Εθνικό Μετσόβιο Πολυτεχνείο

Διατμηματικό Μεταπτυχιακό Πρόγραμμα Εφαρμοσμένων Μαθηματικών Επιστημών

Μέθοδοι ομαδοποίησης στην εξόρυξη πληροφορίας από βάσεις δεδομένων

Clustering methods in Data Mining

Διπλωματική εργασία

της

Ευαγγελίας Τσιώκου

Επιτροπή: 1. Κοκκίνης Βασίλειος, Επίκουρος Καθηγητής

2. Κολέτσος Ιωάννης, Επίκουρος Καθηγητής (επιβλέπων)

3. Στεφανέας Πέτρος, Επίκουρος Καθηγητής

Αθήνα, Μάρτιος 2017

ΠΕΡΙΕΧΟΜΕΝΑ

Ευχαριστίες

Στην οικογένειά μου, στους φίλους μου και στον επιβλέποντα καθηγητή μου.

Περίληψη

Στην παρούσα διπλωματική εργασία αρχικά γίνεται αναφορά στην ανάγκη δημιουργίας μιας νέας δυναμικής τεχνολογίας, γνωστής ως Διαδικασία Ανακάλυψης Γνώσης από Βάσεις Δεδομένων και τη χρησιμότητά της για εξόρυξη χρήσιμης πληροφορίας από μεγάλες βάσεις ετερογενών δεδομένων. Στη συνέχεια παρουσιάζεται η διαδικασία ανάλυσης των δεδομένων σε συστάδες με χρήση της διαιρετικής και της ιεραρχικής μεθόδου συσταδοποίησης.

Εισαγωγή

Τα τελευταία χρόνια έχει παρατηρηθεί ραγδαία αύξηση στην παραγωγή και συλλογή δεδομένων. Ο τεράστιος όγκος ετερογενούς μίγματος δεδομένων μας οδήγησε λοιπόν στην ανάγκη εύρεσης τρόπου διαχωρισμού τους, με στόχο την εξαγωγή κρυμμένης χρήσιμης πληροφορίας από αυτά.

Αν και στη Χημεία έχουμε διδαχθεί πως τα συστατικά των ετερογενών μιγμάτων γίνονται διακριτά με το μάτι, δεν συμβαίνει το ίδιο και στις βάσεις δεδομένων. Θα μπορούσαμε να πούμε πως μοιάζουν περισσότερο με ομοιογενή μίγματα, των οποίων τα συστατικά δεν γίνονται ευδιάκριτα με μια απλή ματιά, επομένως γεννιέται η ανάγκη εύρεσης τεχνικών και εργαλείων, τα οποία έξυπνα θα μπορούσαν να μας εξάγουν χρήσιμη πληροφορία.

Η ανακάλυψη γνώσης από βάσεις δεδομένων, είναι η επεξεργασία και η εξερεύνηση μεγάλου όγκου δεδομένων προκειμένου να ανακαλυφθούν ομοιότητες μεταξύ τους με σκοπό την ομαδοποίηση τους και κατά συνέπεια την πιο εύκολη κατανόηση και διαχείρισή τους. Μια πολύ απλή συσχέτιση που μπορούμε να κάνουμε για να κατανοήσουμε όλοι τη χρησιμότητά της, είναι να φανταστούμε ένα δωμάτιο στο οποίο έχουμε διασκορπίσει όλα τα μικροπράγματα που έχουμε σπίτι μας. Φυσικά και αν ψάξουμε αρκετή ώρα θα βρούμε αυτό που θέλουμε, αλλά φανταστείτε πόσο εύκολα θα κάναμε τη δουλειά μας αν τα είχαμε ομαδοποιήσει με τέτοιο τρόπο ώστε να γνωρίζουμε τι ακριβώς έχουμε και την κατάλληλη στιγμή να μπορούμε εύκολα να το βρούμε. Ένα χαρακτηριστικό παράδειγμα ομαδοποίησης είναι ο τρόπος κατασκευής του σύγχρονου περιοδικού πίνακα, στον οποίο τα στοιχεία του έχουν ταξινομηθεί σε διαφορετικές ομάδες μεταξύ τους, όπου στην κάθε μια ανήκουν στοιχεία που έχουν παρόμοιες ιδιότητες, δηλαδή έχουν ομαδοποιηθεί με τέτοιο τρόπο που εύκολα μπορούμε να τα μελετήσουμε. Έτσι τα στοιχεία που ανήκουν στην ίδια ομάδα έχουν

ομοιότητες στα χαρακτηριστικά τους, ενώ τα στοιχεία που ανήκουν σε διαφορετικές ομάδες παρουσιάζουν ανομοιότητες.

