμέτρους που απαιτούνται από την τεχνική. Τα δεδομένα εκπαίδευσης (training data) αποτελούνται από ένα δείγμα δεδομένων εισόδου καθώς επίσης και από την κατηγοριοποίηση που έχει δοθεί σε αυτά τα δεδομένα. Το πρόβλημα της κατηγοριοποίησης παρουσιάζεται από τον ορισμό 1.1 και από τον ορισμό 1.2. Ο δεύτερος ορισμός περιγράφει με μαθηματικό τρόπο το πρόβλημα.

Ορισμός 1.1: Η κατηγοριοποίηση (classification) είναι η διαδικασία η οποία απεικονίζει ένα σύνολο δεδομένων σε προκαθορισμένες ομάδες. Τις ομάδες αυτές συχνά τις καλούμε ***κατηγορίες ή κλάσεις***.

Ορισμός 1.2: Έστω μια Βάση Δεδομένων DB = {*t1*,*t2*,…,*tn*} πλειάδων (στοιχείων, εγγραφών) και ένα σύνολο από κατηγορίες C = {*C1*,*C2*,…,*Cm*}. Το πρόβλημα της κατηγοριοποίησης είναι ο ορισμός μιας απεικόνισης f: DB , C όπου κάθε *ti* τοποθετείται σε μια κατηγορία. Μια κατηγορία ή κλάση *Cj,* περιέχει ακριβώς αυτές τις πλειάδες όπου έχουν απεικονιστεί σε αυτή, δηλαδή $Cj = \{ti \mid f(ti) = Cj, 1 \le i \le n, \text{ και } ti \in DB\}$.

Οι ορισμοί 1.1 και 1.2 θεωρούν την κατηγοριοποίηση σαν μια απεικόνιση από τη Βάση Δεδομένων στο σύνολο των κατηγοριών. Πρέπει να υπογραμμιστεί ότι οι κατηγορίες είναι προκαθορισμένες, δεν επικαλύπτονται και διαμερίζουν ολόκληρη την Βάση Δεδομένων. Κάθε στοιχείο της Βάσης Δεδομένων τοποθετείται σε ακριβώς μια κατηγορία.

Οι κατηγορίες που υπάρχουν σε ένα πρόβλημα κατηγοριοποίησης είναι στην πραγματικότητα ***κλάσεις ισοδυναμίας (equivalence classes).***

Η επίλυση των προβλημάτων κατηγοριοποίησης περιλαμβάνει δύο βασικά στάδια

- Δημιουργούμε ένα μοντέλο από την αξιολόγηση και την ανάλυση των δεδομένων εκπαίδευσης. Αυτό το βήμα έχει σαν είσοδο τα δεδομένα εκπαίδευσης και σαν έξοδο ένα ορισμό του μοντέλου που αναπτύχθηκε. Το μοντέλο που δημιουργείται από αυτό το στάδιο είναι σε θέση να κατηγοριοποιεί τα δεδομένα εκπαίδευσης με όσο το δυνατό μεγαλύτερη ακρίβεια. Όταν είναι ήδη γνωστές οι κατηγορίες του συνόλου των δεδομένων εκπαίδευσης, δηλαδή το σύνολο των δεδομένων εκπαίδευσης περιλαμβάνει ένα χαρακτηριστικό το οποίο δείχνει την κλάση στην οποία κατηγοριοποιείται η κάθε πλειάδα, τότε το βήμα αυτό καλείται ***εποπτευμένη μάθηση (supervised learning)***, σε αντίθετη περίπτωση, δηλαδή όταν δεν είναι γνωστές οι κατηγορίες του συνόλου των δεδομένων εκπαίδευσης, τότε το βήμα αυτό καλείται ***μη εποπτευμένη μάθηση (unsupervised learning - clustering)***.
- Εφαρμόζουμε το μοντέλο που αναπτύχθηκε στο προηγούμενο βήμα κατηγοριοποιώντας τις πλειάδες της υπό εξέταση Βάσης Δεδομένων (μελλοντικές περιπτώσεις). Εάν και το δεύτερο βήμα στην πραγματικότητα εκτελεί την κατηγοριοποίηση, η περισσότερη έρευνα έχει γίνει για το πρώτο βήμα. Το δεύτερο βήμα είναι συνήθως εύκολο στην υλοποίηση.

Υπάρχουν τρεις βασικές μέθοδοι που χρησιμοποιούνται για να λύσουν το πρόβλημα της κατηγοριοποίησης:

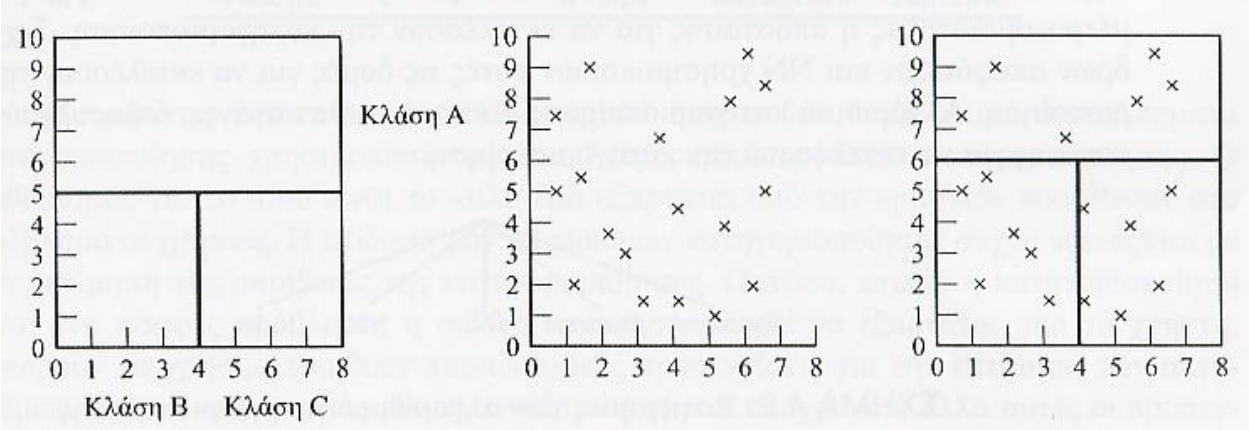
- **Καθορισμός των ορίων:** Η κατηγοριοποίηση εκτελείται με διαίρεση του χώρου της εισόδου των εν δυνάμει πλειάδων της Βάσης Δεδομένων σε περιοχές όπου κάθε περιοχή συνδέεται με μια κατηγορία

- **Χρήση κατανομών πιθανότητας:** Για κάθε κατηγορία που δίνεται *Cj* P(*ti* | *Cj*) είναι η συνάρτηση κατανομής πιθανότητας (probability distribution function) για την κατηγορία υπολογισμένη σε ένα σημείο, *ti*. Αν η πιθανότητα εμφάνισης κάθε κατηγορίας P(*Cj*), είναι γνωστή (ίσως να έχει οριστεί από κάποιον ειδικό του πεδίου εφαρμογής – domain expert), τότε P(*Cj*) P(*ti* | *Cj*) είναι η εκτίμηση της πιθανότητας ότι η *ti* ανήκει στην κατηγορία *Cj*

- **Χρήση εκ των υστέρων πιθανοτήτων:** Με δεδομένη μια τιμή δεδομένων *ti*, θέλουμε να καθορίσουμε την πιθανότητα για την οποία η *ti* ανήκει στην κατηγορία *Cj.* Αυτό υποδηλώνεται με το P(*Cj* | *ti*) που ονομάζεται εκ των υστέρων πιθανότητα (posterior probability). Μια προσέγγιση κατηγοριοποίησης είναι ο καθορισμός της εκ των υστέρων πιθανότητας για κάθε κατηγορία και στη συνέχεια η τοποθέτηση των πλειάδων στην κατηγορία με τη μεγαλύτερη πιθανότητα.

Το παράδειγμα 1.1 ανήκει στην πρώτη κατηγορία όπως επίσης και όλες οι τεχνικές δένδρων απόφασης, ενώ οι προσεγγίσεις των νευρωνικών δικτύων ανήκουν στην Τρίτη κατηγορία.

Ας υποθέσουμε ότι μας δίνεται μια Βάση Δεδομένων που αποτελείται από πλειάδες της μορφής $t = <x,y>$ όπου $0 \leq x \leq 8$ και $0 \leq y \leq 10$. Το σχήμα 1 παρουσιάζει το πρόβλημα της κατηγοριοποίησης. Το σχήμα 1 (α) παρουσιάζει τις προκαθορισμένες κατηγορίες – κλάσεις, το σχήμα 1 (β) παρέχει δείγματα δεδομένων εισόδου και το σχήμα 1 (γ) παρουσιάζει την κατηγοριοποίηση των δεδομένων με βάση τις ορισμένες κατηγορίες.



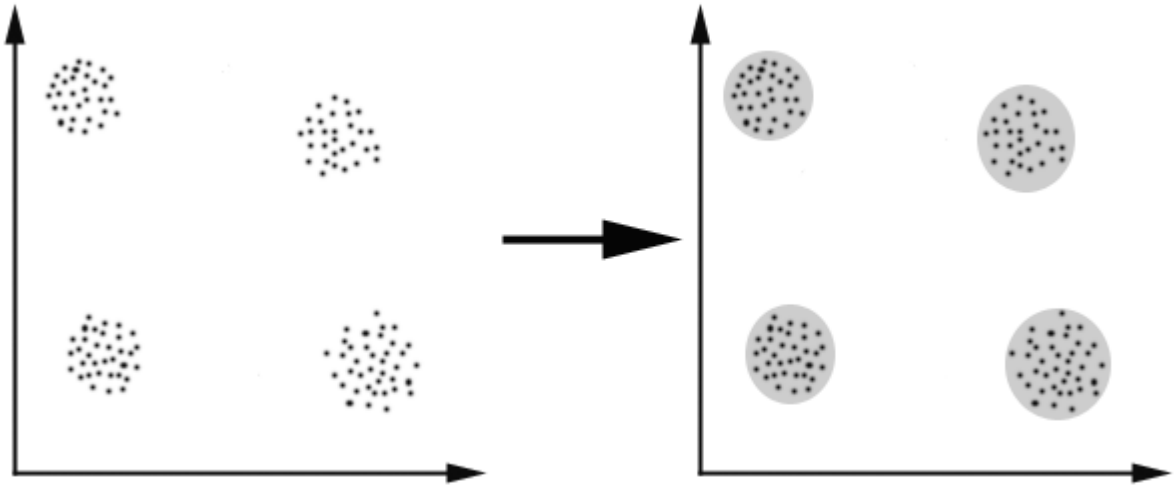
Το πρόβλημα της κατηγοριοποίησης

Εικόνα 1

Ένα πολύ σημαντικό ζήτημα σχετικό με την κατηγοριοποίηση είναι η **υπερπροσαρμογή**. Συγκεκριμένα, λέγοντας υπερπροσαρμογή εννοούμε το φαινόμενο κατά το οποίο η τεχνική κατηγοριοποίηση ταιριάζει ακριβώς στα δεδομένα εκπαίδευσης και ίσως να μη μπορεί να εφαρμοστεί σε πιο ευρύ πληθυσμό δεδομένων. Για παράδειγμα, ας υποθέσουμε ότι τα δεδομένα εκπαίδευσης περιέχουν λανθασμένα δεδομένα ή δεδομένα με θόρυβο. Σε αυτή την περίπτωση, το ακριβές ταίριασμα των δεδομένων δεν είναι επιθυμητό.

1.5 Συσταδοποίηση (Clustering)

Η συσταδοποίηση είναι μια μέθοδος ανάθεσης των στοιχείων ενός συνόλου σε υποσύνολα (συστάδες) έτσι ώστε οι συστάδες που θα δημιουργηθούν να είναι παρόμοιες ως προς κάποιο κριτήριο. Συνήθως δεν υπάρχει καμιά πρότερη γνώση σχετικά με το πόσες ομάδες (συστάδες – clusters) θα δημιουργηθούν ή ποια θα είναι η δομή των συστάδων αλλά όλα αποφασίζονται στην πορεία από αποφάσεις που παίρνονται κατά την εκτέλεση του αλγορίθμου βάσει κάποιων παραμέτρων. Η Συσταδοποίηση (clustering) , είναι η διαδικασία συσταδοποίησης αντικειμένων με όμοια χαρακτηριστικά και η κατάταξη σε κλάσεις ή συστάδες ή συμπλέγματα. Στην Συσταδοποίηση οι συστάδες δεν είναι προκαθορισμένες αλλά προσδιορίζονται από τα δεδομένα. Η συσταδοποίηση αναφέρεται εναλλακτικά και σαν μη εποπτευόμενη μάθηση. Μπορεί να θεωρηθεί σαν μια διαμέριση ή τμηματοποίηση των δεδομένων σε ομάδες που μπορεί να είναι ή να μην είναι διακριτές μεταξύ τους. Η συσταδοποίηση συνήθως επιτυγχάνεται με τον καθορισμό της ομοιότητας, ως προς προκαθορισμένα γνωρίσματα, ανάμεσα στα δεδομένα. Τα πιο σχετικά δεδομένα συσταδοποιούνται σε ίδιες συστάδες.



Clustering

Εικόνα 2

Το clustering μπορεί να θεωρηθεί ως το σημαντικότερο πρόβλημα μη εποπτευόμενης μάθησης (unsupervised learning). Έτσι όπως και κάθε άλλο πρόβλημα του είδους του έχει να κάνει με το να βρούμε μια δομή σε μια συλλογή από δεδομένα τα οποία δεν γνωρίζουμε δηλαδή δεν έχουν κάποιο label (ετικέτα) – για αυτό ονομάζεται και unsupervised learning. Ένας ορισμός του clustering θα μπορούσε να είναι ο εξής: ‘’η διαδικασία του να οργανώσουμε τα δεδομένα μας σε ομάδες, όπου τα μέλη κάθε συστάδας είναι όμοια κατά κάποιο τρόπο’’. Έτσι ονομάζουμε cluster μια ομάδα ή αλλιώς μια συστάδα από δεδομένα που έχουν ομαδοποιηθεί σε αυτό το cluster σύμφωνα με κάποιο κριτήριο ομοιότητας. Συσταδοποίηση, πολλές φόρες μπορεί να θεωρηθεί η μάθηση χωρίς επίβλεψη. Μερικές φορές , διάφορες ομάδες ανθρώπων της έχουν δώσει διάφορα χαρακτηριστικά γνωρίσματα. Για παράδειγμα οι στατιστικολόγοι την συνδέουν με την ταξινόμηση , οι ψυχολόγοι με την διαλογή και οι άνθρωποι του μάρκετινγκ με τον κατακερματισμό της αγοράς .

Πρωταρχικός στόχος της συσταδοποίησης είναι ο διαχωρισμός και η οργάνωση των δεδομένων σε κλάσεις, τέτοιες ώστε μέσα σ' αυτές να υπάρχει υψηλός βαθμός ομοιότητας ή χαμηλός βαθμός ομοιότητας μεταξύ αυτών των στοιχείων της κάθε κατηγορίας. Σε αντίθεση με την ταξινόμηση, στην συσταδοποίηση βρίσκουμε απ' ευθείας από τα δεδομένα τόσο τον αριθμό των κατηγοριών όσο και την ετικέτα των τάξεων στην κάθε κατηγορία. Πιο ανεπίσημα τέλος, ονομάζουμε την εξεύρεση φυσικών ομαδοποιήσεων μεταξύ των αντικειμένων. Υπάρχουν δύο διαφορετικοί τύποι συσταδοποίησης. Ο πρώτος τρόπος έχει να κάνει με τον αλγόριθμο του διαχωρισμού. Με βάση αυτόν τον τύπο αλγορίθμου , κατασκευάζουμε διάφορα τμήματα και στην συνέχεια τα αξιολογούμε με βάση κάποιο κριτήριο ενώ ο δεύτερος τρόπος έχει να κάνει με τον ιεραρχικό αλγόριθμο. Σύμφωνα με αυτόν τον τύπο αλγορίθμου, δημιουργούμε μια ιεραρχική αποσύνθεση του συνόλου των αντικειμένων , χρησιμοποιώντας κάποιο κριτήριο .

Οι επιθυμητές ιδιότητες ενός αλγορίθμου συσταδοποίησης είναι:

- η Επεκτασιμότητα (από άποψη χρόνου και χώρου)
- η Ικανότητα να ασχοληθεί με διαφορετικούς τύπους δεδομένων
- οι Ελάχιστες απαιτήσεις για τις γνώσεις του domain έτσι ώστε να καθορίσει τις παραμέτρους εισόδου
- να είναι σε θέση να ασχοληθεί με το θόρυβο και τις ακραίες τιμές
- να έχει ιδιαίτερη ευαισθησία στην σειρά των αρχείων εισόδου
- να μπορεί ο χρήστης να ενσωματώνει και να καθορίζει τους περιορισμούς
- η χρηστικότητα και η ερμηνευτικότητα

Συνοψίζοντας για τις Μετρήσεις Ομοιότητας ένα χρήσιμο εργαλείο για να εκτιμήσουμε καλύτερα και να αξιολογήσουμε τα παραδείγματα είναι το δενδρόγραμμα. .Η ομοιότητα ανάμεσα σε δύο αντικείμενα σε ένα δενδρόγραμμα παρουσιάζεται ως το ύψος του χαμηλότερου εσωτερικού κόμβου που μοιράζονται .Υπάρχει μόνο ένα σύνολο δεδομένων που μπορεί να είναι απόλυτα συγκεντρωμένο χρησιμοποιώντας μια ιεραρχία. Η Ιεραρχική συσταδοποίηση μπορεί να δείξει μερικές φορές πρότυπα που είναι χωρίς νόημα ή ακόμη και πλαστά.

1.6 Αλγόριθμοι κατηγοριοποίησης

Στην παράγραφο αυτή παρουσιάζονται κάποιοι από τους γνωστούς αλγόριθμους κατηγοριοποίησης που έχουν προταθεί. Καταλαβαίνουμε εύκολα ότι όσο περισσότερα χαρακτηριστικά εμπλέκονται στην διαδικασία της κατηγοριοποίησης, τόσο πιο σύνθετο και πολύπλοκο γίνεται το μοντέλο κατηγοριοποίησης. Μια κατηγοριοποίηση που βασίζεται σε μια μόνο τιμή ενός χαρακτηριστικού δεν είναι αξιόλογη. Μπορούμε να διακρίνουμε πέντε είδη κατηγοριών αλγορίθμων κατηγοριοποίησης.

Συγκεκριμένα, υπάρχουν οι:

- Στατιστικοί αλγόριθμοι κατηγοριοποίησης
- Αλγόριθμοι κατηγοριοποίησης βασισμένοι στην απόσταση
- Αλγόριθμοι κατηγοριοποίησης βασισμένοι στα δένδρα απόφασης
- Αλγόριθμοι κατηγοριοποίησης βασισμένοι στα Νευρωνικά Δίκτυα
- Αλγόριθμοι κατηγοριοποίησης βασισμένη σε κανόνες

1.7 Αλγόριθμοι Βασισμένοι στην απόσταση

Η βασική ιδέα αυτών των αλγορίθμων είναι ότι κάθε στοιχείο του συνόλου δεδομένων που απεικονίζεται στην ίδια κατηγορία θεωρείται ότι είναι πιο κοντά σε στοιχεία της ίδιας κατηγορίας από όσο είναι σε στοιχεία τα οποία ανήκουν σε άλλες κατηγορίες. Έτσι, μπορούν να χρησιμοποιηθούν μέτρα ομοιότητας (ή απόστασης) ώστε να οριστεί η «ομοιότητα» των διαφορετικών στοιχείων της Βάσης Δεδομένων.

Ορισμός 1.3: Η ομοιότητα ανάμεσα σε δύο πλειάδες ti και tj, sim (ti, tj), σε μια Βάση Δεδομένων είναι μια απεικόνιση από το D x D στο διάστημα [0, 1]. Έτσι sim (ti, tj) ∈ [0, 1]. Ο Αντικειμενικός σκοπός είναι να οριστεί η απεικόνιση της ομοιότητας με τρόπο ώστε οι πλειάδες που μοιάζουν μεταξύ τους περισσότερο να έχουν μεγαλύτερη τιμή ομοιότητας.