Οι διαδικασίες επεξεργασίας δεδομένων, χωρίζονται σε δύο κατηγορίες:

1. τις *διαδικασίες ανεύρεσης*, των οποίων στόχος είναι η ανακάλυψη και η κατασκευή συσχετίσεων από τα δεδομένα, σύμφωνα με τις οποίες γίνεται η ομαδοποίησή τους και
2. τις *διαδικασίες επιβεβαίωσης,* των οποίων στόχος είναι η λήψη αποφάσεων δεδομένης της δομής των δεδομένων, σύμφωνα με τις οποίες γίνεται ταξινόμηση τους.

Κοινός στόχος και στις δύο περιπτώσεις είναι η ομαδοποίηση των δεδομένων.

Η *ομαδοποίηση* (clustering) ανήκει στις διαδικασίες ανεύρεσης και είναι μια τεχνική ανάλυσης δεδομένων σε ανόμοιες ομάδες, στην οποία δεν γίνεται καμία υπόθεση σχετικά με τον αριθμό των ομάδων που θα δημιουργηθούν ή τη δομή της ομάδας. Είναι μια περιγραφική διαδικασία που χρησιμοποιείται για την ανίχνευση και κατασκευή ενός πεπερασμένου πλήθους ανόμοιων συστάδων (clusters). Η ομαδοποίηση γίνεται βάσει των ομοιοτήτων ή των αποστάσεων (ανομοιότητες) των δεδομένων που εξετάζονται. Οι είσοδοι που απαιτούνται, είναι μέτρα ομοιότητας ή δεδομένα, από τις ομοιότητες των οποίων μπορούν να υπολογιστούν. Η πιο σημαντική συνθήκη, κάτω από την οποία χρησιμοποιούνται οι τεχνικές δημιουργίας συστάδων είναι η μελέτη των δεδομένων και η ανακάλυψη χρήσιμης ομαδοποίησής τους. Δεν είναι όμως ο μοναδικός λόγος επιλογής αυτής της τεχνικής, αλλά μπορεί να γίνει για παράδειγμα, γιατί το κόστος της απόκτησης ενός αρχικά ταξινομημένου δείγματος, μπορεί να είναι μεγάλο, ή ίσως η δομή των κατηγοριών να μεταβάλλεται με το χρόνο. Αυτό όμως που είναι βασικό να σημειωθεί, είναι πως πάντα θα υπάρχει μια ποικιλία από εναλλακτικές ομαδοποιήσεις ανάλογα με το σύνολο των δεδομένων. Το είδος της ομαδοποίησης θα εξαρτάται σε μεγάλο βαθμό, από τα χαρακτηριστικά των δεδομένων. Κάποιες ομαδοποιήσεις που θα προκύψουν θα είναι

χρήσιμες και κάποιες άλλες όχι τόσο. Για παράδειγμα μια ταξινόμηση καλλυντικών σε ομάδες όπως ματιών, χεριών ή προσώπου, σώματος, πιθανόν να είναι πιο χρήσιμη, από κάποια που βασίζεται στο χρώμα τους.

Η *ταξινόμηση* (classification) ανήκει στις διαδικασίες επιβεβαίωσης και αναφέρεται σε ένα γνωστό αριθμό ομάδων και ο λειτουργικός της στόχος είναι να αναθέτει νέα δεδομένα σε μια από τις προκαθορισμένες ομάδες. Γενικότερα, ο στόχος της διαδικασίας αυτής είναι η δημιουργία ενός μοντέλου, το οποίο αρχικά εκπαιδεύεται σε δοκιμαστικά δεδομένα (test data) για να εξακριβωθεί η ακρίβειά του και στη συνέχεια να μπορεί να χρησιμοποιηθεί για την κατηγοριοποίηση των μελλοντικών δεδομένων.

Κεφάλαιο 1ο: Ανακάλυψη Γνώσης από Δεδομένα

1.1. Η ανάγκη για Εξόρυξη Γνώσης

Την τελευταία εικοσαετία έχει παρατηρηθεί ταχεία αύξηση της παραγωγής και συλλογής δεδομένων εξαιτίας της ευρέως διαδεδομένης χρήσης των υπολογιστών σε κάθε τομέα της ζωής μας. Η χρόνια συλλογή και συνεχής αποθήκευση δεδομένων, με τη χρήση διάφορων τεχνικών, έφερε ως αποτέλεσμα την συσσώρευση τεράστιων όγκων δεδομένων σε μεγάλους αποθηκευτικούς χώρους με ελάχιστο κόστος. Η συγκέντρωση όλων αυτών των ανομοιογενών δεδομένων δημιούργησε προβλήματα σε σχέση με τη γνώση της ποιότητας τους, πράγμα το οποίο οδήγησε τους ερευνητές στη δημιουργία κατάλληλων εργαλείων και μεθόδων για την εξερεύνησή τους. Η Διαδικασία Ανακάλυψης Γνώσης από Βάσεις Δεδομένων (Knowledge Discovery in Databases) και η Εξόρυξη Γνώσης (Data Mining) αποτελούν μία λύση στο συγκεκριμένο πρόβλημα καθώς εξάγουν από τα ακατέργαστα δεδομένα γνώση, με τη μορφή προτύπων. Η εξόρυξη γνώσης αποτελεί στην πραγματικότητα ένα από τα βήματα της διαδικασίας ανακάλυψης γνώσης από βάσεις δεδομένων.

Η Ανακάλυψη Γνώσης από Βάσεις Δεδομένων είναι μια νέα δυναμική τεχνολογία, που βοηθά τον κάτοχο μεγάλης ποσότητας δεδομένων να ανακαλύψει κρυμμένη πληροφορία στα ανεκμετάλλευτα δεδομένα που έχει και να την εκμεταλλευτεί ανάλογα με τις ανάγκες του.

[1] Εικόνα

Η λειτουργία της Ανακάλυψης Γνώσης από Βάσεις Δεδομένων έχει να κάνει ουσιαστικά με δεδομένα που έχουν συλλεχθεί ήδη για κάποιο άλλο σκοπό. Αυτό σημαίνει πως οι στόχοι της εξόρυξης γνώσης δεν επηρεάζουν τον τρόπο με τον οποίο συλλέγονται τα δεδομένα. Αυτή θα μμπορούσε να είναι μία διαφορά της εξόρυξης γνώσης με τις στατιστικές μελέτες, όπου τα δεδομένα συγκεντρώνονται με καθορισμένους τρόπους για την απάντηση συγκεκριμένων ερωτημάτων. Γι αυτόν τον λόγο η μέθοδος της Ανακάλυψης Γνώσης από Βάσεις Δεδομένων συχνά αναφέρεται και ως δευτερεύουσα ανάλυση δεδομένων.

1.2. Σημασία της εξαγωγής πληροφοριών

Ο ρόλος της εξαγωγής πληροφοριών, στα πλαίσια της ανάκτησης πληροφοριών και διαχείρισης γνώσης, είναι η αναγνώριση εξειδικευμένης και στοχευμένης πληροφορίας και η εξαγωγή γνώσης από μη δομημένα δεδομένα με μηχανικό (αυτόματο) τρόπο. [2]

[1] https://www.google.gr/search?q=%CF%83%CE%B7%CE%BC%CE%B1%CF%83%CE%B9%CE%B1+%CE%B5%CE%BE%CE%BF%CF%81%CF%85%CE%BE%CE%B7%CF%82+%CE%B4%CE%B5%CE%B4%CE%BF%CE%BC%CE%B5%CE%BD%CF%89%CE%BD&espv=2&biw=1366&bih=667&source=lnms&tbm=isch&sa=X&ved=0ahUKEwjtopj58NXOAhUHvRoKHT9WAksQ_AUIBigB#imgrc=GlszLvy3QeaVJM%3A

[2] https://el.wikipedia.org/wiki/%CE%95%CE%BE%CE%B1%CE%B3%CF%89%CE%B3%CE%AE_%CF%80%CE%BB%CE%B7%CF%81%CE%BF%CF%86%CE%BF%CF%81%CE%B9%CF%8E%CE%BD

Αντίθετα με την κλασσική ανάκτηση πληροφοριών, σύμφωνα με την οποία η αναζήτηση γίνεται με βάση συγκεκριμένες λέξεις-κλειδιά και το αποτέλεσμα περιλαμβάνει μόνο κείμενα στα οποία βρίσκεται (ενδεχομένως) η χρήσιμη πληροφορία, η εξόρυξη πληροφοριών στοχεύει ακριβώς στην αναγνώριση της χρήσιμης μόνο πληροφορίας και το περιβάλλον (context) στο οποίο αυτή εμφανίζεται.