Η δυσκολία στην εφαρμογή των μέτρων ομοιότητας είναι το πώς αυτά θα εφαρμοστούν στα στοιχεία της Βάσης Δεδομένων και αυτό γιατί τα περισσότερα μέτρα ομοιότητας υποθέτουν ότι οι τιμές είναι αριθμητικές (και συχνά διακριτές) και ίσως είναι δύσκολο να χρησιμοποιηθούν σε περισσότερα γενικά και αφηρημένα είδη δεδομένων. Θα πρέπει να αναφερθεί το ότι η χρήση ενός μέτρου ομοιότητας για μια κατηγοριοποίηση οπού οι κατηγορίες έχουν προκαθοριστεί (εποπτευμένη μάθηση), είναι κάπως απλούστερη από την χρήση ενός μέτρου ομοιότητας σε μια συσταδοποίηση (clustering – μη εποπτευμένη μάθηση), όπου οι κατηγορίες δεν είναι γνωστές εκ των προτέρων. Παρακάτω παρουσιάζονται 3 από τα πιο βασικά είδη αποστάσεων που χρησιμοποιούνται σαν μέτρα ομοιότητας ανάμεσα σε πλειάδες μέσα σε μία Βάσης Δεδομένων. Το κάθε πεδίο μιας εγγραφής θεωρείται σαν μια διαφορετική διάσταση. Έτσι μια πλειάδα θεωρείται ένα σημείο στο χώρο το ν διαστάσεων.

α. Ευκλείδεια απόσταση

Η ευκλείδεια απόσταση αποτελεί την πιο απλή και την πιο γνωστή περίπτωση ανάμεσα σε συνεχή δεδομένα. Μερικές χρήσιμες ιδιότητες είναι πως εξαρτάται από την κλίμακα μέτρησης κι επομένως αλλάζοντας την κλίμακα μπορούμε να πάρουμε ολότελα διαφορετικές αποστάσεις. Επίσης μεταβλητές με μεγάλες απόλυτες τιμές έχουν πολύ μεγαλύτερο βάρος και σχεδόν καθορίζουν την απόσταση ανάμεσα σε παρατηρήσεις. Η ερμηνεία της απόστασης είναι πολύ εύκολο να αποδοθεί γεωμετρικά. Στην πραγματικότητα η απόσταση αγνοεί τις στατιστικές ιδιότητες των παρατηρήσεων όπως για παράδειγμα τη μεταβλητότητα κάθε μεταβλητής. Δεδομένου ότι παίρνουμε τετραγωνικές αποκλίσεις outliers έχουν μεγάλη επίδραση στον υπολογισμό της απόστασης.

β. Απόσταση Manhattan

Η απόσταση Manhattan μοιάζει πολύ με την ευκλείδεια απόσταση με τη διαφορά ότι αντί για τετραγωνικές αποκλίσεις χρησιμοποιούμε απόλυτες αποκλίσεις. Συνήθως λόγω της ομοιότητας με την ευκλείδεια απόσταση δίνει περίπου ίδια αποτελέσματα εκτός από την περίπτωση που υπάρχουν *outliers* όπου επειδή τους δίνει μικρότερο βάρος (εξαιτίας απόλυτης τιμής) μπορεί να οδηγήσει σε πιο ανθεκτικά αποτελέσματα. Και αυτή η απόσταση αγνοεί τις στατιστικές ιδιότητες των δεδομένων.

γ. Chebychev distance

Η απόσταση Chebychev σε αντίθεση με τις υπόλοιπες αποστάσεις που είδαμε δεν χρησιμοποιεί όλες τις αποκλίσεις αλλά μόνο τη μεγαλύτερη εξ αυτών. Η απόσταση αυτή είναι χρήσιμη όταν κανείς θέλει να θεωρήσει δύο διαφορετικές παρατηρήσεις αν έχουν διαφορές τουλάχιστον σε μια μεταβλητή. Επειδή η απόσταση χρησιμοποιεί μόνο τη μεγαλύτερη απόκλιση εξαρτάται πολύ από τις διαφορές στην κλίμακα των μεταβλητών και επομένως αν οι κλίμακες είναι διαφορετικές ουσιαστικά θα αντικατοπτρίζει τη διαφορά στη μεταβλητή με την μεγαλύτερη κλίμακα.

Όλες οι παραπάνω αποστάσεις έχουν το μειονέκτημα ότι δεν λαμβάνουν υπόψη τους τις όποιες διαφορές στην κλίμακα των μεταβλητών όπως επίσης και τις διαφορές στις διακυμάνσεις τους. Επίσης τυχόν συσχετίσεις ανάμεσα στις μεταβλητές δεν λαμβάνονται υπόψη και έτσι κατά κάποιον τρόπο αν υπάρχουν συσχετισμένες μεταβλητές η απόσταση ανάμεσα σε δύο παρατηρήσεις μπορεί να είναι πλασματική.

ΚΕΦΑΛΑΙΟ 2ο

<< Συσταδοποίηση με έμφαση στον αλγόριθμο k-means >>

2.1 Εισαγωγή

Η Συσταδοποίηση δεδομένων (data clustering) είναι μια τεχνική στατιστικής ανάλυσης δεδομένων και χρησιμοποιείται σε πολλούς τομείς όπως: η μηχανική μάθηση, η εξόρυξη δεδομένων, η αναγνώριση προτύπων, η ανάλυση εικόνων και η βιοπληροφορική. Συσταδοποίηση είναι η ταξινόμηση όμοιων αντικειμένων σε διαφορετικές ομάδες ή αλλιώς ο καταμερισμός των δεδομένων σε υποσύνολα (clusters), έτσι ώστε τα δεδομένα να μοιράζονται κοινά χαρακτηριστικά, τα οποία συνήθως σχετίζονται με κάποια μετρική αποστάσεων. Υπό το πρίσμα της εκμάθησης μηχανών μια τέτοια ανάλυση θεωρείται μη επιβλεπόμενη (unsupervised) ενώ υπό το πρίσμα της εξόρυξης δεδομένων θεωρείται μέθοδος ερευνητικής ανάλυσης δεδομένων (exploratory data analysis). Ένα συχνό δίλημμα που προκύπτει στην συσταδοποίηση είναι πιο αλγόριθμος είναι ο κατάλληλος για τα διαθέσιμα δεδομένα. Το σίγουρο είναι ότι δεν υπάρχουν «καλοί» και «κακοί» αλγόριθμοι, αλλά η ποιότητα του καθενός μπορεί να κριθεί από τα αποτελέσματα που δίνει σε μια συγκεκριμένη εφαρμογή.

2.2 Αλγόριθμοι Συσταδοποίησης

Οι αλγόριθμοι συσταδοποίησης μπορεί να είναι ιεραρχικοί (hierarchical) ή μη ιεραρχικοί (non-hierarchical/partitional). Οι ιεραρχικοί αλγόριθμοι βρίσκουν διαδοχικές ομάδες, χρησιμοποιώντας κάθε φορά ήδη καθιερωμένες ομάδες, ενώ οι μη ιεραρχικοί καθορίζουν τις ομάδες αμέσως. Οι ιεραρχικοί αλγόριθμοι χωρίζονται στους συσσωρευτικούς (agglomerative) και στους διαχωριστικούς (divisive). Οι πρώτοι αντιμετωπίζουν κάθε στοιχείο σαν μια συστάδα από μόνο του και στη συνέχεια συγχωνεύεται σε μεγαλύτερες ομάδες. Οι δεύτεροι ξεκινούν με ολόκληρο το σύνολο και το διασπούν σε μικρότερες ομάδες. Συν-συσταδοποίηση (co-clustering, two-way clustering, bi-clustering) είναι η συσταδοποίηση, όπου εκτός από τα αντικείμενα ομαδοποιούνται και τα χαρακτηριστικά των αντικειμένων. Αν δηλαδή τα δεδομένα απεικονίζονται σε ένα πίνακα δεδομένων, τότε ομαδοποιούνται και οι γραμμές και οι στήλες. Μια άλλη σημαντική διάκριση είναι αν η συσταδοποίηση χρησιμοποιεί συμμετρικές ή ασύμμετρες αποστάσεις. Π.χ. μια ιδιότητα του Ευκλείδειου χώρου είναι ότι οι αποστάσεις είναι συμμετρικές, κάτι που δεν είναι εφικτό σε όλες τις εφαρμογές.

2.3 Ιεραρχική Συσταδοποίηση

Τα βασικά βήματα του ιεραρχικού αλγόριθμου (Johnson, 1967). πάνω σε ένα σύνολο αντικειμένων N και έναν πίνακα αποστάσεων N*N είναι τα εξής:

1. Κάθε αντικείμενο ανατίθεται σε μια συστάδα, έτσι ώστε αν έχουμε N αντικείμενα να έχουμε και N συστάδες. Οι αποστάσεις μεταξύ των συστάδων είναι ίδιες με τις αποστάσεις μεταξύ των αντικειμένων που περιέχονται στις συστάδες.

2. Εύρεση του πιο όμοιου ζευγαριού συστάδων και συγχώνευσή του σε μια συστάδα, έτσι ώστε να υπάρχει μια συστάδα λιγότερη.

3. Υπολογισμός των αποστάσεων μεταξύ της νέας συστάδας και των παλαιών συστάδων.

4. Επανάληψη των βημάτων 2 και 3, έως ότου όλα τα αντικείμενα να ανήκουν σε μια συστάδα μεγέθους N.

Το βήμα 3 γίνεται με 3 διαφορετικούς τρόπους:

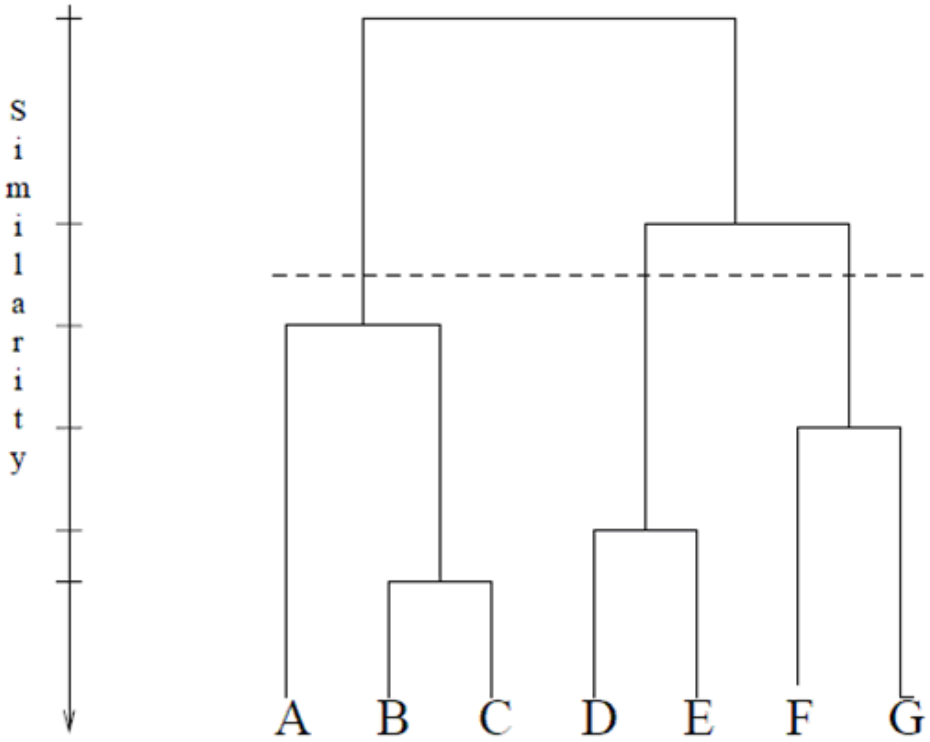
- Στην συσταδοποίηση **μονής σύνδεσης (single linkage clustering)** θεωρούμε την απόσταση μεταξύ μιας συστάδας και μιας άλλης να είναι ίση με την μικρότερη απόσταση μεταξύ του κάθε μέλους μιας συστάδας και των μελών της άλλης συστάδας.

- Στην συσταδοποίηση **ολικής σύνδεσης (complete linkage clustering)** θεωρούμε απόσταση μεταξύ δυο συστάδων την μέγιστη απόσταση μεταξύ του κάθε μέλους της μιας συστάδας με τα μέλη της άλλης συστάδας.

- Στην συσταδοποίηση **μέσης σύνδεσης (average linkage clustering)** η απόσταση μεταξύ μιας συστάδας και μιας άλλης είναι ίση με τη μέση απόσταση του κάθε μέλους από τα μέλη της άλλης συστάδας.

Το πρώτο βήμα αυτής της συσταδοποίησης είναι επιλογή της μετρικής αποστάσεων. Ένα είδος μετρικής είναι η απόσταση manhattan, ίση με το σύνολο των απόλυτων διαφορών για κάθε μεταβλητή. Άλλο είδος είναι η ευκλείδεια απόσταση, όπου υπολογίζεται η τετραγωνική απόσταση στην κάθε μεταβλητή, υπολογίζεται το άθροισμα όλων αυτών και τέλος υπολογίζεται η τετραγωνική ρίζα αυτού του αθροίσματος. Το δεύτερο βήμα μετά την επιλογή της μετρικής αποστάσεων είναι ο συνδυασμός των στοιχείων. Η ιεραρχική συσταδοποίηση δημιουργεί είτε με συσσωρευτικό (agglomerative) τρόπο είτε με διαχωριστικό τρόπο (divisive) την ιεραρχία των συστάδων. Η κλασσική μορφή αυτής της ιεραρχίας είναι το δενδρόγραμμα, όπου τα μεμονωμένα στοιχεία είναι από τη μια μεριά και η συστάδα με το κάθε στοιχείο από την άλλη. Οι συσσωρευτικοί αλγόριθμοι ξεκινάνε από την κορυφή του δέντρου, ενώ οι διαχωριστικοί από τον πάτο.

Το αποτέλεσμα του αλγόριθμου ιεραρχικού clustering είναι ένα και μοναδικό cluster που περιλαμβάνει όλα τα δεδομένα. Για την οπτικοποίηση του αποτελέσματος συνήθως χρησιμοποιούμε μια δεντρική δομή, η οποία φαίνετε στην παρακάτω εικόνα.



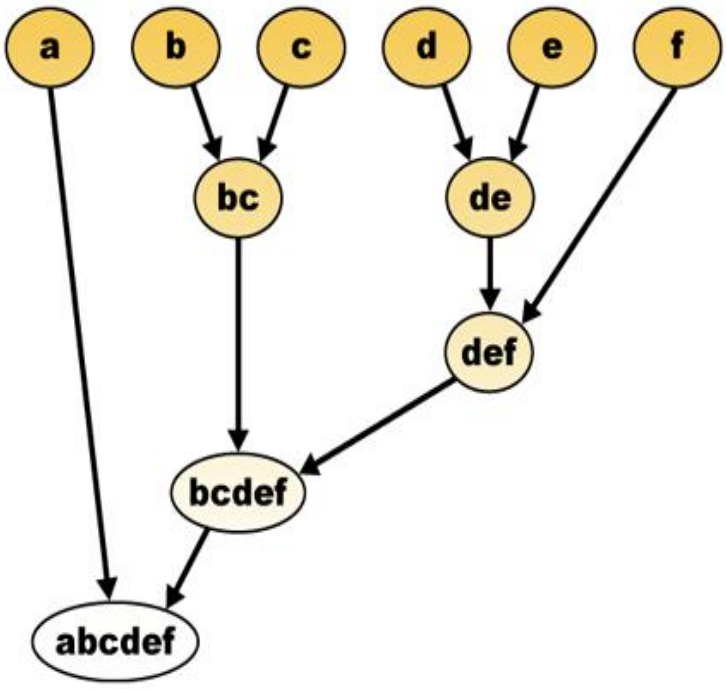
Agglomerative ιεραρχικό clustering

Εικόνα 3

Όπως βλέπουμε αρχικά στο κάτω μέρος υπάρχουν 7 cluster-αντικείμενα a, b, c, d, e, f, και g τα οποία ανεβαίνοντας από κάτω προς τα πάνω ομαδοποιούνται σε ένα και μοναδικό cluster. Υπάρχουν 2 βασικές στρατηγικές οι οποίες χρησιμοποιούνται για την οπτικοποίηση του παραπάνω δέντρου:

•Agglomerative: είναι μία προσέγγιση από κάτω προς τα πάνω, όπως φαίνετε και στην εικόνα. Αρχικά κάθε αντικείμενο είναι ένα cluster και όσο ανεβαίνουμε προς τα πάνω στην ιεραρχία ζευγαρώνει με άλλα cluster.

•Divisive: είναι η αντίθετη προσέγγιση από πάνω προς τα κάτω, όπως φαίνετε στην παρακάτω εικόνα όπου έχουμε 6 διαφορετικά clusters, a, b, c, d, e, f τα οποία συσταδοποιούντε από πάνω προς τα κάτω



Divisive ιεραρχικό clustering

Εικόνα 4

2.4 Παράδειγμα χρήσης του Hierarchical clustering

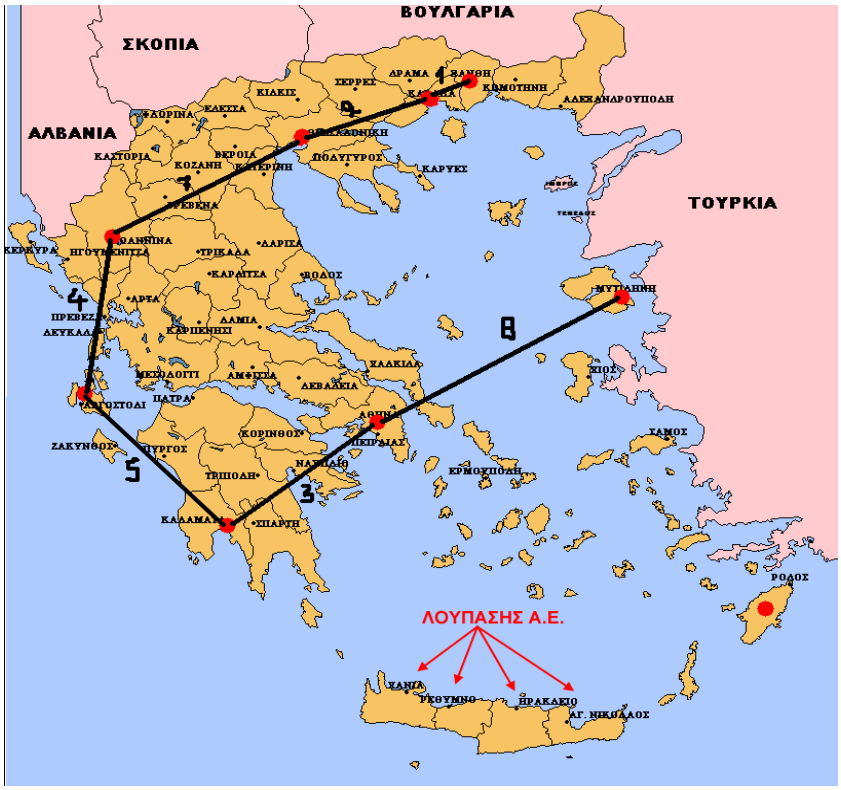
Παράδειγμα εικόνας πριν την εφαρμογή του ιεραρχικού clustering

Εικόνα 5

Παραπάνω βλέπουμε έναν χάρτη της Ελλάδας στον οποίο έχουν τονιστεί με μια κόκκινη κουκίδα ορισμένες πόλεις (Καβάλα, Θεσσαλονίκη, Ξάνθη και άλλες).

Θα προσπαθήσουμε να εφαρμόσουμε την τεχνική του hierarchical clustering χρησιμοποιώντας single linkage στρατηγική σε αυτές τις πόλεις της Ελλάδας. Αρχικά ενώνουμε με μια γραμμή τις δύο κοντινότερες πόλεις όπως εμφανίζονται στον χάρτη, φυσικά ανάλογα με την χιλιομετρική τους απόσταση. Οι πιο κοντινές πόλεις εμφανίζονται να είναι η Ξάνθη με την Καβάλα οπότε ενώνουμε αυτές πρώτα.