Δεδομένου του μεγάλου όγκου πληροφοριών που παράγονται και διακινούνται σήμερα (κύριο χαρακτηριστικό του διαδικτύου), το ζητούμενο στις μέρες μας είναι όχι απλώς η κατοχή της πληροφορίας και η πρόσβαση σε αυτή, ο οποιοσδήποτε σήμερα μπορεί να έχει πρόσβαση σε σχεδόν οποιαδήποτε πληροφορία, αλλά η διαχείριση της πληροφορίας και ο εντοπισμός της «σχετικής» πληροφορίας. Έτσι, ενώ με μια κλασσική μηχανή αναζήτησης ο ενδιαφερόμενος θα λάβει ως απάντηση ένα σύνολο κειμένων που ενδεχομένως περιέχουν την απάντηση που περιμένει, η εξόρυξη πληροφοριών στοχεύει στην απάντηση και μόνο σε αυτή.

1.3. Βήματα Ανακάλυψης Γνώσης από Βάσεις Δεδομένων

Η Ανακάλυψη Γνώσης από Βάσεις Δεδομένων αποτελείται από συγκεκριμένα στάδια. Πρόκειται για την ανακάλυψη ή παραγωγή λειτουργικής γνώσης μέσα από την ανάλυση δεδομένων. Αναφέρεται σε ολόκληρη τη διαδικασία, από τη συλλογή των δεδομένων μέχρι και την αξιοποίηση των αποτελεσμάτων σε πιο πρακτικό επίπεδο.

Τα *βασικά βήματα* από τα οποία αποτελείται η διαδικασία ανεύρεσης γνώσης είναι τα ακόλουθα:

1. ***Συλλογή δεδομένων (Data collection)*:** Το πρώτο βήμα είναι η επιλογή των δεδομένων που θα χρησιμοποιηθούν στη διαδικασία της εξόρυξης πληροφοριών. Στο στάδιο αυτό γίνεται η συλλογή και η αποθήκευση των δεδομένων. Η συλλογή των δεδομένων συνήθως γίνεται είτε αυτόματα, π.χ. Με χρήση αισθητήρων, είτε μη αυτόματα, π.χ. Με χρήση ερωτηματολογίων. Δυσλειτουργία στους αισθητήρες ή αδυναμία απάντησης κάποιας ερώτησης στα ερωτηματολόγια μπορεί να οδηγήσει σε θορυβώδη ή ελλιπή δεδομένα. Τα συγκεκριμένα προβλήματα, που ενδεχομένως να προκύψουν κατά τη συλλογή δεδομένων, αναλαμβάνει να τα αντιμετωπίσει επόμενο στάδιο.

***2. Ενσωμάτωση δεδομένων (Data integration)*:** Σε αυτό το βήμα τα ανομοιογενή δεδομένα που έχουν συλλεχθεί από πολλές διαφορετικές πηγές, ενσωματώνονται σε μια κοινή βάση δεδομένων.

***3. Επιλογή δεδομένων (Data selection)*:** Από όλα εκείνα τα δεδομένα που έχουμε συλλέξει στην κοινή βάση δεδομένων, ανακτώνται εκείνα που σχετίζονται με το προς ανάλυση πρόβλημα που θα ακολουθήσει.

***4. Προεπεξεργασία δεδομένων (Data preprocessing)*:** Σε αυτό το πολύ σημαντικό βήμα, αυτό που μας απασχολεί είναι η αξιοπιστία των δεδομένων. Στην προσπάθεια δημιουργίας ενός τέτοιου αξιόπιστου συνόλου πραγματοποιείται ο καθαρισμός τους, αφαιρώντας από τη βάση δεδομένων αυτά που παράγουν θόρυβο, δηλαδή όλα εκείνα τα στοιχεία που μπορούν να επηρεάσουν ή και να διαστρεβλώσουν το αποτέλεσμα. Το βήμα αυτό μπορεί να απαιτήσει έως και το 60% της συνολικής προσπάθειας και αυτό γιατί, αν τα δεδομένα δεν είναι <<καθαρά>>, δεν έχει νόημα να μιλάμε για ποιότητα αποτελεσμάτων. Η διαδικασία αυτή περιλαμβάνει την απομάκρυνση των λανθασμένων

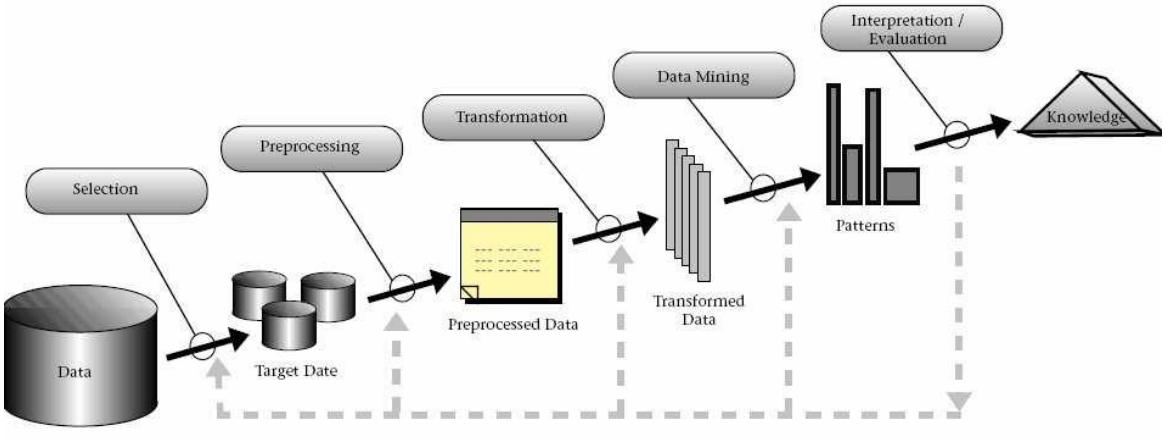
δεδομένων - θορύβου (noise) με τη χρήση στατιστικών μεθόδων ή με τη βοήθεια ενός αλγορίθμου εξόρυξης δεδομένων.

***5. Τροποποίηση δεδομένων (Data transformation)*:** Μέσω αυτού του βήματος τα δεδομένα που έχουμε επιλέξει προσεκτικά μετασχηματίζονται έτσι ώστε η μορφή τους να είναι κατάλληλη για την διαδικασία της εξόρυξης. Για να πετύχουμε αυτό το σκοπό εφαρμόζουμε μεθόδους μείωσης διαστάσεων και μετασχηματισμού χαρακτήρων.

***6. Εξόρυξη δεδομένων (Data mining)*:** Είναι το σημαντικότερο από τα βήματα της διαδικασίας και αυτό γιατί στο συγκεκριμένο στάδιο, ποικίλες εξελιγμένες τεχνικές (ταξινόμηση, συσταδοποίηση, παλινδρόμηση) χρησιμοποιούνται για την εξαγωγή προτύπων.

***7. Αξιολόγηση προτύπων (Pattern evaluation)*:** Στο βήμα αυτό αναγνωρίζονται χρήσιμα πρότυπα που αναπαριστούν γνώση, βάσει συγκεκριμένων μέτρων αξιολόγησης (evaluation measures).

***8. Αναπαράσταση γνώσης (Knowledge representation)*:** Στο τελικό αυτό στάδιο, η γνώση που έχει ανακαλυφθεί παρουσιάζεται στον χρήστη, βοηθώντας τον έτσι να κατανοήσει και να ερμηνεύσει τα αποτελέσματα της εξόρυξης δεδομένων.



[3] Εικόνα

Τα βήματα 2 έως 5 είναι διαφορετικές μορφές προεπεξεργασίας δεδομένων, όπου τα δεδομένα προετοιμάζονται για την εξόρυξη. Το βήμα εξόρυξη δεδομένων μπορεί να αλληλεπιδράσει με το χρήστη ή μια βάση δεδομένων. Τα ενδιαφέροντα πρότυπα παρουσιάζονται στον χρήστη και μπορεί να αποθηκευτούν ως νέα γνώση στη βάση