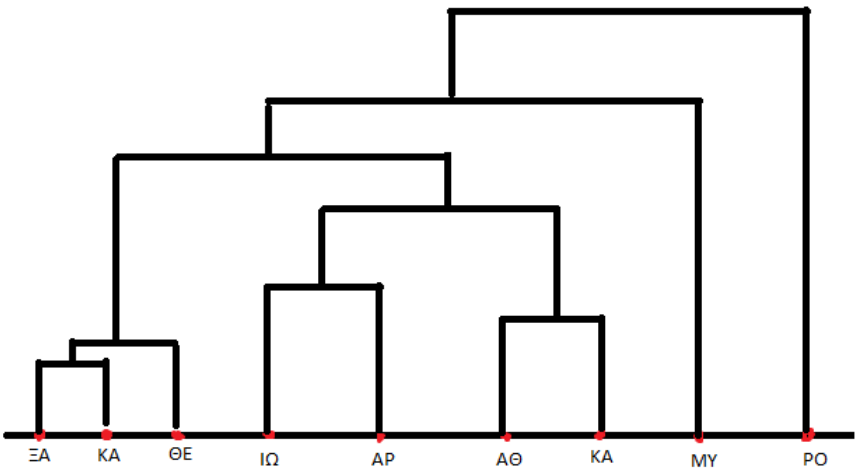
Έτσι τώρα η Καβάλα και η Ξάνθη αποτελούν μαζί ένα καινούργιο cluster. Συνεχίζουμε ενώνοντας τις κοντινότερες πόλεις μεταξύ τους. Αυτή τη φορά μετράμε την απόσταση οποιασδήποτε πόλης από την Καβάλας και την Ξάνθη ως την μικρότερη απόσταση από οποιαδήποτε από αυτές τις δύο πόλεις. Συνεχίζοντας έτσι μέχρι να ενώσουμε όλες τις πόλεις τελικά παίρνουμε τον χάρτη:



Παράδειγμα εικόνας μετά την εφαρμογή του ιεραρχικού clustering

Εικόνα 6

Τέλος το ιεραρχικό δέντρο που έχει ως αποτέλεσμα η εφαρμογή του αλγόριθμου στις παραπάνω πόλεις τις Ελλάδας είναι το παρακάτω:



Παράδειγμα Ιεραρχικού δέντρου

Εικόνα 7

2.5 Ο Αλγόριθμος K-means

Ο k-means είναι ένας από τους πιο γνωστούς και πιο απλούς αλγόριθμους που λύνουν το πρόβλημα του clustering. Δημοσιεύτηκε για πρώτη φορά από τον McQueen το 1967. Φυσικά ανήκει στην κατηγορία του unsupervised learning δηλαδή τα δεδομένα μας δεν έχουν καμία ετικέτα και δεν γνωρίζουμε τίποτα για αυτά. Ο αλγόριθμος ακολουθεί μια απλή και εύκολη διαδικασία για να κατηγοριοποιήσει τα δοσμένα δεδομένα σε έναν συγκεκριμένο αριθμό από clusters. Η κύρια ιδέα του αλγορίθμου είναι να καθορίσουμε εμείς έναν συγκεκριμένο αριθμό από k κέντρα των clusters (centroids) όταν θα ξεκινάει ο αλγόριθμος, που θα συμβολίζουν φυσικά και τον αριθμό των τελικών clusters που θα έχουμε ως έξοδο του αλγορίθμου.

Το επόμενο βήμα που κάνει ο αλγόριθμος είναι να αναθέσει κάθε δεδομένο (datum) στο κοντινότερο του centroid. Όταν ανατεθούν όλα τα δεδομένα, ένα πρώιμο clustering έχει γίνει. Σε αυτό το σημείο επαναϋπολογίζουμε τα centroids με βάση τα καινούργια clusters που έχουν δημιουργηθεί και τα τοποθετούμε έτσι ώστε να κατοπτρίζουν το κέντρο των δεδομένων που ανήκουν στο cluster τους. Έτσι όταν έχουμε k καινούργια centroids επαναϋπολογίζουμε τις θέσεις των δεδομένων και τα εναποθέτουμε στο κοντινότερο σε αυτά κέντρο. Με τον τρόπο αυτό βλέπουμε πως έχει δημιουργηθεί ένας βρόχος ο οποίος τερματίζετε όταν πλέον τα κέντρα δεν κουνιούνται από την θέση τους.

Με τον όρο K-means εννοούμε έναν αλγόριθμο ο οποίος αναθέτει κάθε αντικείμενο, στην συστάδα που έχει το κοντινότερο σε αυτήν μέσο. Η διαδικασία που ακολουθείται αποτελείται από τα εξής βήματα:

- 1. Χωρίζουμε τα αντικείμενα σε *K* αρχικές συστάδες και υπολογίζουμε τον μέσο(κέντρο) της συστάδας
- 2. Ανατρέχουμε όλα τα αντικείμενα, αναθέτοντας καθένα από αυτά στη συστάδα με της οποίας το μέσο είναι πιο κοντά. Ο υπολογισμός αυτής της απόστασης μεταξύ κάθε αντικειμένου και του κέντρου της συστάδας είναι συνήθως η Ευκλείδεια απόσταση όπως ορίστηκε παραπάνω.
- 3. Αφού γίνουν οι νέες αναθέσεις των αντικειμένων σε συστάδες σύμφωνα με αυτό τον κανόνα, υπολογίζουμε εκ νέου το κέντρο κάθε συστάδας.
- 4. Επαναλαμβάνουμε τη διαδικασία έως ότου να μην μπορούν να γίνουν περαιτέρω ανακατατάξεις.

Αντί να ξεκινάμε με μία διαίρεση όλων των αντικειμένων σε Κ αρχικές συστάδες θα μπορούσαμε να ορίσουμε *K* αρχικά κομβικά σημεία και να προχωρήσουμε από εκεί στο δεύτερο βήμα. Η τελική ανάθεση των αντικειμένων σε συστάδες θα εξαρτάται σε κάποιο βαθμό από τον αρχικό διαχωρισμό των συστάδων ή την επιλογή των σημείων. Στην πράξη οι περισσότερες αλλαγές στην ανάθεση των αντικειμένων συμβαίνουν κατά το πρώτο βήμα ανακατανομής.

Λόγοι για να μην έχουμε σταθερό αριθμό συστάδων *K* κατά την εκτέλεση του αλγορίθμου

- Αν δύο ή περισσότερα κομβικά σημεία βρίσκονται σε μία συστάδα οι παραγόμενες συστάδες θα είναι ανεπαρκώς διαφοροποιημένες
- Η ύπαρξη μιας ακραίας τιμής (outlier) μπορεί να δώσει τουλάχιστον μια συστάδα με πολύ διασκορπισμένα αντικείμενα.
- Ακόμη και αν ο πληθυσμός αποτελείται από ***K*** συστάδες, η δειγματοληπτική μέθοδος μπορεί να είναι τέτοια που δεδομένα από την πιο σπάνια συστάδα να μην εμφανίζονται στο δείγμα. Το να εξαναγκάζουμε τα δεδομένα σε ***K*** συστάδες μπορεί να οδηγήσει στη δημιουργία συστάδων που δεν έχουν νόημα.

Αν ο αλγόριθμος απαιτεί να οριστεί ο αριθμός των συστάδων , είναι προτιμότερο να τρέχουμε τον αλγόριθμο για αρκετές τιμές του ***K ,***έτσι ώστε να καταλήξουμε σε ένα ασφαλές συμπέρασμα όσον αφορά την τελική σύνθεση των συστάδων. Ο αλγόριθμος k-means (k-μέσων) είναι ένας αλγόριθμος που ομαδοποιεί αντικείμενα βάσει των χαρακτηριστικών των και μεριδίων. Αποτελεί μεταβλητή του αλγόριθμου μεγιστοποίησης αναμονής (expectation-maximization algorithm-EM), όπου σκοπός είναι να οριστεί ο k-means δεδομένων που προήλθαν από Gaussian κατανομές. Ο αλγόριθμος υποθέτει ότι τα χαρακτηριστικά του αντικειμένου δημιουργούν ένα χώρο διανυσμάτων και ο σκοπός του είναι να ελαχιστοποιήσει τη συνολική διακύμανση της συστάδας ή τη συνάρτηση τετραγωνικού σφάλματος.

Όπου υπάρχουν k ομάδες *Si*, *i* = 1,2,...,*k* και μ*i* είναι το κεντροειδές ή το μεσαίο σημείο από όλα τα σημεία

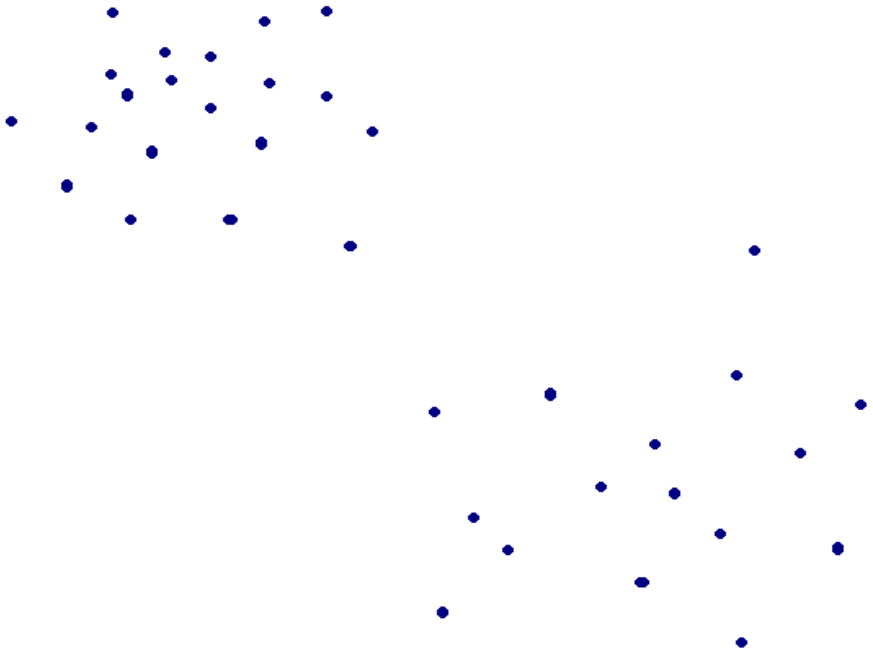
Τα βασικά βήματα του αλγόριθμου είναι τα εξής:

1. Επιλογή του αριθμού των συστάδων.
2. Τυχαία δημιουργία k συστάδων και ορισμός των κεντροειδών των συστάδων.
3. Μεταβίβαση του κάθε σημείου στο κεντροειδές της κοντινότερης συστάδας.
4. Υπολογισμός των νέων κεντροειδών των συστάδων.
5. Επανάληψη μέχρι να συγκλίνει ο αλγόριθμος σε κάποιο κριτήριο.

Ο αλγόριθμος ξεκινά διαχωρίζοντας τα αρχικά σημεία σε k αρχικά σύνολα είτε τυχαία είτε χρησιμοποιώντας ευριτικά δεδομένα. Στη συνέχεια υπολογίζει το μεσαίο ή το κεντροειδές του κάθε συνόλου, υλοποιεί νέο διαχωρισμό ώστε το κάθε σημείο να σχετίζεται με το κοντινότερο κεντροειδές. Έπειτα τα κεντροειδή ξανά υπολογίζονται για τις νέες ομάδες, ο αλγόριθμος επαναλαμβάνει τα δυο βήματα ωσότου τα σημεία δεν μπορούν να αλλάξουν ομάδες (ή εναλλακτικά τα κεντροειδή παραμένουν αμετάβλητα). Ο αλγόριθμος αυτός παραμένει διάσημος επειδή τείνει σε κάποιο όριο πολύ γρήγορα. Όσον αφορά την απόδοση ο αλγόριθμος δεν εγγυάται ότι θα αγγίξει το βέλτιστο. Η ποιότητα της τελική λύσης εξαρτάται πολύ από το αρχικό σύνολο συστάδων και μπορεί να είναι πολύ χαμηλότερη από το συνολικό βέλτιστο. Επίσης ένα άλλο μειονέκτημα του αλγόριθμου είναι ότι ο αριθμός των συστάδων πρέπει να οριστεί εξαρχής.

2.6 Παράδειγμα χρήσης του k-means

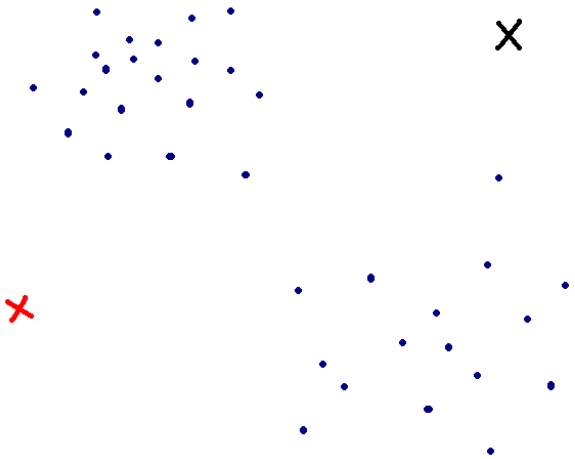
Όπως βλέπουμε στην παρακάτω εικόνα έχουμε ένα σύνολο δεδομένων, τα οποία και θα συσταδοποιήσουμε.



Σύνολο δεδομένων προς συσταδοποίηση

Εικόνα 8

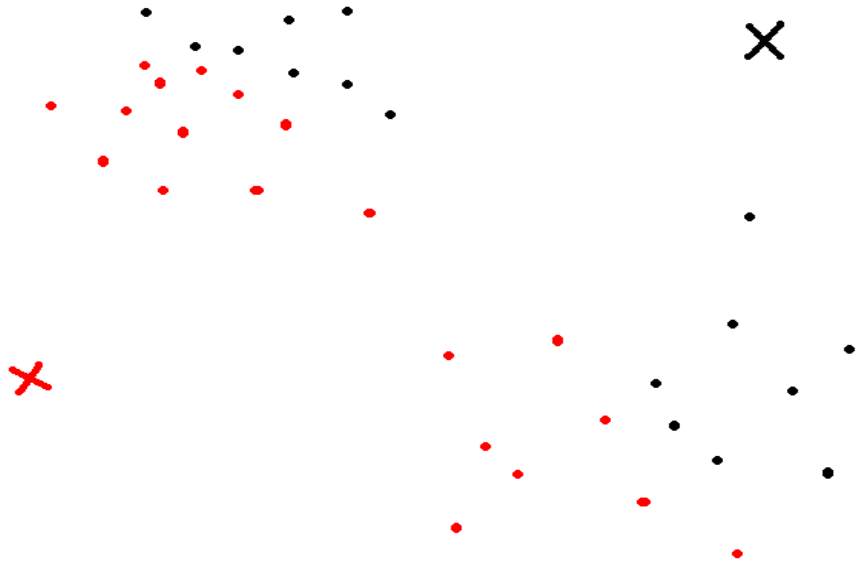
Αρχικά τοποθετούμε 2 σημεία στον χώρο των δεδομένων. Αυτά τα σημεία αντιπροσωπεύουν τα αρχικά κέντρα των clusters.



Αρχικά κέντρα των clusters

Εικόνα 9

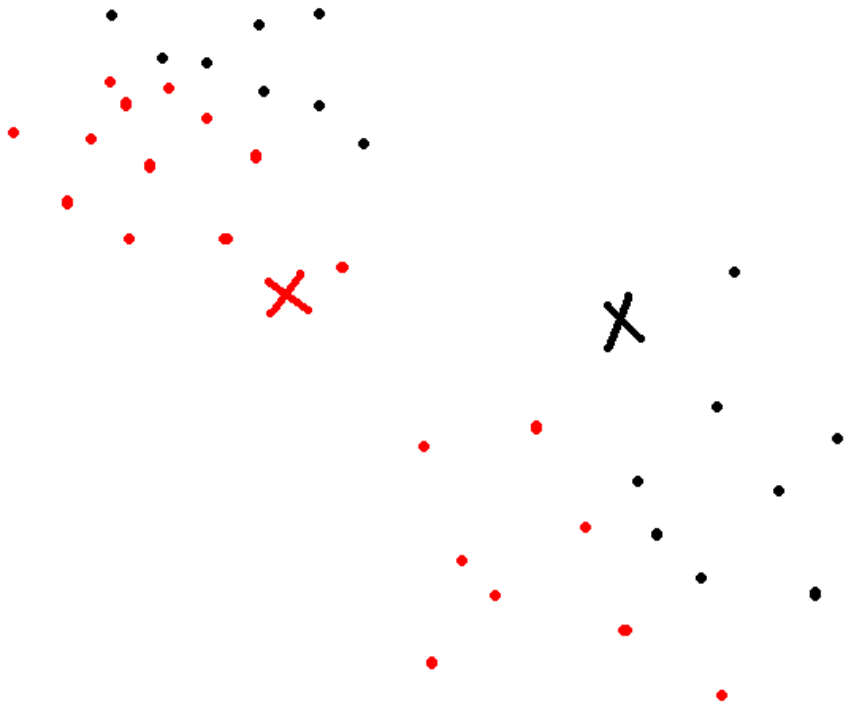
Έπειτα αναθέτουμε κάθε δεδομένο στην συστάδα που έχει το κοντινότερο κέντρο, όπως φαίνετε στην εικόνα.



Ανάθεση δεδομένων στις συστάδες

Εικόνα 10

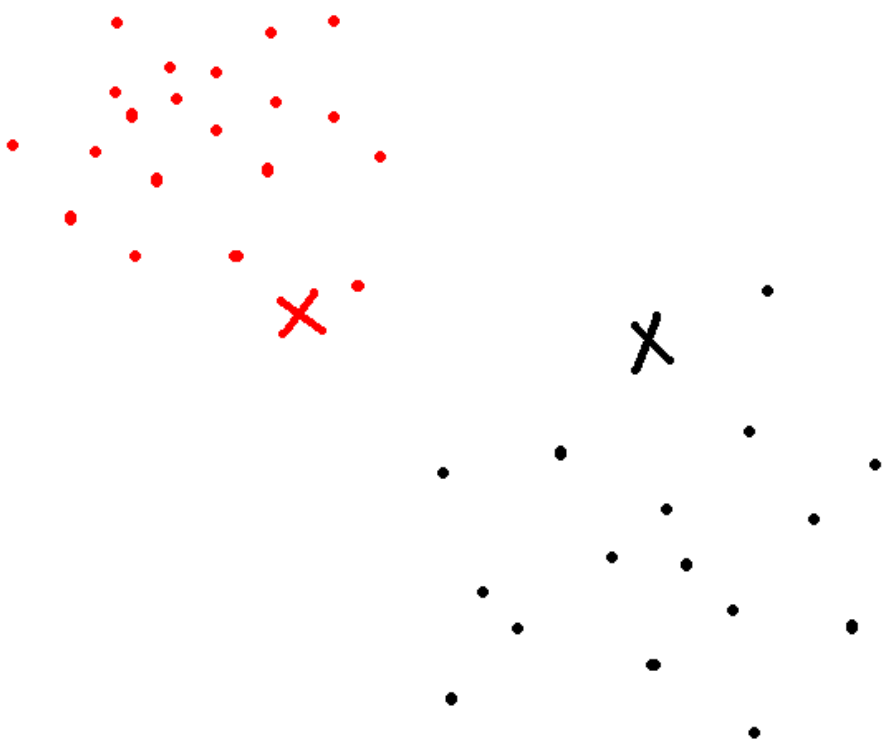
Στην συνέχεια επαναϋπολογίζουμε τα κέντρα των συστάδων.



Επαναυπολογισμός κέντρων

Εικόνα 11

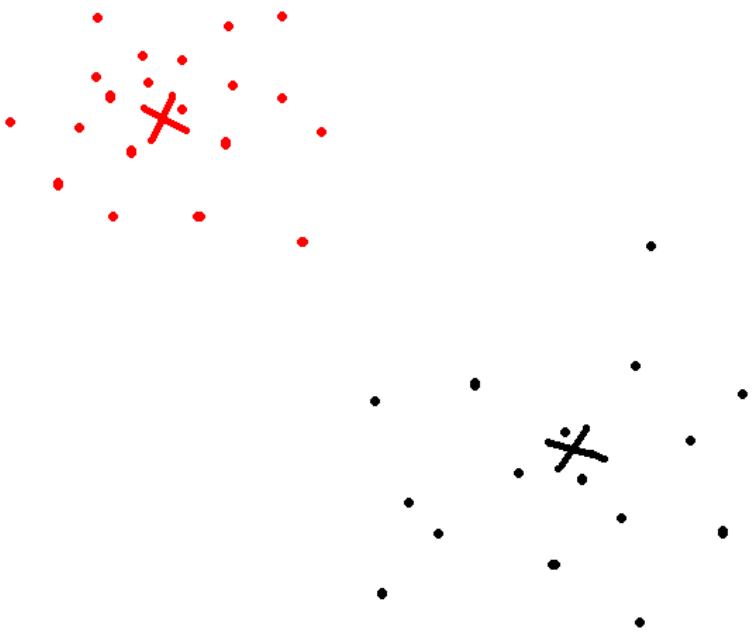
Έπειτα αναθέτουμε κάθε δεδομένο στην συστάδα που έχει το κοντινότερο κέντρο.



Ανάθεση δεδομένων στην κοντινότερη συστάδα

Εικόνα 12

Τέλος επαναϋπολογίζουμε τις θέσεις των κέντρων.



Επαναυπολογισμός κέντρων

Εικόνα 13

Όσες φορές και να εκτελεστεί ο αλγόριθμος από εδώ και πέρα τα κέντρα δεν θα αλλάξουν θέση. Δηλαδή ο αλγόριθμος τερματίζεται.

ΚΕΦΑΛΑΙΟ 3ο

<< Μέθοδοι κατηγοριοποίησης - K-NN (k nearest neighbors)>>

3.1 Εισαγωγή

Ο αλγόριθμος k-NN [Mitchell 1997] αποτελεί έναν από τους πιο χαρακτηριστικούς αλγόριθμους για μάθηση βασισμένη σε παραδείγματα. Η πλήρης ονομασία του είναι αλγόριθμος των k κοντινότερων γειτόνων (Nearest Neighbor – NN).

3.2 Ο αλγόριθμος k-NN

Ο αλγόριθμος k-NN (k nearest neighbor = k κοντινότερου γείτονα) είναι επίσης ένας απλός αλγόριθμος. Είναι ένας instance-based αλγόριθμος μάθησης, δηλαδή η διαδικασία μάθησης αφορά απλά την αποθήκευση των δεδομένων εκπαίδευσης και θεωρείται ένας από τους πιο απλούς αλγόριθμους μάθησης. Τα δεδομένα εκπαίδευσης τυγχάνουν επεξεργασίας όταν εμφανιστεί ένα νέο instance για αυτό και ονομάζεται Lazy Learning. Κάθε φορά που ένα νέο instance πρόκειται να ταξινομηθεί, υπολογίζεται η ομοιότητα του με κάθε ένα από τα αποθηκευμένα δεδομένα εκπαίδευσης.

Ο αλγόριθμος KNN βασίζεται σε μια συνάρτηση απόστασης όπως είναι η Ευκλείδεια απόσταση και η απόσταση συνημίτονου, μεταξύ κάθε εγγράφου εκπαίδευσης και του εγγράφου που πρόκειται να ταξινομηθεί. Ο αλγόριθμος KNN αναθέτει το έγγραφο που πρόκειται να ταξινομηθεί στην κατηγορία στην οποία ανήκει η πλειοψηφία των k κοντινότερων γειτόνων, όπου το k είναι ένας θετικός ακέραιος, συνήθως μικρός και δίνεται ως παράμετρος. Οι κοντινότεροι γείτονες υπολογίζονται χρησιμοποιώντας κάποια συνάρτηση απόστασης. Τα έγγραφα συνήθως αναπαριστούνται με το vector space μοντέλο, δηλαδή ως διανύσματα όπου κάθε συνιστώσα τους αντιστοιχεί σε κάποιο όρο που εμφανίζεται στο λεξικό, μαζί με κάποιος βάρος για κάθε όρο. Αν ένας όρος δεν εμφανίζεται σε κάποιο συγκεκριμένο έγγραφο, το βάρος του είναι μηδέν. Στη συνέχεια υπολογίζεται η ομοιότητα(similarity) κάθε εγγράφου με το έγγραφο το οποίο πρόκειται να ταξινομηθεί. Αυτό είναι δυνατό με την χρήση κάποιας συνάρτησης που υπολογίζει την απόσταση μεταξύ δυο διανυσμάτων. Μερικά παραδείγματα είναι: Euclidean distance(ευκλείδεια απόσταση) και Manhattan distance.
Από τους προηγούμενους υπολογισμούς θα εξαχθούν τα k έγγραφα τα οποία έχουν την μεγαλύτερη ομοιότητα με το έγγραφο που πρόκειται να ταξινομηθεί(δηλαδή την μικρότερη απόσταση). Αυτοί ονομάζονται οι κοντινότεροι γείτονες. Τέλος, το έγγραφο θα ανατεθεί στην κλάση στην οποία ανήκουν οι περισσότεροι κοντινότεροι γείτονες. Κάποιες παραλλαγές στον αλγόριθμο μπορεί να είναι οι εξής: Αντί να λαμβάνεται υπόψη μόνο αν σε ποια κλάση ανήκουν οι k κοντινότεροι γείτονες, είναι καλύτερο να λαμβάνεται υπόψη και η απόσταση που έχουν από το έγγραφο που θα ταξινομηθεί. Με άλλα λόγια αντί το έγγραφο να ανατίθεται στην κλάση στην οποία ανήκουν οι περισσότεροι από τους k κοντινότερους γείτονες του, να ανατίθεται στη κλάση με την μεγαλύτερη ομοιότητα.

Το πλεονέκτημα του αλγορίθμου αυτού είναι ότι είναι εύκολος να κατανοηθεί και να υλοποιηθεί. Ο αλγόριθμος k-NN δουλεύει καλά σε περιπτώσεις multi-modal κλάσεων και σε εφαρμογές όπου κάποιο αντικείμενο μπορεί να ανήκει σε περισσότερο από μια κλάση. Επίσης, είναι πολύ αποδοτικός σε περιπτώσεις όπου τα δεδομένα εκπαίδευσης περιέχουν θόρυβο(noisy) και σε περιπτώσεις όπου τα δεδομένα εκπαίδευσης είναι πολλά. Το σημαντικότερο μειονέκτημα του είναι ότι είναι instance-based αλγόριθμος μάθησης, οπότε δεν γίνεται οποιαδήποτε εκπαίδευση, μέχρι να φτάσει κάποιο έγγραφο για ταξινόμηση και επίσης έχει μεγάλο κόστος υπολογισμού γιατί πρέπει να υπολογίσει την απόσταση κάθε όρου του κειμένου που θα ταξινομηθεί με όλα τα έγγραφα εκπαίδευσης. Ένα άλλο σημαντικό μειονέκτημα είναι ότι χρειάζεται να καθοριστεί κάποια τιμή για το k. Το k παίζει αρκετά σημαντικό ρόλο στην αποδοτικότητα του ταξινομητή και είναι δύσκολο να προσδιοριστεί. Αν είναι πολύ μικρό, το αποτέλεσμα μπορεί να είναι ευαίσθητο σε θορυβώδη δεδομένα. Αν είναι πολύ μεγάλο, το αποτέλεσμα των κοντινότερων γειτόνων μπορεί να περιέχει πολλά έγγραφα από άλλες κατηγορίες. Το k μπορεί να οριστεί χρησιμοποιώντας διάφορες τεχνικές οι οποίες χρησιμοποιούν ευρετικά. Τέλος, πρέπει να καθοριστεί πια συνάρτηση απόστασης πρέπει να εφαρμοστεί για να προκύψουν τα καλύτερα αποτελέσματα.

Για τη χρήση και την αξιολόγηση του αλγορίθμου θα πρέπει να υπάρχουν δυο σύνολα από αντικείμενα: το σύνολο εκπαίδευσης train και το σύνολο ελέγχου test. Τα αντικείμενα των συνόλων αυτών πρέπει να έχουν καταταγεί χειρωνακτικά σε δύο ή περισσότερες κατηγορίες (c1, c2, c3, …, ck), των οποίων η τομή ανά δύο πρέπει να είναι κενή, δηλαδή δεν μπορεί ένα αντικείμενο να ανήκει σε περισσότερες από μία κατηγορίες. Επίσης θα πρέπει να έχει οριστεί ένα σύνολο ιδιοτήτων $A = \{ X1 , X 2 , X 3 ,..., X m\}$. Κάθε αντικείμενο των συνόλων train και test παριστάνεται από ένα διάνυσμα $x = x1 , x2 , x3 ,..., xm\ r$, κάθε συντεταγμένη του οποίου αποτελεί την τιμή μιας συγκεκριμένης ιδιότητας xi.

Ο αλγόριθμος «εκπαιδεύεται» στα αντικείμενα του συνόλου train, ώστε να προβλέπει τη σωστή κατηγορία κάθε αντικειμένου βάση των τιμών του διανύσματός του. Η ακρίβεια του ταξινομητή που προκύπτει από την εκπαίδευση αξιολογείται στο σύνολο test, συγκρίνοντας τις αποφάσεις του ταξινομητή με τις σωστές κατηγορίες. Κατά την εκπαίδευση, ο αλγόριθμος k-NN απλά αποθηκεύει σε μια μνήμη όλα τα διανύσματα των αντικειμένων του συνόλου train και τις σωστές κατηγορίες τους. Η κατάταξη νέων αντικειμένων, των οποίων δεν είναι γνωστές οι κατηγορίες, γίνεται ως εξής: Υπολογίζεται η απόσταση του διανύσματος του νέου αντικειμένου από τα διανύσματα όλων των αντικειμένων εκπαίδευσης. Επιλέγονται τα k αντικείμενα εκπαίδευσης με τις μικρότερες αποστάσεις (οι k κοντινότεροι γείτονες) και το νέο αντικείμενο κατατάσσεται στην κατηγορία που πλειοψηφεί μεταξύ των k αντικειμένων. Ως μέτρο απόστασης μπορεί να χρησιμοποιηθεί, για παράδειγμα, η απόσταση Manhattan. Η απλότητα αλλά και η γενικότητα του παραπάνω αλγορίθμου παρέχει στον χρήστη την ευκολία να τον τροποποιήσει, προσθέτοντας δικά του μέτρα απόστασης, που θα επηρεάσουν την ευαισθησία του αλγορίθμου ως προς του k γείτονες και θα φέρουν τα αποτελέσματα πιο κοντά στις επιθυμητές τιμές. Η αναγνώριση των προτύπων , στον αλγόριθμο k-NN (k-κοντινότερο γείτονα) είναι μια μέθοδος για την ταξινόμηση των αντικειμένων με βάση ,το πιο κοντινό παράδειγμα , την κατάρτιση στο χώρο των χαρακτηριστικών .Ο αλγόριθμος k-NN είναι ένας τύπος παραδείγματος βασισμένος σε μια μορφή μάθησης , ή τεμπέλικους μάθησης όπου η λειτουργία του είναι μόνο κατά προσέγγιση σε τοπικό επίπεδο και όλος ο υπολογισμός αναβάλλεται μέχρι την ταξινόμηση. Ο αλγόριθμος k-NN (k-κοντινότερο γείτονα) είναι ο πιο απλός από όλους τους αλγόριθμους μηχανικής μάθησης: ένα αντικείμενο που έχει χαρακτηριστεί από την πλειοψηφία των γειτόνων του, με το αντικείμενο που υπάγονται στην κλάση πιο κοινά μεταξύ k κοντινότερους γείτονές της (k είναι ένας θετικός ακέραιος, συνήθως μικρό). Αν k = 1, τότε το αντικείμενο απλά υπάγεται στην κλάση του πλησιέστερου γείτονά της.

Η ίδια μέθοδος μπορεί να χρησιμοποιηθεί για την οπισθοδρόμηση , με την απλή ανάθεση της αξίας του ακινήτου για το αντικείμενο που είναι ο μέσος όρος των τιμών των k πλησιέστερων γειτόνων της. Μπορεί να είναι χρήσιμο για τη στάθμιση των εισφορών των γειτόνων, έτσι ώστε οι γείτονες πιο κοντά συμβάλλει περισσότερο με το μέσο όρο από ό, τι το πιο μακρινά. (Ένα κοινό σύστημα στάθμισης είναι να δοθεί σε κάθε γείτονα βάρους 1 / d, όπου d είναι η απόσταση από το γείτονα. Το σύστημα αυτό είναι μια γενίκευση της γραμμικής παρεμβολής.) Οι γείτονες που λαμβάνονται από ένα σύνολο αντικειμένων για τα οποία η σωστή κατάταξη (ή, στην περίπτωση της παλινδρόμησης, η αξία του ακινήτου) είναι γνωστή. Αυτό μπορεί να θεωρηθεί ως η εκπαίδευση που για τον αλγόριθμο, αν και καμία ρητή βήμα εκπαίδευση είναι απαραίτητη. Ο k-κοντινότερος γείτονας αλγόριθμος είναι ευαίσθητος στην τοπική δομή των δεδομένων. Οι Κοντινότεροι κανόνες γείτονα σε ισχύ υπολογίζουν το όριο απόφασης με έμμεσο τρόπο. Είναι επίσης δυνατό να υπολογιστεί το όριο απόφασης ρητά και μάλιστα με αποτελεσματικό τρόπο, έτσι ώστε η πολυπλοκότητα είναι συνάρτηση της πολυπλοκότητας του ορίου. Ο k-NN αλγόριθμος μπορεί επίσης να προσαρμοστεί για την χρήση στην εκτίμηση συνεχών μεταβλητών. Μια τέτοια εφαρμογή χρησιμοποιείται για μια αντίστροφη απόσταση σταθμισμένου μέσου όρου των k-κοντινότερων πολυμεταβλητων γειτόνων. Αυτός ο αλγόριθμος λειτουργεί ως εξής:

1. Υπολογίστε την Ευκλείδεια απόσταση από τους στόχος σε εκείνα τα σημεία που περιλήφθηκαν στο δείγμα.
2. Καταγράψτε τα δείγματα για αποστάσεις που υπολογίσατε .
3. Επιλέξτε τον βέλτιστο πλησιέστερο γείτονα στο *k.*
4. Υπολογίστε μια αντίστροφη απόσταση σταθμισμένου μέσου όρου με το *k*-κοντινότερο γείτονα.

Χρησιμοποιώντας έναν σταθμισμένο k-NN επίσης βελτιώνει σημαντικά τα αποτελέσματα: η κατηγορία (ή την αξία, σε προβλήματα παλινδρόμησης) από καθένα από τα σημεία k πλησιέστερο πολλαπλασιάζεται με συντελεστή ανάλογο με το αντίστροφο της απόστασης μεταξύ του σημείου και το σημείο για το οποίο η κατηγορία πρέπει να προβλεφθεί. Η κατηγοριοποίηση αυτή βασίζεται στην εύρεση και στην εξέταση κοντινότερων γειτόνων (κατηγοριοποίηση KNN) είναι μια από τις πιο γνωστές μεθόδους κατηγοριοποίησης. Ο αριθμός κοντινότερων γειτόνων, ο οποίος χρησιμοποιείται για την επίτευξη της κατηγοριοποίησης με τη μεγαλύτερη δυνατή ακρίβεια, είναι σταθερός και γνωστός εκ των προτέρων. Όταν το μέγεθος της Βάσης Δεδομένων όπου αναζητούνται οι κοντινότεροι γείτονες είναι μεγάλο, οι αλγόριθμοι της σειριακής και της δυαδικής αναζήτηση δε μπορούν να εφαρμοστούν εξαιτίας του χρόνου που απαιτούν. Έτσι είναι απαραίτητη η χρήση κάποιας δομής δεδομένων, όπως το R-Tree, που να δεικτοδοτεί τα δεδομένα έτσι ώστε να επιτρέπεται η γρήγορη αναζήτηση. Επίσης, αν ο αριθμός των κοντινότερων γειτόνων που δίνει την υψηλότερη τιμή ακρίβειας κατηγοριοποίησης είναι μεγάλος, τότε, ακόμη και όταν τα δεδομένα δεικτοδοτούνται, η διαδικασία της αναζήτησης εξακολουθεί να είναι χρονοβόρα. Η εργασία αυτή, εκτός από τη βιβλιογραφική παρουσίαση των μεθόδων κατηγοριοποίησης, δεικτοδότησης δεδομένων (χωρικών ή μη) και αναζήτησης κοντινότερων γειτόνων σε δομές δεδομένων δεικτών, προτείνει μια νέα εκδοχή KNN κατηγοριοποίησης, η οποία ονομάζεται κατηγοριοποίηση με βάση δυναμικό αριθμό κοντινότερων γειτόνων και σκοπεύει στη διακοπή του αλγόριθμου αναζήτησης κοντινότερων γειτόνων όταν ικανοποιούνται κάποια κριτήρια κόστους και ακρίβειας.

Η εργασία προτείνει τρεις ευρεστικές μεθόδους που χρησιμοποιούνται για την υλοποίηση της νέας προσέγγισης. Οι μέθοδοι αυτοί εισάγουν κριτήρια διακοπής και σκοπεύουν στη μεγαλύτερη δυνατή μείωση του χρόνου που απαιτείται από τη διαδικασία κατηγοριοποίησης κρατώντας την ακρίβεια σε υψηλά επίπεδα. Το σημείο όπου θα γίνει η διακοπή εξαρτάται από το αν ο αριθμός των γειτόνων που έχουν βρεθεί μέχρι τη δεδομένη χρονική στιγμή, αρκεί ώστε να εκτελεστεί κατηγοριοποίηση ενός αντικειμένου με υψηλή ακρίβεια. Εύκολα καταλαβαίνει κανείς ότι ο αριθμός γειτόνων που χρησιμοποιείται είναι δυναμικός. Δηλαδή κάθε αντικείμενο προς κατηγοριοποίηση, τοποθετείται σε κατηγορία βάσει διαφορετικού και μη προκαθορισμένου αριθμού κοντινότερων γειτόνων. Η καταλληλότητα της νέας προσέγγισης KNN κατηγοριοποίησης αποδεικνύεται από τα πειραματικά αποτελέσματα που παρουσιάζονται σε αυτή την εργασία.

3.3 Δένδρα απόφασης

Ο αλγόριθμος ID3 (**Iterative Dichotomiser 3**) είναι ένας αλγόριθμος που χρησιμοποιείται για να δημιουργήσει ένα δέντρο αποφάσεων και εφευρέθηκε από τον Ross Quinlan. Είναι ένας αρκετά διαδεδομένος αλγόριθμος επαγωγικής μάθησης, με τον οποίο κατασκευάζονται δένδρα απόφασης. Η λογική του βασίζεται στην χρήση των ιδιοτήτων, μία προς μία, για την διάσπαση του συνόλου των αντικειμένων εκπαίδευσης στους κλάδους ενός δένδρου. Σε γενικές γραμμές, εκτελούνται αναδρομικά τα παρακάτω βήματα :

1. Επιλέγεται η ιδιότητα με το μεγαλύτερο πληροφοριακό κέρδος, εκτιμώντας τις πιθανότητες στο σύνολο εκπαίδευσης.
2. Τοποθετείται στη ρίζα του δένδρου ένας έλεγχος για την ιδιότητα με το μεγαλύτερο πληροφοριακό κέρδος και δημιουργείται ένας κλάδος κάτω από τη ρίζα για κάθε μία δυνατή τιμή της ιδιότητας.
3. Τα αντικείμενα του συνόλου εκπαίδευσης κατανέμονται στους κλάδους, ανάλογα με την τιμή της ιδιότητας που χρησιμοποιήθηκε στη ρίζα.
4. Κάθε κλάδος οδηγεί σε ένα υποδένδρο που κατασκευάζεται αναδρομικά, χρησιμοποιώντας ως σύνολο εκπαίδευσης το υποσύνολο που αντιστοιχεί στον κλάδο και ως σύνολο ιδιοτήτων το αρχικό μείον την ιδιότητα που χρησιμοποιήθηκε στην ρίζα.

Το δένδρο απόφασης διαχωρίζει τα αντικείμενα εκπαίδευσης σε ομάδες. Κατά το στάδιο της κατάταξης αντικειμένων των οποίων η κατηγορία είναι άγνωστη, εντοπίζεται πρώτα η «κοντινότερη» συστάδα χρησιμοποιώντας το δέντρο απόφασης και το αντικείμενο κατατάσσεται στην κατηγορία που πλειοψηφεί σε εκείνη την συστάδα αντί (όπως στην περίπτωση του k-NN) για την κατηγορία που πλειοψηφεί στο σύνολο των αντικειμένων εκπαίδευσης.