δεδομένων. Η προηγούμενη προβολή εμφανίζει την εξόρυξη δεδομένων ως ένα βήμα της διαδικασίας ανακάλυψης γνώσης και περιλαμβάνει την εφαρμογή αλγορίθμων ανάλυσης δεδομένων και Εξόρυξης Γνώσης που παράγουν μία συγκεκριμένη απαρίθμηση των προτύπων (ή μοντέλων) πάνω στα δεδομένα. Ωστόσο στη βιομηχανία, στα μέσα μαζικής ενημέρωσης και στην έρευνα περιβάλλοντος, ο όρος εξόρυξη δεδομένων χρησιμοποιείται συχνά για να αναφέρεται σε ολόκληρη τη διαδικασία ανακάλυψης γνώσης (ίσως επειδή ο όρος είναι μικρότερος από την ανακάλυψη γνώσης από δεδομένα). Ως εκ τούτου, έχουμε υιοθετήσει μια ευρεία άποψη της εξόρυξης δεδομένων λειτουργικά. Εξόρυξη δεδομένων είναι η διαδικασία ανακάλυψης ενδιαφερόντων προτύπων και γνώσης από μεγάλες πηγές δεδομένων. Οι πηγές δεδομένων μπορεί να περιλαμβάνουν βάσεις δεδομένων, αποθήκες δεδομένων ή άλλες αποθήκες πληροφοριών.

[3] http://docplayer.gr/docs-images/17/163502/images/5-0.jpg

1.4. Μέθοδοι Εξόρυξης Γνώσης από δεδομένα

Υπάρχει μια μεγάλη ποικιλία μεθόδων εξόρυξης γνώσης από δεδομένα.[4] Ανάλογα με το είδος των δεδομένων και το είδος της γνώσης που εξάγεται, αυτές διαχωρίζονται σε διαφορετική κατηγορία. Κάποιες από τις κατηγορίες, είναι:

- ***Κανόνες συσχέτισης*** (*Mining Association Rules*): Οι κανόνες συσχέτισης ανακαλύπτουν κρυμμένες «συσχετίσεις» μεταξύ των γνωρισμάτων ενός συνόλου των δεδομένων. Αυτοί οι συσχετισμοί παρουσιάζονται στη μορφή $A \to B$, όπου τα A και B αποτελούν σύνολα που αναφέρονται στα χαρακτηριστικά του συνόλου δεδομένων που αναλύουμε. Δεδομένου ενός συνόλου από δεδομένα, ένας κανόνας συσχέτισης $A \to B$ προβλέπει την εμφάνιση των χαρακτηριστικών του συνόλου B δεδομένης της εμφάνισης των χαρακτηριστικών του συνόλου A.

Κλασικό πεδίο εφαρμογής των κανόνων συσχέτισης είναι η ανάλυση του καλαθιού της αγοράς (market basket analysis), όπου σκοπός είναι να αναγνωρισθούν προϊόντα που οι καταναλωτές τείνουν να αγοράζουν μαζί. Οι συναλλαγές μπορεί να είναι, για παράδειγμα: {ψωμί, γάλα}, {ψωμί, πάνες, μπύρα, αυγά}, {γάλα, πάνες, μπύρα, σόδα}, {ψωμί, γάλα, πάνες, μπύρα} και {ψωμί, γάλα, πάνες, σόδα}, και κάποιοι κανόνες συσχέτισης σε αυτές είναι {Πάνες} → {μπύρα}, {μπύρα, ψωμί} → {γάλα}, {γάλα, ψωμί} → {αυγά, σόδα}. Ο τελευταίος κανόνας, για παράδειγμα, φανερώνει ότι είναι πολύ πιθανό όποιος αγοράζει γάλα και ψωμί να αγοράσει, επίσης, αυγά και σόδα.

[4] https://el.wikipedia.org/wiki/%CE%9A%CE%B1%CF%84%CE%B7%CE%B3%CE%BF%CF%81%CE%B9%CE%BF%CF%80%CE%BF%CE%AF%CE%B7%CF%83%CE%B7

Εξάγοντας χρήσιμα συμπεράσματα μέσω των κανόνων συσχέτισης, το τμήμα προώθησης του παντοπωλείου μπορεί να τοποθετήσει κατάλληλα τα προϊόντα στα ράφια, να κάνει την κατάλληλη καμπάνια προώθησης τους και να διαχειριστεί πιο αποδοτικά τα αποθεματικά του.