3.4 Αλγόριθμος (NNS) - Nearest neighbor search

Η αναζήτηση κοντινότερη γείτονα (NNS), γνωστή και ως αναζήτηση εγγύτητας, αναζήτηση ομοιότητα ή αναζήτηση του πλησιέστερου σημείου, είναι ένα πρόβλημα βελτιστοποίησης για την εύρεση του πλησιέστερου σημείου σε μετρικούς χώρους. Το πρόβλημα είναι το εξής: δίνεται ένα σύνολο S των σημείων σε ένα μετρικό χώρο M και ένα ερώτημα το σημείο $q \in M$, βρείτε το πλησιέστερο σημείο στο S για το q. Σε πολλές περιπτώσεις , το M θεωρείται ότι είναι σε d-διάστατο Ευκλείδειο χώρο και η απόσταση μετράται από Ευκλείδεια απόσταση ή απόσταση Manhattan. Η ταξινόμηση Κοντινής Γειτονιάς είναι μία στατιστική μέθοδος επιβλεπόμενης ταξινόμησης, δηλαδή είναι γνώστες οι κατηγορίες του προβλήματος και ένας αριθμός προτύπων που ανήκουν σε αυτές. Ο αλγόριθμος της μεθόδου ταξινομεί κάθε πρότυπο σε εκείνη τη κατηγορία από την οποία απέχει λιγότερο, με βάση μία απόσταση, π.χ. Ευκλείδεια απόσταση. Η μέθοδος αυτή μπορεί να έχει τη μορφή της ταξινόμησης του κοντινότερου γείτονα ή των k-κοντινότερων γειτόνων.

3.5 Μέθοδος Κατηγοριοποίησης -Ταξινόμησης

Η μέθοδος ταξινόμησης Κοντινότερου Γείτονα (Nearest Neighbor-NN) περιγράφεται παρακάτω για το πρόβλημα της ταξινόμησης σε δύο κατηγορίες ή και σε περισσότερες από δύο κατηγορίες. Ας θεωρήσουμε δύο κατηγορίες K1, K2 και το προς ταξινόμηση πρότυπο **h**. Το πρότυπο **h** θα ταξινομηθεί στη κατηγορία, η οποία έχει κάποιο στοιχείο που να απέχει το λιγότερο δυνατό από αυτό. Μπορεί, δηλαδή, να οριστεί μια συνάρτηση απόφασης *f*(**h**) ως εξής:

f(**h**) = (Μικρότερη απόσταση από K1) - (Μικρότερη απόσταση από K2) ,όπου η απόσταση μεταξύ δύο διανυσμάτων x, y διάστασης *n* μπορεί να είναι μια από τις Εξής:

Ευκλείδεια Απόσταση : E =

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{3.1}$$

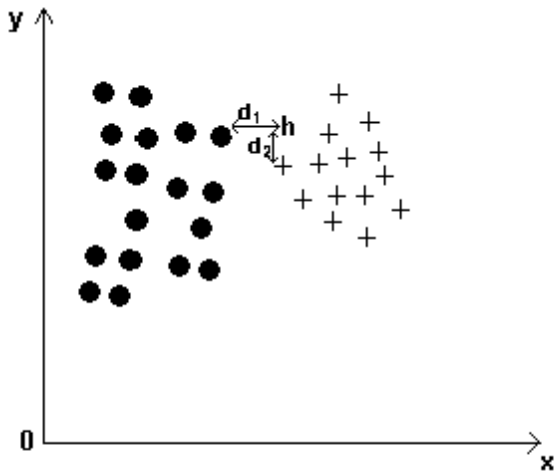
‘Ομοιως

$$\left(\sum_{i=1}^{n} |x_i - y_i|^2 \right)^{1/2} \tag{3.2}$$

Οπότε:

- Εάν f(**h**) < 0 τότε το **h** ανήκει στην κατηγορία K1
- Εάν f(**h**) > 0 τότε το **h** ανήκει στην κατηγορία K2

Στο παρακάτω σχήμα φαίνεται η ταξινόμηση ενός δισδιάστατου προτύπου **h** σε μία από τις κατηγορίες σύμφωνα με τη μέθοδο του κοντινότερου γείτονα.



Ταξινόμηση με τη μέθοδο K-NN

Εικόνα 14

Ταξινόμηση κοντινότερου γείτονα. Το πρότυπο **h** ταξινομείται στην κατηγορία K2, γιατί ισχύει **d2<d1.**

Στην πράξη το πρόβλημα δεν είναι τόσο απλό, γιατί οι κατηγορίες K_1 και K_2 δεν είναι τόσο ξεκάθαρα διαχωρίσιμες, όπως προϋποθέτει το παραπάνω κριτήριο διχοτομίας. Σε πολλές περιπτώσεις ένα πρότυπο μπορεί να ανήκει σε μια κατηγορία αλλά να βρίσκεται πλησιέστερα σε μια άλλη. Για να αποφευχθεί η δυσκολία αυτά μετράται η απόσταση του **h** από πολλά δείγματα προτύπων κάθε κατηγορίας, έτσι ώστε η επίδραση οποιουδήποτε διφορούμενου προτύπου να εξομαλυνθεί. Ο τρόπος αυτός ταξινόμησης λέγεται ταξινόμηση του k- κοντινότερου γείτονα, όπου k είναι το πλήθος των γειτονικών δειγμάτων, ως προς τα οποία μετράται η απόσταση και με βάση τα οποία γίνεται η ταξινόμηση του προτύπου.

Διάφορες λύσεις για το πρόβλημα NNS έχουν προταθεί. Η ποιότητα και η χρησιμότητα των αλγορίθμων που καθορίζεται από την χρονική πολυπλοκότητα των ερωτημάτων καθώς και την πολυπλοκότητα χώρου των οποιονδήποτε δομών δεδομένων αναζήτησης που πρέπει να διατηρηθούν. Η άτυπη παρατήρηση που συνήθως αναφέρεται ως η **κατάρα της διαστατικότητας** δηλώνει ότι δεν υπάρχει μια ακριβή γενικής χρήσης λύση για τον αλγόριθμο NNS σε μεγάλες διαστάσεις με Ευκλείδειο χώρο και με προεπεξεργασία πολυωνύμου και πολυλογαριθμικού χρόνου αναζήτησης.

3.6 Η κατάρα της διαστατικότητας

Η **κατάρα της διαστατικότητας** αναφέρεται σε διάφορα φαινόμενα που προκύπτουν κατά την ανάλυση και την οργάνωση μεγάλων διαστάσεων χώρου (συχνά με εκατοντάδες ή χιλιάδες διαστάσεις) που δεν απαντούν σε συνθήκες χαμηλού διαστάσεων ρυθμίσεις, όπως το φυσικό χώρο που συνήθως διαμορφωθεί με μόνο τρεις διαστάσεις. Υπάρχουν πολλαπλά φαινόμενα που αναφέρονται με αυτό το όνομα σε τομείς όπως η δειγματοληψία , η Συνδυαστική , η μηχανική μάθηση και η εξόρυξη δεδομένων . Ο κοινός παρονομαστής αυτών των προβλημάτων είναι ότι όταν αυξάνεται η διαστατικότητα, ο όγκος του χώρου αυξάνεται τόσο γρήγορα ώστε τα διαθέσιμα δεδομένα να αραιώνουν και να καθίστανται προβληματικά για κάθε μέθοδο που απαιτεί στατιστική σημασία. Αυτό έχει ελάχιστες αναφορές και είναι πρόβλημα για κάθε μέθοδο που απαιτεί στατιστική σημασία. Για να επιτευχθεί ένα στατιστικά ορθό ασφαλές και αξιόπιστο αποτέλεσμα, ο όγκος των δεδομένων θα πρέπει να υποστηρίξει το αποτέλεσμα το οποίο αυξάνεται συχνά εκθετικά με βάση τη διαστατικότητα.

Επίσης, η οργάνωση και η αναζήτηση των δεδομένων συχνά βασίζεται στον εντοπισμό περιοχών όπου τα αντικείμενα αποτελούν ομάδες με παρόμοιες ιδιότητες. Σε υψηλά επίπεδα διαστάσεων των δεδομένων, πάντως όλα τα αντικείμενα φαίνονται να έχουν αραιή και ανόμοια συμπεριφορά με πολλούς τρόπους που δεν επιτρέπουν κοινές στρατηγικές οργάνωσης των δεδομένων για να είναι αποτελεσματικά.

Ο όρος *κατάρα της διαστατικότητας* επινοήθηκε από Bellman Richard E. κατά την εξέταση των προβλημάτων στη βελτιστοποίηση δυναμικού.

Η "κατάρα της διαστατικότητας" χρησιμοποιείται συχνά ως δικαιολογία για να μην ασχολούνται με μεγάλες διαστάσεις δεδομένων. Ωστόσο, δεν είναι οι συνέπειες ακόμη πλήρως κατανοητές από την επιστημονική κοινότητα, και υπάρχει σε εξέλιξη η έρευνα. Από τη μία πλευρά, η έννοια της εγγενής διάστασης αναφέρεται στο γεγονός ότι οι χαμηλές διαστάσεις χώρου των δεδομένων μπορούν επιπόλαια να μετατραπούν σε ένα υψηλότερο επίπεδο διαστάσεων του χώρου, με την προσθήκη περιττών (π.χ. εις διπλούν) ή τυχαιοποιημένων διαστάσεων, και με τη σειρά τους πολλά σύνολα δεδομένων μεγάλων διαστάσεων μπορούν να μειωθούν σε κάποιο χαμηλότερο επίπεδο διαστάσεων των δεδομένων, χωρίς σημαντική απώλεια πληροφοριών. Αυτό φαίνεται και από την αποτελεσματικότητα της μείωσης της διάστασης των μεθόδων, όπως η ανάλυση κυρίων συστατικών σε πολλές καταστάσεις. Για τις λειτουργίες της απόστασης και της αναζήτησης του πλησιέστερου γείτονα, πρόσφατη έρευνα έδειξε επίσης ότι τα σύνολα δεδομένων που παρουσιάζουν την κατάρα των ιδιοτήτων διαστατικότητας μπορεί ακόμα να υποβάλλονται σε επεξεργασία, εκτός αν υπάρχουν πάρα πολλές άσχετες διαστάσεις, ενώ οι σχετικές διαστάσεις μπορεί να κάνει κάποια προβλήματα, όπως η ανάλυση διασποράς στην πραγματικότητα ευκολότερη.

Δεύτερον, οι μέθοδοι, όπως η Markov Chain Monte Carlo ή κοινές μεθόδους πλησιέστερο γείτονα , συχνά λειτουργούν πολύ καλά σε δεδομένα που θεωρήθηκε δυσεπίλυτο με άλλες μεθόδους, λόγω των υψηλών διαστάσεων του. Ο αλγόριθμος του πλησιέστερου γείτονα είναι εύκολο να εφαρμοστεί και εκτελεί γρήγορα, αλλά μερικές φορές μπορεί να χαθούν μικρότερες διαδρομές οι οποίες είναι εύκολα να παρατηρηθούν με την ανθρώπινη γνώση, λόγω του "άπληστου" της φύσης. Ως γενικός οδηγός, αν τα τελευταία στάδια της περιοδείας είναι συγκρίσιμα σε μήκος με το πρώτο στάδιο, τότε η περιήγηση είναι λογική. Αν είναι πολύ μεγαλύτερη, τότε είναι πιθανό ότι υπάρχουν πολύ καλύτερες εκδρομές. Ένας άλλος έλεγχος είναι να χρησιμοποιήσετε έναν αλγόριθμο, είναι να εφαρμόζεται το κατώτερο όριο του αλγόριθμου για να εκτιμηθεί αν σ’ αυτόν η περιοδεία του είναι αρκετά καλή. Στη χειρότερη περίπτωση, τα αποτελέσματα του αλγόριθμου σε μια περιοδεία είναι πολύ μεγαλύτερα από την βέλτιστη περιήγηση. Για την ακρίβεια, για κάθε σταθερά r είναι ένα παράδειγμα του επιβατικού προβλήματος πώλησης έτσι ώστε το μήκος του μήκους της περιοδείας να υπολογίζεται από τον πλησιέστερο γείτονα αλγόριθμο είναι ώστε να είναι μεγαλύτερος από το r τόσες φορές όσο και το μήκος του βέλτιστου περιηγητή.

3.7 Η Μέθοδος Naive Bayes

Η μέθοδος Bayes είναι μια πιθανολογική προσέγγιση της επαγωγικής μάθησης και ανήκει στην γενικότερη κατηγορία των Bayesian ταξινομητών. Οι Bayesian ταξινομητές χρησιμοποιούνται για να λύνουν προβλήματα ταξινόμησης. Στηρίζονται στην θεωρία των πιθανοτήτων και στο θεώρημα του Bayes. Η Bayesian στατιστική χρησιμοποιεί την πιθανότητα για να παρουσιάσει την αβεβαιότητα στις σχέσεις που αντλήθηκαν από τα δεδομένα. Επίσης η έννοια του «προγενέστερου» είναι πολύ σημαντική, καθώς αντιπροσωπεύει την εκ των προτέρων γνώση μας για το ποια μπορεί να είναι η πραγματική σχέση. Οι προκύπτουσες εκ των υστέρων πιθανότητες είναι ανάλογες με τις προκύπτουσες εκ των προτέρων πιθανότητες. Οι Bayesian ταξινομητές θεωρούν κάθε χαρακτηριστικό γνώρισμα και κατηγορία σαν μια τυχαία μεταβλητή.

Το μοντέλο λοιπόν εκτιμά τις προκύπτουσες εκ των υστέρων πιθανότητες P(c/d) του εγγράφου d που ανήκει στην κατηγορία c, στηριζόμενο στην εκ των προτέρων πιθανότητα P(c) παρατήρησης κάποιου έγγραφου στην κατηγορία c, στην πιθανότητα P(d/c) παρατήρησης του εγγράφου d δεδομένης της κατηγορίας c και στην πιθανότητα P(d) παρατήρησης του d. Υπολογίζεται λοιπόν η Δεσμευμένη Πιθανότητα με βάση το Θεώρημα Bayes ως εξής:

$$P(c|d) = \frac{P(c)P(d|c)}{P(d)} \quad (3.3)$$

Ο βασικός στόχος του ταξινομητή , είναι να προβλέψει σε ποια κατηγορία c ανήκει το έγγραφο d και να βρει την τιμή του c που θα μεγιστοποιήσει την πιθανότητα P(c/d):

$$c = argmax_{c_j} \frac{P(c_j)P(d|c_j)}{P(d)} \quad (3.4)$$

Για να υπολογίσει την δεσμευμένη πιθανότητα ο ταξινομητής Bayes, θεωρεί ότι όλα τα γνωρίσματα του εγγράφου d, δηλαδή οι λέξεις ή τα «σημεία» είναι ανεξάρτητα μεταξύ τους. Τα σημαντικότερα προτερήματα των ταξινομητών αυτών , είναι ότι είναι αρκετά ισχυροί ώστε να απομονώνουν τα περιττά χαρακτηριστικά ή τα σημεία θορύβου και ότι χειρίζονται τιμές που λείπουν , αγνοώντας το στιγμιότυπο κατά τον υπολογισμό των πιθανοτήτων.

Ο αλγόριθμος Bayes είναι ένας από τους πιο απλούς αλγόριθμους ταξινόμησης. Ονομάζεται Naïve Bayes, δηλαδή αφελής επειδή γίνεται η υπόθεση ότι η δεσμευμένη πιθανότητα κάποιας λέξης που εμφανίζεται σε κάποια κατηγορία είναι ανεξάρτητη από τις δεσμευμένες πιθανότητες άλλων λέξεων δοθέντος αυτής της κατηγορίας. Ακόμα και απλός, έχει πολύ δυνατά σημεία. Κάποια από αυτά είναι:

- Είναι εύκολο να κατασκευαστεί
- Είναι γρήγορος
- Εύκολα κατανοητός
- Αν και απλός, είναι πολύ αποτελεσματικός

Ο αλγόριθμος δουλεύει πολύ καλά στην πράξη, ακόμα και αν κάνει κάποιες υποθέσεις που στην πράξη δεν ισχύουν(π.χ. ότι τα χαρακτηριστικά είναι ανεξάρτητα μεταξύ τους). Ένα μειονέκτημα του αλγορίθμου αυτού είναι ότι δεν μπορεί να χρησιμοποιηθεί σε πιο πολύπλοκες περιπτώσεις ταξινόμησης. Αποτελέσματα δείχνουν ότι το μοντέλο Naïve Bayes έχει καλύτερες επιδόσεις σε σχέση με άλλες τεχνικές, και ειδικά σε εφαρμογές που έχουν να κάνουν με μεγάλα λεξιλόγια. Ωστόσο, παρουσιάζει αδυναμίες όταν υπάρχουν διαφορετικά μεγέθη στα έγγραφα του training set έχουν και όταν οι κατηγορίες είναι λίγες γιατί δεν υπάρχουν αρκετά training δεδομένα. Τέλος , ο ταξινομητής Naïve Bayes εφαρμόζεται σε πολλά συστήματα προτάσεων με βάση το περιεχόμενο .

3.8 Γραμμικοί Ταξινομητές

Οι Γραμμικοί ταξινομητές είναι αλγόριθμοι που χρησιμοποιούν τα γραμμικά όρια αποφάσεων, όπως για παράδειγμα να διαχωρίσουν τις περιπτώσεις σε έναν πολυδιάστατο χώρο, και εφαρμόζονται ευρέως σε κατηγοριοποιήσεις κειμένων. Όλοι οι γραμμικοί ταξινομητές λειτουργούν στα πλαίσια μιας κοινής φιλοσοφίας. Η διαδικασία μάθησης του αλγορίθμου μοντελοποιείται με ένα v-διάστατο διάνυσμα βαρών w, του οποίου το εσωτερικό γινόμενο με ένα στιγμιότυπο, όπως για παράδειγμα ένα έγγραφο κειμένου που εκπροσωπείται από το μοντέλο του Διανυσματικού Χώρου , δίνει ως αποτέλεσμα μια αριθμητική πρόβλεψη. Η αριθμητική αυτή πρόβλεψη οδηγεί σε μια προσέγγιση γραμμικής παλινδρόμησης. Κάποιες φορές ωστόσο μπορεί να χρησιμοποιηθεί ένα όριο ώστε οι συνεχείς προβλέψεις να μετατραπούν σε διακριτές κατηγορίες.

Αυτό το γενικό πλαίσιο λειτουργίας ισχύει για όλους τους γραμμικούς ταξινομητές. Οι διαφοροποιήσεις εμφανίζονται στις μεθόδους εκπαίδευσης των αλγορίθμων, που χρησιμοποιούνται για να υπολογίζουν το διάνυσμα βαρών w. Ένα σημαντικό πλεονέκτημα των παραπάνω μεθόδων εκπαίδευσης των γραμμικών αλγορίθμων είναι ότι μπορούν να εφαρμοστούν διαδικτυακά on-line. Το πλεονέκτημα αυτό είναι υψίστης σημασίας για εφαρμογές που λειτουργούν σε πραγματικό χρόνο, καθώς για παράδειγμα το εκάστοτε διάνυσμα βαρών μπορεί να μετατρέπεται σταδιακά, ενώ όλο και περισσότερα στιγμιότυπα γίνονται διαθέσιμα και προστίθενται στην εφαρμογή. Μια άλλη σημαντική παρατήρηση σχετικά με τους γραμμικούς ταξινομητές είναι ότι ενώ οι μέθοδοι που χρησιμοποιούν τείνουν να συγκλίνουν σε γραμμικά όρια που χωρίζουν σε κατηγορίες τα δεδομένα εκμάθησης με ακρίβεια, η γενική απόδοση αυτών των ορίων δεν είναι βέλτιστη.

ΚΕΦΑΛΑΙΟ 4ο

<< Περιγραφή των πειραμάτων, των αντίστοιχων συνόλων δεδομένων >>

4.1 Εισαγωγή

Η Ομοιότητα αναζήτησης είναι ένα σημαντικό έργο με πολλές εφαρμογές, όπως για παράδειγμα την ανάκτησης βάση περιεχομένου, τη διερευνητική ανάλυση των δεδομένων, τα μοντέλα προβλέψεων και την εξόρυξη δεδομένων. Το βασικό πρόβλημα μπορεί να διατυπωθεί ως εξής: δίνεται ένα σύνολο αντικειμένων, βρείτε τα πιο παρόμοια συστήματα με βάση ένα συγκεκριμένο ερώτημα αντικειμένου. Για παράδειγμα, κάποιος μπορεί να ενδιαφέρεται για την ανάκτηση των πιο παρόμοιων εικόνων από μια βάση δεδομένων ή τον προσδιορισμό των εν λόγω αποθεμάτων των οποίων οι τιμές εξελίχθηκαν παρόμοια με ένα ειδικό τον τελευταίο χρόνο. Η Ανάκτηση δεδομένων από αυτά τα αντικείμενα στηρίζεται στην “ομοιότητα” και όχι σε “ακρίβεια”. Η κύρια έρευνα στον τομέα αυτό επικεντρώνεται στην ανάπτυξη των μεθόδων που μπορούν να υποστηρίξουν αποτελεσματικά την αναζήτηση ομοιότητας, αφού κοινές εφαρμογές περιλαμβάνουν ένα πολύ μεγάλο όγκο δεδομένων. Ο όγκος των δεδομένων δεν εξαρτάται μόνο από τον αριθμό των αντικειμένων, αλλά και από τις διαστάσεις τους. Συνήθως, ένα αντικείμενο μπορεί να θεωρηθεί ως ένα σημείο σε ένα n-διάστατο Ευκλείδειο χώρο. Μια σημαντική κατηγορία των μεθόδων για την αποδοτική αναζήτηση ομοιότητας αποτελείται από πολυδιάστατα συστήματα ευρετηρίασης που μπορεί να χρησιμοποιηθούν για γρήγορη πρόσβαση σε αυτά τα σημεία.