- ***Συσταδοποίηση***: είναι μια περιγραφική μέθοδος. Έχοντας ένα σύνολο δεδομένων, στόχος της συσταδοποίησης είναι η δημιουργία συστάδων, δηλαδή ομάδων, οι οποίες θα περιέχουν όμοια ή παρεμφερή δείγματα. Ο αριθμός των τελικών συστάδων δεν είναι γνωστός από την αρχή. Ένα παράδειγμα συσταδοποίησης, είναι ο διαμερισμός των ανθρώπων σε ομάδες με βάση την οικονομική τους κατάσταση και το βιοτικό τους επίπεδο. Σε επόμενο κεφάλαιο θα γίνει αναλυτική ανάπτυξη της χρησιμότητάς της.

- ***Ταξινόμηση***: πρόκειται για μια προγνωστική μέθοδο που συναντάται στη βιβλιογραφία και ως κατηγοριοποίηση. Είναι μια τεχνική της εξόρυξης γνώσης από πηγές δεδομένων, κατά την οποία ένα στοιχείο ανατίθεται σε ένα προκαθορισμένο σύνολο κατηγοριών. Γενικότερα στόχος της είναι η ανάπτυξη ενός μοντέλου που αργότερα θα μπορεί να χρησιμοποιηθεί για την ταξινόμηση μελλοντικών δεδομένων. Τέτοια παραδείγματα είναι ο διαχωρισμός των emails με βάση την επικεφαλίδα τους ή το περιεχόμενό τους, η πρόβλεψη καρκινικών κυττάρων χαρακτηρίζοντάς τα ως καλοήθη ή κακοήθη, η κατηγοριοποίηση πελατών μιας τράπεζας ανάλογα με την πιστωτική τους ικανότητα κ.α.

Η ταξινόμηση μπορεί να περιγραφεί ως μια διαδικασία δύο βημάτων:

1. **Εκμάθηση** (Learning): Στο πρώτο βήμα της διαδικασίας προσδιορίζεται το μοντέλο με βάση ένα σύνολο προκατηγοριοποιημένων δεδομένων, που ονομάζονται δεδομένα εκπαίδευσης (training data). Τα δεδομένα εκπαίδευσης αναλύονται από ένα αλγόριθμο ταξινόμησης, προκειμένου να σχηματιστεί το μοντέλο. Λόγω του ότι τα δεδομένα

εκπαίδευσης ανήκουν σε μία προκαθορισμένη κατηγορία, η οποία είναι γνωστή, η ταξινόμηση αποτελεί μέθοδο εποπτευόμενης μάθησης (supervasid learning). Το μοντέλο, που λέγεται και αλλιώς ταξινομητής (classifier), αναπαρίσταται με τη μορφή κανόνων ταξινόμησης (classification rules), δέντρων απόφασης (decision trees) ή μαθηματικών τύπων.

2. **Κατηγοριοποίηση** (Classification): Μετά τη δημιουργία του μοντέλου, το επόμενο βήμα είναι η αξιολόγησή του. Για να επιτευχθεί αυτό, χρησιμοποιούμε τα δοκιμαστικά δεδομένα (test data) για να εξετάσουν την ακρίβεια του μοντέλου. Το μοντέλο κατηγοριοποιεί τα δοκιμαστικά δεδομένα. Έπειτα, η κατηγορία που σχηματίστηκε με βάση τα δοκιμαστικά δεδομένα συγκρίνεται με την πρόβλεψη που έγινε για τα δεδομένα εκπαίδευσης, τα οποία είναι ανεξάρτητα από αυτά της δοκιμής. Η ακρίβεια του μοντέλου υπολογίζεται με το ποσοστό των δειγμάτων δοκιμής που κατηγοριοποιήθηκαν σωστά σε σχέση με το υπό εκπαίδευση μοντέλο.

Το προερχόμενο μοντέλο μπορεί να αναπαρασταθεί σε διάφορες μορφές, όπως οι κανόνες ταξινόμησης (δηλαδή κανόνες IF-THEN), δέντρα απόφασης, μαθηματικούς τύπους κ.α. Για παράδειγμα ένα δέντρο απόφασης, είναι ένα διάγραμμα ροής, σαν δομή δέντρου, όπου κάθε κόμβος δηλώνει μια δοκιμή σε μια τιμή παραμέτρου, κάθε κλάδος αντιπροσωπεύει ένα αποτέλεσμα μιας δοκιμής και τα φύλλα των δέντρων αντιπροσωπεύουν τις κατηγορίες ή κατανομή κατηγορίας.