Σε γενικές γραμμές, συστήματα όπως τα δέντρα διαιρούν το Ν-διάστατο χώρο σε επικαλυπτόμενες υποπεριφέρειες που περιέχουν τα υποσύνολα των αντικειμένων. Όταν ένα ερώτημα αντικειμένου φτάνει, δεν οδηγείται σε μια από αυτές τις επιμέρους περιοχές, και στη συνέχεια, γίνεται μια αναζήτηση για τον πλησιέστερο γείτονα, προκειμένου να εντοπιστούν τα παρόμοια αντικείμενα που μπορούν να διαμένουν όχι μόνο σε αυτή την υποπεριφέρεια, αλλά και σε όμορες περιοχές. Αν και η προσέγγιση ευρετηρίασης μπορεί να είναι εξαιρετικά γρήγορη, η αποτελεσματικότητα του υποβαθμίζεται ραγδαία με την αύξηση των διαστάσεων. Υπάρχουν διάφορα αποτελέσματα της έρευνας που καταδεικνύουν ότι οι αρνητικές επιπτώσεις της αύξησης διαστάσεων στις δομές δείκτη αναφέρουν ότι, όπως η διάσταση αυξάνει από 5 έως 10, η εκτέλεση μιας αναζήτησης για τον πλησιέστερο γείτονα υποβαθμίζει κατά ένα συντελεστή 12 για διάφορα πολυδιάστατες δομές δείκτη. Η κοντινότερη αναζήτηση γείτονα μπορεί να γίνει ασταθές και με μόνο 10-20 διαστάσεις. Το φαινόμενο αυτό, γνωστό και ως κατάρα διαστάσεων, σημαίνει ότι μια απλή διαδοχική σάρωση εκτελεί συνήθως καλύτερα σε υψηλότερες διαστάσεις από δομές δείκτη. Μια λύση για την επίτευξη αποτελεσματικών αναζητήσεων ομοιότητας με την παρουσία υψηλών διαστάσεων είναι να συμπυκνώσει τα δεδομένα, εφαρμόζοντας μια τεχνική μείωσης διαστάσεων (δηλαδή Ανάλυση κύριο συστατικό). Η ιδέα είναι να χαρτογραφήσει τα αρχικά δεδομένα σε ένα χαμηλότερο τομέα διάστασης, χωρίς να χάσει μεγάλο μέρος των πληροφοριών. Η προσέγγιση της μείωσης των διαστάσεων μπορεί να είναι πολύ χρήσιμη με διάφορους τρόπους. Το γεγονός ότι μειώνει τις απαιτήσεις αποθήκευσης επηρεάζει άμεσα την κλιμακωτή απόδοση. Επίσης, αυτή η προσέγγιση μπορεί να οδηγήσει σε μια σειρά από διαστάσεις που επιτρέπουν την αποτελεσματική εφαρμογή των πολυδιάστατων δομών ευρετηρίασης. Επιπλέον, υπάρχουν πολλές εφαρμογές όπου η μείωση διαστάσεων βελτιώνει τις επιδόσεις της αναζήτησης ομοιότητας όσον αφορά την ποιότητα των αποτελεσμάτων.

Για παράδειγμα, μια τεχνική μείωσης των διαστάσεων , είναι ικανή να εξαλείψει τα υψηλά επίπεδα θορύβου των παρόντων δεδομένων και μπορεί να βελτιώσει την ποιότητα της αναζήτησης ομοιότητας. Σε αυτό το κεφάλαιο, εξετάζουμε την περίπτωση της αναζήτησης ομοιότητας στατιστικών δεδομένων στο πλαίσιο ενός κοντινότερο γείτονα (1NN) ταξινόμησης. Πίνακες με στατιστικά δεδομένα διαφέρουν από άλλους τομείς, δεδομένου ότι αποτελούνται από πολλές διαστάσεις, γεγονός που καθιστά απαραίτητη την εφαρμογή της τεχνικής μείωσης των διαστάσεων. Επιπλέον, τα αποτελέσματα από το κεφάλαιο 4 δείχνουν ότι η απόδοση της αναζήτησης ομοιότητα βελτιώνεται όταν τα αρχικά δεδομένα αντιστοιχίζονται σε χαμηλότερο τομέα διάστασης. Η μέθοδος της 1NN ταξινόμησης απαιτεί την αναζήτηση σε μια βάση δεδομένων για το πιο παρόμοιο αντικείμενο σε ένα ανά ένα δεδομένο. Το κύριο μειονέκτημα της μεθόδου αυτής είναι ότι έχουμε να συγκρίνουμε ένα ερώτημα αντικειμένου με κάθε ένα αντικείμενο σε μια βάση δεδομένων για να βρεθεί έτσι το πιο παρόμοιο αντικείμενο. Η προσέγγιση αυτή καθίσταται απαγορευτική, όταν η βάση δεδομένων αναφοράς είναι εξαιρετικά μεγάλη. Η αποτελεσματικότητα αυτής της μεθόδου επηρεάζεται από τον αριθμό των αντικειμένων που υπάρχουν στη βάση δεδομένων, καθώς και, από το χαρακτήρα της , αυτά τα αντικείμενα έχουν ένα μέτρο απόστασης για να υπολογίζουν την μέτρηση της εγγύτητας των αντίστοιχων αντικειμένων.

4.2 Η Μέθοδος Leave One Out

Μια πολύ απλή , και μια ευρέως χρησιμοποιούμενη μέθοδος είναι η μέθοδος "leave-one-out" («αφήστε-μια-έξω).Η ιδέα πίσω από αυτή τη μέθοδο κρύβεται, όπως υποδηλώνει και το όνομά της , στο να περιλαμβάνει τη χρησιμοποίηση ενός και μόνο παρατηρητή από το αρχικό δείγμα και οι υπόλοιπες παρατηρητές να χρησιμοποιούνται ως data train. Αυτό επαναλαμβάνεται για κάθε έναν παρατηρητή στο δείγμα και χρησιμοποιείται μία φορά και μόνο την επικύρωση των δεδομένων. Τα αποτελέσματα αυτής της μεθόδου είναι συνήθως πολύ ακριβά από μια υπολογιστική άποψη, λόγω του μεγάλου αριθμού των φορών που η εκπαιδευτική διαδικασία επαναλαμβάνεται.

4.3 Η Προτεινόμενη Προσέγγιση

Η προτεινόμενη μέθοδος αποτελείται από τρεις φάσεις οι οποίες περιγράφονται στη συνέχεια αναλυτικά. Η πρώτη φάση περιλαμβάνει την εφαρμογή μιας τεχνικής μείωσης της διαστατικότητας των αρχικών δεδομένων. Αυτό γίνεται για δύο λόγους. Πρώτον, περιμένουμε να βελτιωθεί η ποιότητα στα αποτελέσματα της αναζήτησης ομοιότητας. Δεύτερον, η επακόλουθη συσταδοποίηση της ανάλυση γίνεται πιο αποτελεσματική, δεδομένου ότι το πρόβλημα των υψηλών τρισδιάστατων δεδομένων μετριάζεται. Ουσιαστικά, κάθε τεχνική μπορεί να επιλεγεί σε αυτό το βήμα, ωστόσο, η επιλογή αυτής της αίτησης εξαρτάται, από διαφορετικές τεχνικές που μπορούν να οδηγήσουν σε αναπαραστάσεις της υψηλότερης ποιότητας σε διαφορετικές εφαρμογές. Σε αυτό το έργο προτείνουμε να εφαρμοστεί η Κατά-τμηματική Συνολική Προσέγγιση (PAA), γιατί είναι απλή και γρήγορη για τον υπολογισμό και έχει αποδειχθεί εμπειρικά ότι είναι τόσο αποτελεσματική όσο άλλες πιο σύνθετες προσεγγίσεις.

Η δεύτερη φάση περιλαμβάνει την εφαρμογή ενός αλγόριθμου συγκέντρωσης, που μετατραπεί τα δεδομένα (δηλαδή k means) και είναι ένας από τους πιο μελετημένους και δημοφιλέστερους τρόπους συσταδοποίησης. Το αποτέλεσμα αυτής της φάσης αποτελείται από το κέντρο βάρους (C i) της παραγόμενης συστάδας κατά μήκος με τις ακτίνες τους(R i). Η ακτίνα ενός συμπλέγματος ή μιας συστάδας ορίζεται ως η απόσταση από το απώτατο αντικείμενο ενός συμπλέγματος στο αντίστοιχο κέντρο βάρους. Εκτός από αυτά, καταγράφουμε την συμμετοχή του κάθε συμπλέγματος του αντικειμένου και την απόστασή του από το αντίστοιχο κέντρο βάρους .Τα αντικείμενα που αναδιατάσσονται στο σύνολο δεδομένων, όσον αφορά στην ένταξη της διασποράς και της απόστασης τους από το κέντρο βάρους της διασποράς τους. Σημειώστε ότι οι δύο προηγούμενες φάσεις αποτελούν την προεπεξεργασία που εκτελείται off-line. Η τρίτη φάση περιλαμβάνει τη διαδικασία της ομοιότητας της αναζήτησης στα παράγωγα συμπλέγματα. Λαμβάνοντας υπόψη ένα ερώτημα αντικειμένου (q),τα απαιτούμενα βήματα που πρέπει να ακολουθούνται παρέχονται παρακάτω:

1. Υπολογισμός των αποστάσεων της Κατά-τμηματικής Συνολικής Προσέγγισης που μετέτρεψαν το ερώτημα του αντικειμένου στο κέντρο βάρους των συμπλεγμάτων ($d(q,c(i))$).

2. Ρύθμιση του συμπλέγματος με το πλησιέστερο κέντρο βάρους, ως την τρέχον συστάδα. Ας δηλώσουμε αυτήν την συστάδα C^ (i), όπου i = 1.

3. Υπολογισμός της απόστασης του ερωτήματος αντικειμένου με κάθε μία από τις άλλες ομάδες. Η απόσταση αυτή ορίζεται να είναι η διαφορά της μεταξύ τους απόστασης από το ερώτημα του αντικείμενου για το κέντρο βάρους και την αντίστοιχη ακτίνα.

Εάν το ερώτημα αντικειμένου βρίσκεται μέσα στην συστάδα , τότε η απόσταση αυτή είναι ίση με μηδέν (1).

$$d\,(q,C^{(i)})=\max\{0\,,\,d\,(q,c(i))-r(i)\}\,,\ i=2,3,4,\ldots,k \quad (4.1)$$

4. Οι ομάδες ταξινομούνται σε αύξουσα σειρά ως προς τις αποστάσεις τους από το ερώτημα του αντικειμένου. Οι δεσμοί μπορεί να σπάσουν, σύμφωνα με τις αποστάσεις των κέντρων βάρους των συσπειρώσεων του ερωτήματος του αντικειμένου. Ας δηλώσουμε αυτές τις ομάδες (i) C, όπου i = 2,3 ,..., k με το k να αντιστοιχεί με την συστάδα που είναι η πλέον απομακρυσμένη από το ερώτημα.

5. Αναζητήστε την τρέχουσα συστάδα διαδοχικά, προκειμένου να εντοπίσετε τον τρέχον πλησιέστερο γείτονα (NN) στο ερώτημα του αντικείμενου. Καταγράψτε την αντίστοιχη απόσταση d (Q, nn).

6. Αν η απόσταση του ερωτήματος του αντικειμένου για τον τρέχον πλησιέστερο γείτονα είναι μικρότερη ή ίση με την απόσταση του ερωτήματος αντικείμενου στην επόμενη συστάδα (2), τότε ο πραγματικός πλησιέστερος γείτονας έχει βρεθεί και ο αλγόριθμος σταματά.

$$d\,(q,nn)\leq d\,(q,C^{(I+1)}) \quad (4.2)$$

Διαφορετικά, ας αυξήσουμε το i,(i = i +1) και προχωράμε στο επόμενο βήμα.

7. Υπολογίστε τη διαφορά μεταξύ της απόστασης των ερωτημάτων αντικείμενου με το κέντρο βάρους του (i) C και την απόσταση του ερωτήματος αντικείμενου στον τρέχον πλησιέστερο γείτονα. Εάν αυτή η διαφορά είναι θετική, τότε η αναζήτηση αντικειμένων του (i) ότι η C απόστασή τους να ic είναι μεγαλύτερη από τη διαφορά αυτή. Σε αντίθετη περίπτωση, να αναζητήσει όλη την συστάδα. Μετά από αυτό, ο τρέχον πλησιέστερος γείτονας (NN) έχει αλλάξει. Αν i = k, ο αλγόριθμος σταματά. Σε αυτό το βήμα, έχουμε καθορίσει ένα κάτω φράγμα για τις αποστάσεις των αντικειμένων από το κέντρο βάρους, προκειμένου να μειωθεί ο χώρος αναζήτησης της τρέχουσας συστάδας (3).

$$\textit{χαμηλότερο\_κάτω φράγμα} = \max\{0 , d(q,c) - (q,nn)\}, \quad (4.3)$$

8.Μεταβείτε στο βήμα 6.

Για την αξιολόγηση της απόδοσης της προτεινόμενης προσέγγισης, μπορούμε να εκτελέσουμε την ταξινόμηση one-nearest neighbor (ένας κοντινότερος γείτονας)(1-NN) και να επικυρώσει αυτά που αφήνει η one-out διαδικασία. Εμείς καταγράφουμε την ακρίβεια ταξινόμησης και το ποσοστό του όγκου των δεδομένων αναζήτησης. Ο αντίστοιχος μέσος όρος τιμών υπολογίζεται πάνω από το συνολικό αριθμό των σειρών στο σύνολο δεδομένων. Έχουν χρησιμοποιηθεί ευρέως ως σημείο αναφοράς σύνολα δεδομένων για τις δοκιμές αλγορίθμων ταξινόμησης και για το λόγο αυτό, είναι χωρισμένα σε εκπαιδευτικά και δοκιμαστικά σύνολα. Σε αυτό το έργο, εμείς συγχωνεύουμε αυτές τις δύο ομάδες προκειμένου να αυξηθεί το μέγεθος του κάθε σύνολο δεδομένων.

Όσον αφορά την Κατά-τμηματική Συνολική Προσέγγιση (PAA), ο αριθμός των τμημάτων είναι ίσος με πολλαπλάσια του 2, που κυμαίνονται από 4 έως 16. Υποθέτουμε ότι η αναλογία του μήκους της σειράς με τον αριθμό των τμημάτων είναι ακέραιος αριθμός. Σε αντίθετη περίπτωση, ο αναγκαίος αριθμός των μηδενικών μπορεί να προστεθεί στο τέλος της αρχικής σειράς πριν εφαρμοστεί η Κατά-τμηματική Συνολική Προσέγγιση (PAA). Η τροποποίηση αυτή δεν επηρεάζει τα αποτελέσματα της αναζήτησης. Ο αλγόριθμος K-means εφαρμόζεται με αριθμούς από συστάδες που κυμαίνονται από 2 έως 16 και είναι πολλαπλάσια του 2. Διαισθητικά, όσο πιο πολύ είναι οι αριθμοί των συστάδων, τόσο μικρότερο είναι το κλάσμα του συνόλου δεδομένων που πρέπει να αναζητηθούν.

Τέλος, όλοι οι απαραίτητοι κώδικες και τα πειράματα αναπτύχθηκαν σε MATLAB και υπάρχουν διαθέσιμοι στα Παραρτήματα Α ,Β,Γ , στο τέλος αυτής της εργασίας

4.4 Πληροφορίες από Αρχεία Δεδομένων

- **Land sat Satellite**

Αυτή η βάση δεδομένων αποτελείται από πολυ-φασματικές τιμές των pixels σε 3x3 «γειτονιές» σε μια δορυφορική εικόνα, και κατά την ταξινόμηση κάθε μια από αυτές συνδέεται με το κεντρικό pixel σε κάθε «γειτονιά». Στόχος είναι να προβλέψει αυτή την κατάταξη, δεδομένης της πολυ-φασματικές τιμής. Στο δείγμα βάσης δεδομένων, η κλάση του ενός pixel κωδικοποιείται ως αριθμός. Ο δορυφόρος δεδομένων Landsat είναι μία από τις πολλές πηγές πληροφοριών που υπάρχουν για μια σκηνή. Η ερμηνεία μιας σκηνής για την ενσωμάτωση χωρικών δεδομένων των διαφορετικών τύπων και των ψηφισμάτων συμπεριλαμβάνει τις πολυφασματικών τιμές , τα δεδομένα από ραντάρ , τους χάρτες που δείχνουν την τοπογραφία, το έδαφος κλπ. Η χρήση τους αναμένεται να αναλάβει μεγαλύτερη σημασία με την έναρξη μιας εποχής που χαρακτηρίζεται από ενοποιητικές προσεγγίσεις για την τηλεπισκόπηση (για παράδειγμα, η Γη της NASA ένα σύστημα παρακολούθησης που αρχίζει αυτή τη δεκαετία). Οι υπάρχουσες στατιστικές μέθοδοι απαιτούν εφόδια για τη διεκπεραίωση αυτών των διαφορετικών τύπων δεδομένων. Σημειώστε ότι αυτό δεν ισχύει για τα δεδομένα Landsat MSS τα οποία και εξετάζονται μεμονωμένα (όπως σε αυτό το δείγμα βάσης δεδομένων). Συνεπώς, για αυτά τα δεδομένα, είναι ενδιαφέρον να τα συγκρίνουμε τις επιδόσεις των άλλων μεθόδων κατά την στατιστική προσέγγιση. Ένα πλαίσιο εικόνων Landsat MSS αποτελείται από τέσσερις ψηφιακές εικόνες της ίδιας σκηνής σε διάφορες φασματικές ζώνες. Δύο από αυτές αντιστοιχούν στην ορατή περιοχή (που αντιστοιχεί περίπου σε πράσινες και κόκκινες περιοχές του ορατού φάσματος) και οι υπόλοιπες δύο αντιστοιχούν στο υπέρυθρο. Κάθε pixel αποτελείται από μια 8-bit δυαδική λέξη , από 0 που αντιστοιχεί στο μαύρο έως και το 255 που αντιστοιχεί στο λευκό. Η χωρική ανάλυση ενός pixel είναι περίπου 80m x 80m. Κάθε εικόνα περιέχει 2340 x 3380 pixels τέτοια.

- **Shuttle**

Περίπου το 80% των δεδομένων ανήκει στην κατηγορία 1. Επομένως, η ακρίβεια είναι περίπου 80%. Ο στόχος εδώ είναι να επιτευχθεί μια ακρίβεια της τάξης του 99 - 99.9%. Τα παραδείγματα στο αρχικό σύνολο δεδομένων ήταν για τον καιρό, και αυτό για του χρόνου θα μπορούσε πιθανώς να είναι χρήσιμο για την ταξινόμηση. Ωστόσο, αυτά δεν θεωρούνται σημαντικά για τον σκοπό της StatLog, έτσι ώστε με τη σειρά τους τα παραδείγματα στο αρχικό σύνολο δεδομένων ήταν τυχαιοποιημένα, και ένα μέρος του αρχικού συνόλου δεδομένων αφαιρείται.

.

ΚΕΦΑΛΑΙΟ 5ο

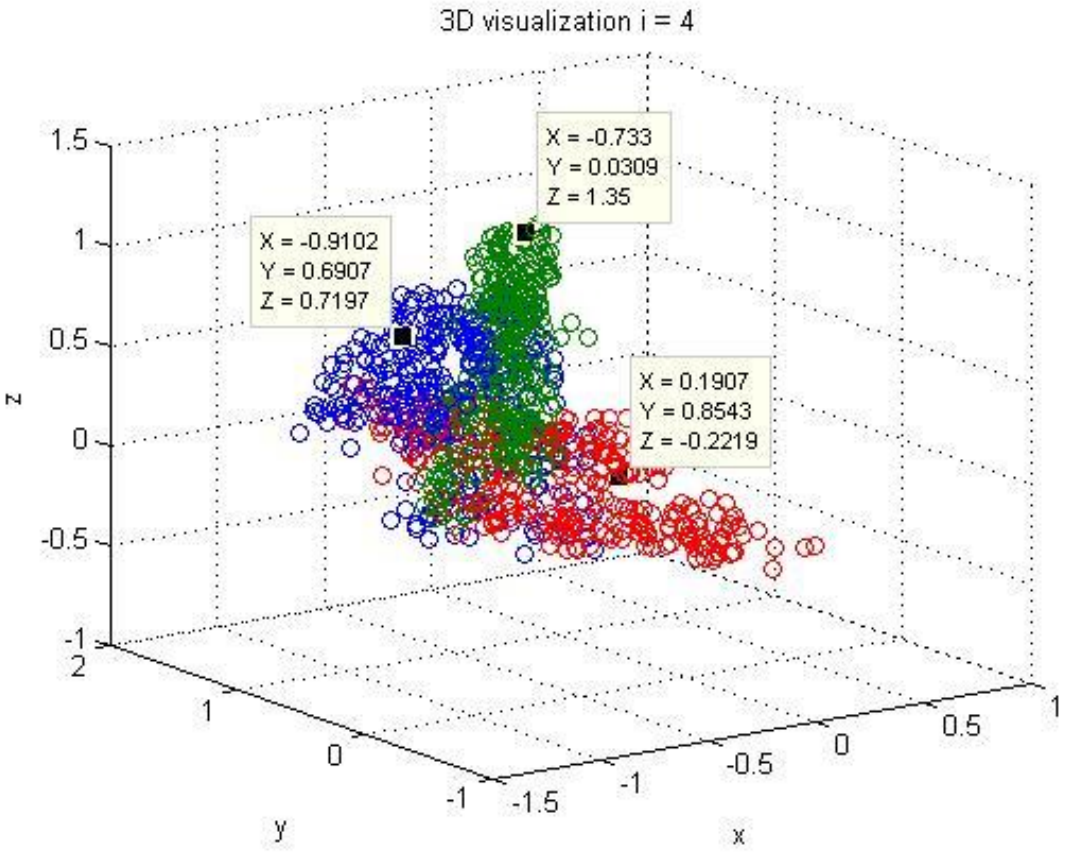
5.1 Αποτελέσματα

Στο πρώτο μέρος από τα αποτελέσματα, παρέχουμε τα ποσοστά σφάλματος ταξινόμησης και το ποσοστό του όγκου των δεδομένων που έψαχναν για ένα σταθερό αριθμό από συστάδες (K = 10). Ο στόχος είναι να εξεταστεί η σχέση τους κάτω από μία ποικιλία διαστάσεων. Στα παρακάτω κεφάλαια [5.2] , [5.3] παρουσιάζονται τα ποσοστά σφάλματος ταξινόμησης για διαφορετικές διαστάσεις όταν ο αριθμός των συστάδων που δημιουργούνται ορίζεται ίσος με το 10. Η βασική παρατήρηση είναι ότι το χαμηλότερο ποσοστό σφάλματος επιτυγχάνεται σε υψηλές διαστάσεις. Επίσης , κάτω από κάθε εικόνα εκτός από τα ποσοστά σφάλματος ταξινόμησης για διαφορετικές διαστάσεις παρατίθενται και τα ποσοστά του όγκου των δεδομένων όταν αναζητείται ο αριθμός των συστάδων που δημιουργούνται όταν αυτός ορίζεται ίσος με το 10. Υπάρχει μια αυξανόμενη τάση του όγκου των δεδομένων που ερευνήθηκαν, όπως αυξάνεται και η διαστατικότητα. Η βασική παρατήρηση που προκύπτει από τον συνδυασμό είναι ότι καθώς αυξάνεται η διαστατικότητα ,το ποσοστό σφάλματος ταξινόμησης μειώνεται με το κόστος του αυξανόμενου ποσοστού του όγκου των δεδομένων που πρέπει να αναζητηθούν. Παρόμοιες παρατηρήσεις μπορούν να γίνουν, αν δημιουργούν λιγότερες ή περισσότερες από 10 συστάδες.

Για συντομία των αποτελεσμάτων, δεν παρουσιάζονται άλλες τιμές για το k. Στο δεύτερο μέρος των αποτελεσμάτων, παρέχουμε το ποσοστό του όγκου των δεδομένων που έψαξαν για διαφορετικό αριθμό από συστάδες. Η διαστατικότητα έχει οριστεί ίση με εκείνον τον αριθμό για τον οποίο επιτυγχάνεται το χαμηλότερο ποσοστό σφάλματος .Σε γενικές γραμμές, αυτός ο αριθμός ποικίλλει μεταξύ των διαφόρων συνόλων δεδομένων. Σε αυτό το σύνολο των πειραμάτων, παρέχουμε αποτελέσματα για την CLUREP. Σημείωση ότι η τελευταία διαφέρει από την πρώτη εφόσον η αναζήτηση της γίνει σε ολόκληρη την συστάδα που είχε επισκεφθεί κάποτε. Εξ ορισμού, η μέθοδος CLUREP αναμένεται να δώσει καλύτερα αποτελέσματα. Ωστόσο, ένας από τους στόχους μας είναι να πειραματιστούμε ποσοτικά την αναμενόμενη βελτίωση. Η απόδοση της CLUREP παρουσιάζεται σε σχέση με το ποσοστό του όγκου των δεδομένων που είναι προς αναζήτηση. Η πρώτη παρατήρηση είναι ότι όταν ο αριθμός των παραγόμενων συσπειρώσεων αυξάνεται, αντίστοιχα μειώνεται η ένταση. Η δεύτερη παρατήρηση είναι ότι και η μέθοδος έχει καλύτερες επιδόσεις από την διαδοχική σάρωση. Η CLUREP εκτελείται με συνέπεια μεταξύ του συνόλου των δεδομένων για οποιοδήποτε αριθμό από συστάδες.

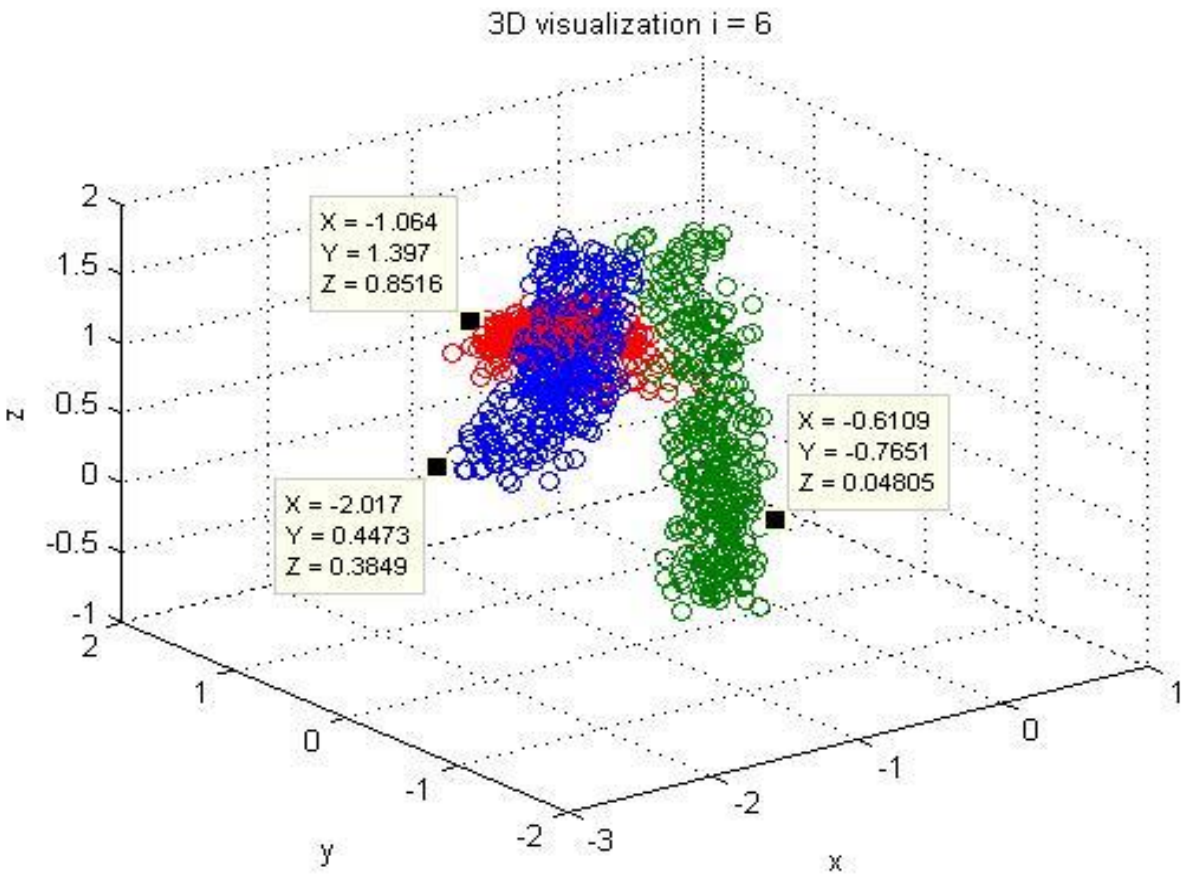
5.2 Οπτικοποίηση των αποτελεσμάτων

Το παρακάτω έχουμε μια οπτική εικόνα των αποτελεσμάτων . Η οπτική αναπαράσταση γίνεται σε τρεις διαστάσεις.



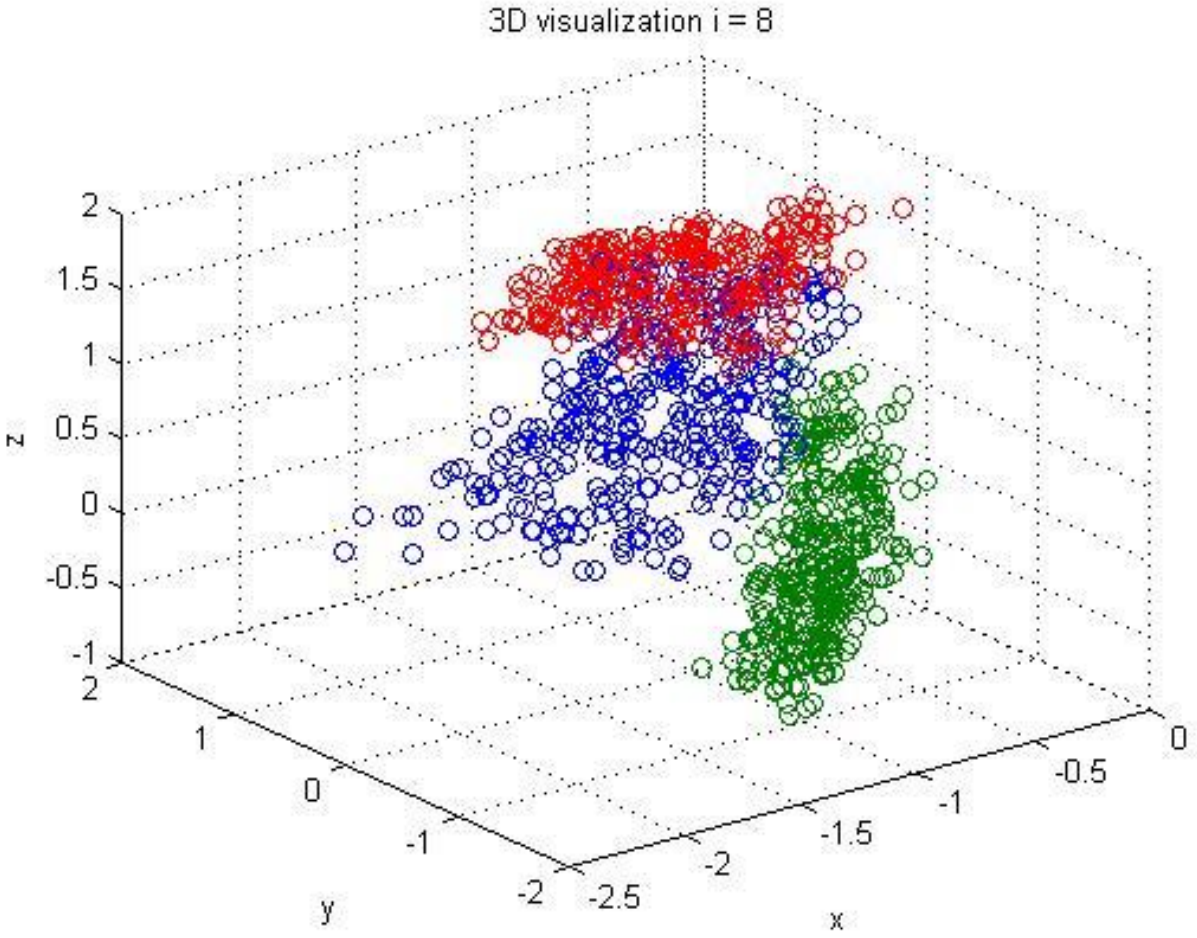
Error =10.5376, average space=12.1134

Εικόνα 15



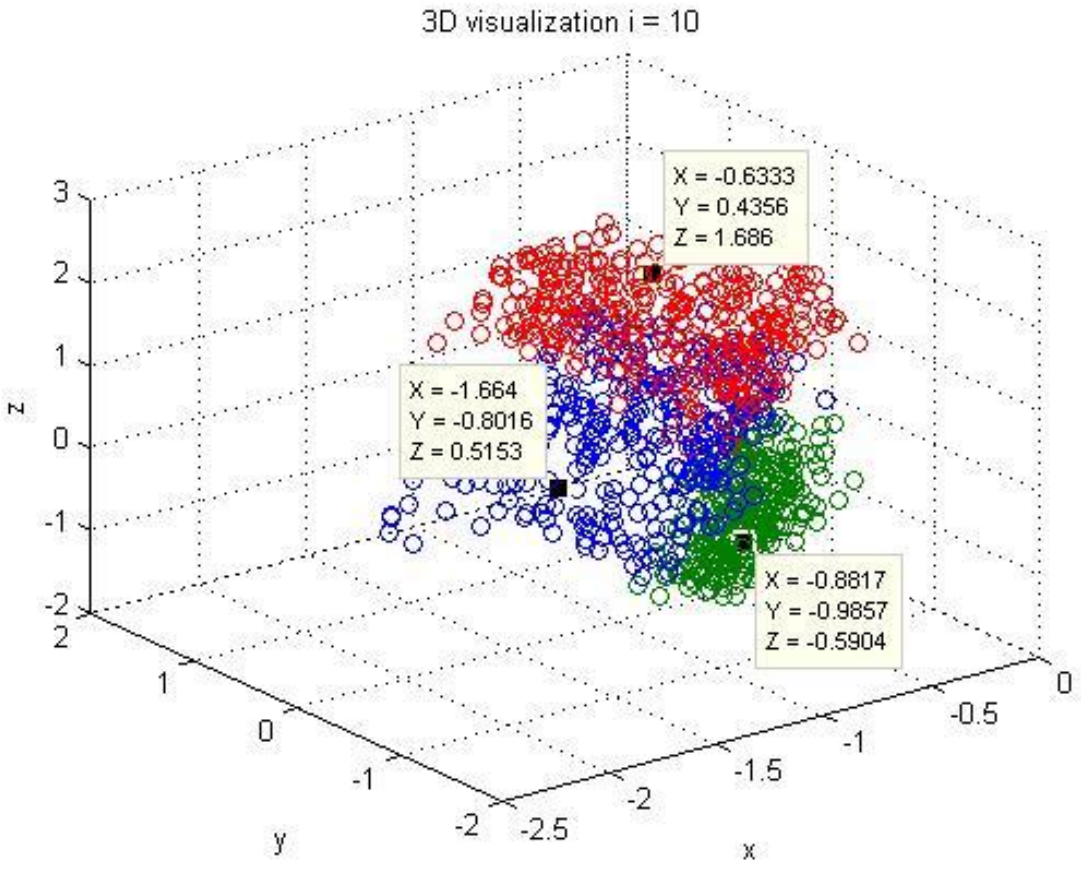
Error= 0.4301, average space=13.3050

Εικόνα 16



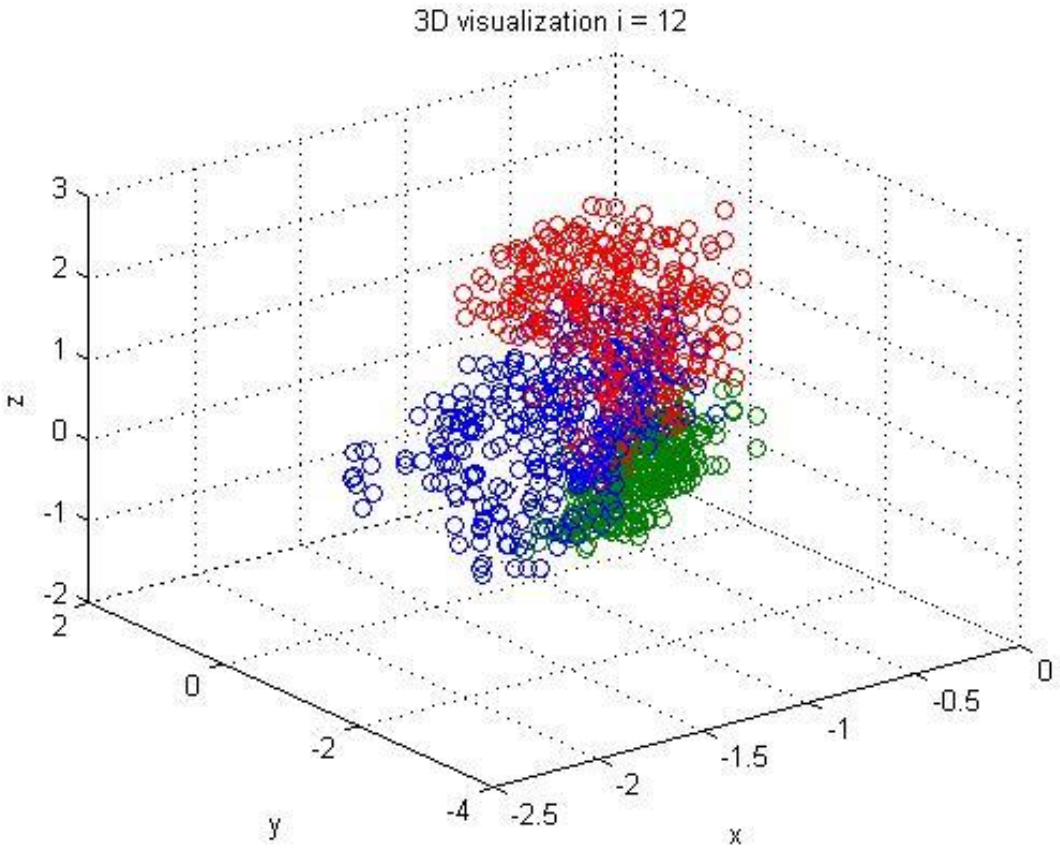
Error=0, average space=13.8788

Εικόνα 17



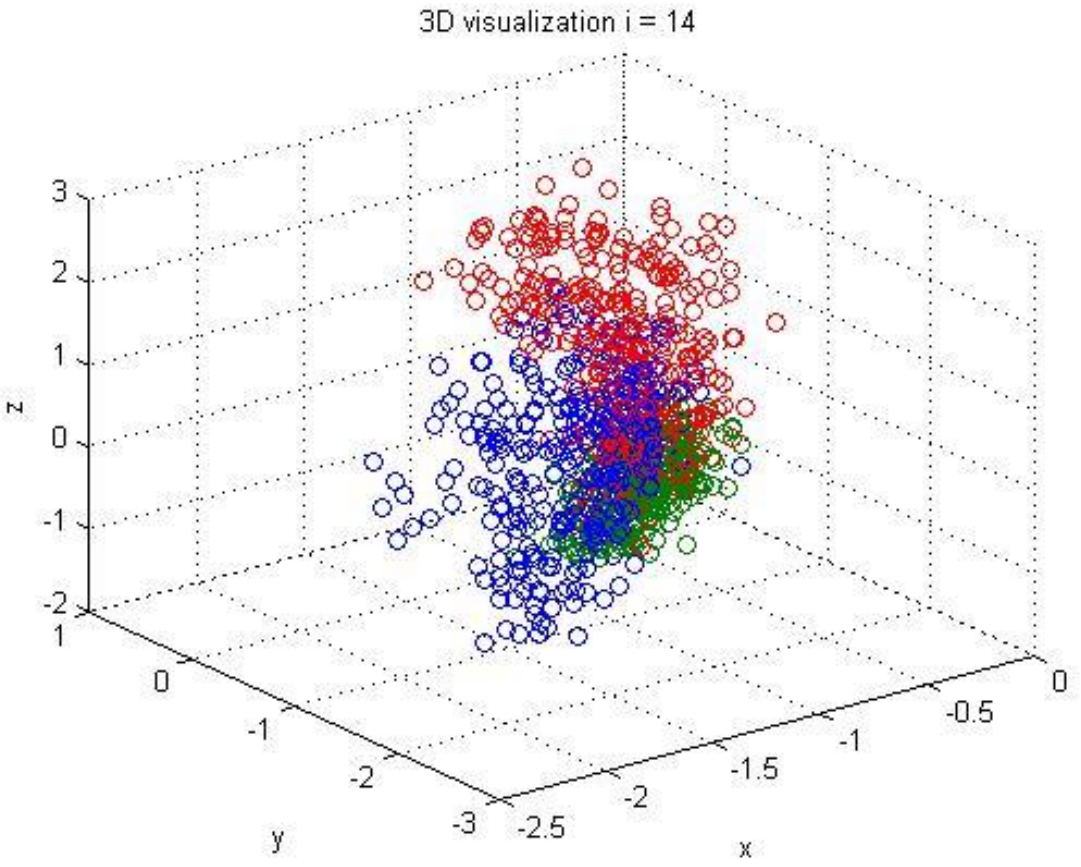
Error= 0.1075, average space=14.7267

Εικόνα 18



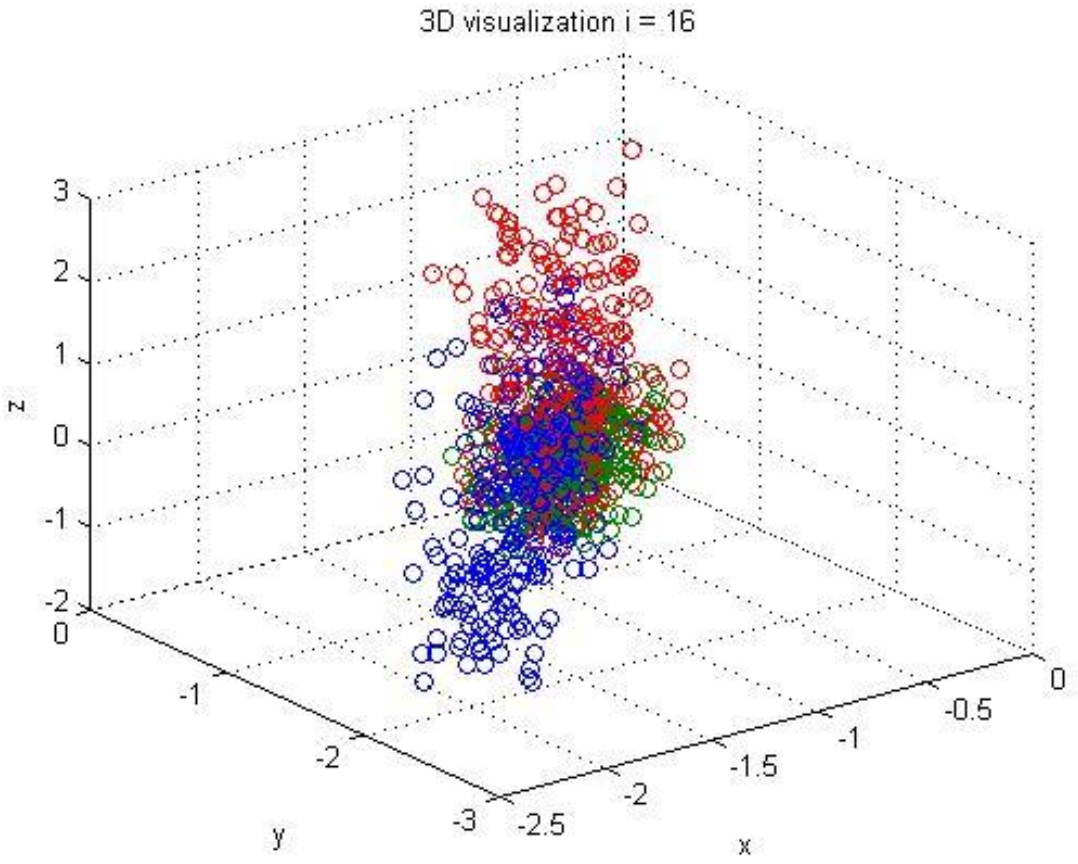
Error= 0, average space=16.2031

Εικόνα 19



Error=0, average space=15.5621

Εικόνα 20



Error= 0, average space=17.1762

Εικόνα 21

5.3 Πίνακες αποτελεσμάτων

Πίνακας 5.1 1-NN classification error rates (%) ($k = 10$)

		Dimensionality						
ID	Dataset	4	6	8	10	12	14	16
1	LandSat-Satellite	89.38	85.62	84.20	86.81	82.06	82.06	82.06
2	CBF	10.54	0.43	0.00	0.11	0.00	0.00	0.00

Στον παραπάνω πίνακα παρατηρούμε ότι το CBF dataset δίνει σαφώς καλύτερα αποτελέσματα από ότι το LandSat-Satellite dataset. Το LandSat-Satellite dataset δεν δίνει μεγάλη ακρίβεια καθώς ο αριθμός που παρουσιάζουν τα errors είναι μεγάλος. Στο παραπάνω πείραμα κρατήσαμε σταθερό των αριθμό των συστάδων (k = 10) και αυξάνουμε σταδιακά ανά δυο την διαστατικότητα. Παρατηρούμε , στο dataset CBF, ότι στις τιμές 8,12,14,16 επιτυγχάνουμε άριστη απόδοση καθώς τα σφάλματα είναι μηδενικά. Επίσης βλέπουμε ότι μειώνεται σταδιακά άλλα με αργώ ρυθμό και ο αριθμός των σφαλμάτων στο άλλο dataset. Κάτι ακόμη που αξίζει να σημειωθεί, είναι ότι ενώ υπάρχει μια μειώσει του σφάλματος και στα 2 datasets όταν το i παίρνει την τιμή 10 έχουμε μια αύξηση της τιμής της διαστατικότητας.

Πίνακας 5.2 1-NN classification average space (%) ($k=10$)

		Dimensionality						
ID	Dataset	4	6	8	10	12	14	16
1	LandSat-Satellite	13.87	16.14	16.73	18.14	18.99	19.02	18.96
2	CBF	12.12	13.31	13.88	14.73	16.21	15.57	17.18

Στον πίνακα [5.2] παρουσιάζονται τα αποτελέσματα από το average space.Το συμπέρασμα που μπορούμε να βγάλουμε είναι ότι όσο αυξάνεται η διαστατικότητα τόσο αυξάνεται και η τιμή του average space. Σε κάθε ένα dataset έχουμε για την μικρότερη τιμή του i χαμηλότερο ποσοστό για το average space.

Πίνακας 5.3 1-NN classification error rates (%) (i=4)

		Number of clusters						
ID	Dataset	4	6	8	10	12	14	16
1	LandSat-Satellite	89.38	89.38	89.38	89.38	89.38	89.38	89.38
2	CBF	10.54	10.54	10.54	10.54	10.54	10.54	10.54

Στον παραπάνω πίνακα έχουμε κρατήσει σταθερή την διαστατικότητα και αυξάνουμε σταδιακά τον αριθμό των συστάδων ανά 2. Αυτό όμως όπως μπορούμε να διαπιστώσουμε από τον πίνακα 5.3 δεν παίζει κανένα ρόλο στα αποτελέσματα μας καθώς και στα 2 dataset ο αριθμός του error παραμένει σταθερός.

Πίνακας 5.4 1-NN classification average space (%) (i=4)

		Number of Clusters						
ID	Dataset	4	6	8	10	12	14	16
1	LandSat-Satellite	28.67	20.33	15.25	12.36	10.31	8.75	7.90
2	CBF	27.13	20.05	16.15	13.88	12.25	10.96	9.95

Στον πίνακα [5.4] παρουσιάζουμε το average space για σταθερό ι=4 και αυξανόμενο αριθμό από συστάδες με πολλαπλάσια του 2 από 4 μέχρι 16. Τα αποτελέσματα που παίρνουμε είναι ότι όσο αυξάνεται από αριθμός από τα cluster τόσο καλύτερα αποτελέσματα μας επιστρέφει κ το average space.Δηλαδή, η απόδοση του average space βελτιώνεται καθώς αυξάνονται οι συστάδες γιατί υπάρχει μειώσει στο ποσοστό του χώρου που αναζητήσαμε.

ΚΕΦΑΛΑΙΟ 6ο

6.1 Γενικά Συμπεράσματα

Η ανάκτηση κοντινότερου γείτονα είναι μια κοινή λειτουργία σε ευρύ φάσμα των πραγματικών εφαρμογών. Η αυξανόμενη ποσότητα των αποθηκευμένων δεδομένων απαιτεί την ανάπτυξη τεχνικών και αλγορίθμων που θα είναι σε θέση να εκτελέσουν αποτελεσματικά και με ακρίβεια τις αναζητήσεις για τον πλησιέστερο γείτονα. Στην εργασία αυτή, προτείνουμε μία συστάδα με βάση τη μέθοδο (CLUREP) για την αναζήτηση ομοιότητας για τους σκοπούς της 1NN ταξινόμηση. Μέρος αυτής της μεθόδου μπορεί να χρησιμοποιηθεί σε συνδυασμό με μια πολυδιάστατη δομή ευρετηρίασης. Ωστόσο, εξετάζουμε την αποτελεσματικότητα αυτής της προσέγγισης σε μια διαδοχική αναζήτηση διότι η μείωση διαστάσεων είναι μία τεχνική που εφαρμόζεται σε στοιχεία που μπορούν όμως να μην είναι επαρκής για να παρέχουν τα μέσα για ένα πολυδιάστατο ευρετήριο ξεπερνώντας έτσι την διαδοχική σάρωση. Το κύριο συμπέρασμα είναι ότι, σε σύγκριση με διαδοχική αναζήτηση, η προτιμώμενη μέθοδος CLUREP είναι πολύ πιο γρήγορη, στην πλειοψηφία της από το σύνολο των δεδομένων που εξετάζονται στα πειράματα αυτά. Ένα δεύτερο συμπέρασμα είναι ότι όταν ο αριθμός των παραγόμενων συστάδων (clusters) αυξάνεται, η απόδοση του CLUREP βελτιώνεται. Ωστόσο, τα πειράματα δείχνουν ότι ο ρυθμός βελτίωσης "επιβραδύνεται", μετά την παραγωγή του από 8 σε 12 ομάδες. Ένα γενικότερο συμπέρασμα είναι ότι, όπως η διαστατικότητα από τα μετασχηματισμένα δεδομένα αυξάνεται έτσι και το σφάλμα του ποσοστού ταξινόμησης μειώνεται , με το κόστος του αυξανόμενου ποσοστού του όγκου των δεδομένων που θα πρέπει να αναζητηθούν. Στο πλαίσιο της εξόρυξη δεδομένων, μπορεί κανείς να πραγματοποιήσει αλλαγές μεταξύ της ακρίβειας και της αποδοτικότητα σε σχέση με τις απαιτήσεις που έχει η εξεταζόμενη αίτηση.

7. Βιβλιογραφία

Ιστοσελίδες:

1. http://en.Wikipedia.org/wiki/Cluster_analysis
2. http://en.Wikipedia.org/wiki/Data_mining
3. http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm
4. http://en.wikipedia.org/wiki/K-means_clustering
5. http://en.wikipedia.org/wiki/Hierarchical_clustering
6. http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/AppletH.html
7. http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/AppletKM.html
8. http://www.mathworks.com/products/matlab/
9. www.datamininglab.com/
10. http://www.cs.ucr.edu/~eamonn
11. http://en.wikipedia.org/wiki/Naive_Bayes_classifier
12. http://en.wikipedia.org/wiki/Bayes%27_theorem
13. http://archive.ics.uci.edu/ml/datasets.html
14. http://archive.ics.uci.edu/ml/datasets/MAGIC+Gamma+Telescope
15. http://archive.ics.uci.edu/ml/datasets/Statlog+%28Shuttle%29
16. http://archive.ics.uci.edu/ml/datasets/Statlog+%28Landsat+Satellite%29
17. http://archive.ics.uci.edu/ml/datasets/Letter+Recognition
18. http://archive.ics.uci.edu/ml/datasets/PenBased+Recognition+of+Handwritten+Digits
19. http://www.cs.uoi.gr/~pitoura/
20. http://en.wikipedia.org/wiki/Curse_of_dimensionality
21. http://en.wikipedia.org/wiki/Linear_classifier

22. http://en.wikipedia.org/wiki/Cross-validation_%28statistics%29
23. http://en.wikipedia.org/wiki/Decision_tree
24. http://en.wikipedia.org/wiki/Decision_tree_learning
25. http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/mlbook/ch3.pdf
26. http://decisiontrees.net/

Βιβλία & Αναφορές:

[1]. Evangelidis Georgios**,** Leonidas Karamitopoulos **Cluster-Based Similarity Search in Time Series**

[2]. Ougiaroglou Stefanos, Georgios Evangelidis, and Dimitris A. Dervos , An Adaptive Hybrid and Cluster-Based Model for speeding up the *k*-NN Classier,

[3]. Theodoridis S., Koutroumbas K. (2008) "*Pattern Recognition & Matlab Intro: Pattern Recognition*", 4th Edition, Academic Press

[4]. Fayyad, U.-M., Piatetsky-Shapiro, G,. Smyth, P(1996)..: "*From Data Mining to Knowledge Discovery: An Overview. In: Advances in Knowledge Discovery and Data Mining*", AAAI / MIT Press, pp. 1-30

[5]. MATLAB: The Language of Technical Computing

[6].MATLAB TUTORIAL: A practical Time-Series Tutorial with MATLAB

[7]. Pavel Berkhin, A Survey of Clustering Data Mining Techniques,

[8]. Ramakrishnan, R. (2002) “Database Management Systems”, 2nd Edition, Εκδόσεις Τζιόλα.

[9] Zezoula P., Amato G., Dohnal V., Batko M. (2006) “*Similarity Search: The Metric Space Approach*”, Springer Science & Business Media, Inc.

ΠΑΡΑΡΤΗΜΑ Α: Αλγόριθμος PAA

MATLAB codes – PAA

```
% Piecewise Aggregate Approximation
% x: m*(n+1) array of m time series of length n. The first column contains
the class label of each time series
% seg: the number of segments that will be generated
% When (n/seg) is not an integer, the time series is augmented by adding
zeros and then represented by PAA

function [y] = paa(x,seg)

x_label = x(:,1);% create a vector of the class labels
x(:,1) = []; % pull out the class labels from the dataset
points = size(x,2)/seg;    % find the number of points that will consist each
segment
if mod(size(x,2),seg) == 0 % if (n/segments) is an integer
    for j= 1:size(x,1)
        y(j,:) = mean(reshape(x(j,:),points,seg)); % PAA
    end
else    % if (n/seg) is not an integer
    for j= 1:size(x,1)
        a = x(j,:); % create a temporarily
        pad_points = (ceil(points) * seg) - length(a); % find the number of
zeros to be added
        a = [a zeros(1,pad_points)]; % create the augmented time series
        y(j,:) = mean(reshape(a,ceil(points),seg));    % PAA
    end
end
y = [x_label y];    % add the class labels to the tranformed dataset

end
```

ΠΑΡΑΡΤΗΜΑ Β: Αλγόριθμος 1-NN

MATLAB codes – Cluster-based 1-NN search

```
% The function "nn1_clurep" applies the algorithm k-means in order to partition
% the dataset into clusters, and then similarity search proceeds at each
% cluster sequentially.
% The algorithm reduces the search space for each cluster.
% k-means: k is the number of clusters

function [error, average_space ] = nn1_clurep(dataset,k)

correct = 0;
for j = 1 : size(dataset,1)
    query = dataset(j,:);
    % Pull-out label from the query
    query_label = query(1);
    % Remove label from the query
    query(1) = [];
    train = dataset;
    % Remove the query from the dataset
```

```
    train(j,:) = [];

    % Pull-out labels from the training set

    train_labels = train(:,1);

    % Remove labels from the training set

    train(:,1) = [];

    [cluster_id centroids sumd distances] = kmeans(train,k);

    distances = sqrt(distances);

    % insert 1) cluster_id, 2) distances fron centroids and 3) class_labels to

      the training set

    temp = [cluster_id distances train_labels train];

    % sort training set with respect to cluster_id

    temp = sortrows(temp,1);

    for i = 1 : k

       first = find(temp(:,1) == i , 1 , 'first');

       last  = find(temp(:,1) == i , 1 , 'last');

       cluster(i).id       = i;

       cluster(i).dist     = temp (first:last , 1+i);

       cluster(i).class    = temp (first:last , 2+k);

       cluster(i).data     = temp (first:last , 3+k:end);

       cluster(i).radius   = max( cluster(i).dist );

       cluster(i).mean     = mean( cluster(i).dist );

       cluster(i).std      = std( cluster(i).dist );

       cluster(i).ub       = cluster(i).mean + 3 * cluster(i).std;

    end


    %%%%%%%% calculate distances of the query from each centroid and sort them%
```

```
for i=1:k

    dqc(i,1) = i;

    dqc(i,2) = sqrt(sum((query-centroids(i,:)).^2));

end

dqc = sortrows(dqc,2);

primary_cluster = dqc(1,1);

% sort by cluster id (1,2,3,...k)

dqc = sortrows(dqc,1);

%%%%%%% end %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%


%%% calculate distances of the query from each cluster and sort them %%%%%%

for i=1:k

    dqcl(i,1) = i;

    dqcl(i,2) = max( 0 , dqc(i,2)-cluster(i).radius );

end




dqcl(primary_cluster,:) = [];

% sort by distance from cluster

dqcl = sortrows(dqcl,2);

%%%%% %%%%%%%%%%%%%%%%%%%%%%%%%% end %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%


%%%%%%%% 1-NN Classification for the Primary Cluster %%%%%%%%%%%%

vis = 0;

best_so_far = inf;
```

```
    for i =  1 : size (cluster(primary_cluster).data , 1)

        vis = vis +1;

        d = sqrt( sum ((query - cluster(primary_cluster).data(i,:)).^2));

        if d < best_so_far

            best_so_far = d;

            predicted_class = cluster(primary_cluster).class(i);

        end

    end

    %%%%%%%% 1-NN Classification for the Rest Clusters %%%%%%%%%%%

    for s = 1 : k-1

        if dqcl(s,2) < best_so_far

            if dqcl(s,2) > 0

                x = best_so_far - dqcl(s,2);

                upperb = max (cluster(dqcl(s,1)).radius - x , 0);

            else

                x = best_so_far - dqcl(s,2);

                upperb = max (  dqc(dqcl(s,1),2) - x , 0);

            end;

            for i = 1 : size (cluster(dqcl(s,1)).data , 1)

                if cluster(dqcl(s,1)).dist(i) >= upperb

                    vis = vis +1;

                    d = sqrt(sum((query - cluster(dqcl(s,1)).data(i,:)).^2));

                    if d < best_so_far

                        best_so_far = d;

                        predicted_class = cluster(dqcl(s,1)).class(i);

                    end
```

```
                end

            end

        end

    end

    if predicted_class == query_label

        correct = correct+1;

    end

    visited(j) = vis * 100 / size (train , 1)  ;

end

accuracy = correct / size(dataset,1);

error = (1 - accuracy)*100;

average_space = mean (visited);


end
```

Figure B2. Cluster-based 1-NN similarity search (partial cluster search)

ΠΑΡΑΡΤΗΜΑ Γ: Οπτικοποίηση των Αποτελεσμάτων

MATLAB codes

```
clear all
i=4,6,8,10,12,14,16;
train = importdata('CBF_TRAIN');
test = importdata('CBF_TEST');
dataset = [ train ; test ];
a = paa(dataset,i);
clust = a(:,1);
x = a(:,2);
y = a(:,3);
z = a(:,4);
colors = ['r','g','b','y'];


for j = 1:3
    scatter3(x((clust==j),1),y(clust==j),z(clust==j))
    hold on
end
title(sprintf('3D visualization i = %d',i))
xlabel('x')
ylabel('y')
zlabel('z')

hold off
```