Στην περίπτωση που το μοντέλο κριθεί αποδεκτό, τότε μπορεί να χρησιμοποιηθεί για την κατηγοριοποίηση μελλοντικών δειγμάτων, των οποίων η κατηγοριοποίηση είναι άγνωστη.

- ***Παλινδρόμηση*** (*Regression*): είναι μια ευρέως χρησιμοποιούμενη στατιστική τεχνική μοντελοποίησης για την έρευνα της συσχέτισης μεταξύ μίας εξαρτώμενης μεταβλητής και μιας ή περισσότερων ανεξάρτητων μεταβλητών. Χρησιμοποιείται με

σκοπό την εκχώρηση δεδομένων σε μία πραγματική μεταβλητή πρόβλεψης, όπως ισχύει και στην περίπτωση της κατηγοριοποίησης όταν είναι διακριτή, αλλιώς καλείται παλινδρόμηση αν η μεταβλητή είναι συνεχής. Η παλινδρόμηση προϋποθέτει ότι τα σχετικά δεδομένα ταιριάζουν με μερικά γνωστά είδη συνάρτησης και μετά καθορίζει την καλύτερη συνάρτηση αυτού του είδους που μοντελοποιεί τα δεδομένα που έχουν δοθεί. Αποτέλεσμα της παλινδρόμησης, όταν χρησιμοποιείται ως τεχνική εξόρυξης δεδομένων, αποτελεί ένα μοντέλο που χρησιμοποιείται αργότερα για να προβλέψει τις τιμές της κατηγορίας για τα νέα δεδομένα. Τέτοια παραδείγματα εφαρμογής της παλινδρόμησης αποτελεί η πρόβλεψη της συζήτησης για ένα νέο προϊόν ή υπηρεσία συναρτήσει των δαπανών διαφήμισης ή ο υπολογισμός της ταχύτητας του ανέμου σε σχέση με την θερμοκρασία, την υγρασία και την ατμοσφαιρική πίεση του περιβάλλοντος.

1.5. Τομείς εφαρμογής

Οι τομείς εφαρμογής[5] της εξόρυξης γνώσης είναι πάρα πολλοί και ιδιαιτέρως σημαντικοί για τον άνθρωπο. Παρακάτω γίνεται μια μικρή αναφορά σε κάποιους βασικούς.

- ***Ιατρική***

Τα τελευταία χρόνια με την ανάπτυξη κλάδων της Ιατρικής, όπως η γενετική και η βιοϊατρική, αναδείχθηκε η χρησιμότητα της Εξόρυξης Δεδομένων στην Ιατρική. Στον τομέα της γενετικής, στόχος είναι η κατανόηση και η χαρτογράφηση της σχέσης μεταξύ της μεταβολής των ακολουθιών του ανθρώπινου DNA και της προδιάθεσης κάποιας ασθένειας. Η Εξόρυξη Δεδομένων είναι ένα εργαλείο, το οποίο μπορεί να βοηθήσει στη βελτίωση της διάγνωσης, της πρόληψης και κατ' επέκταση της θεραπείας ασθενειών.

[5] https://el.wikipedia.org/wiki/%CE%95%CE%BE%CF%8C%CF%81%CF%85%CE%BE%CE%B7_%CE%B4%CE%B5%CE%B4%CE%BF%CE%BC%CE%AD%CE%BD%CF%89%CE%BD

Ένας από τους κύριους στόχους, που συνδέεται με την ανάλυση του DNA, είναι η σύγκριση ποικίλων ακολουθιών και η αναζήτηση ομοιοτήτων μεταξύ των δεδομένων του DNA. Η σύγκριση γίνεται μεταξύ της γονιδιακής ακολουθίας υγιών και βλαβερών ιστών, για να βρεθούν τυχόν διαφορές ανάμεσα τους.

Τα εργαλεία οπτικοποίησης παίζουν, επίσης, έναν σημαντικό ρόλο στην εξόρυξη δεδομένων ως προς την βιοϊατρική. Τα εργαλεία αυτά μπορούν να παρουσιάσουν πολύπλοκες δομές γονιδίων σε γράφους, δένδρα και αλυσίδες. Η οπτική παρουσίαση συμβάλει στην καλύτερη κατανόηση αυτών των δομών και στην εξερεύνηση των δεδομένων.

- ***Οικονομία***

Τα δεδομένα που συλλέγονται από διάφορα οικονομικά ινστιτούτα, όπως οι τράπεζες, συγκεντρώνονται αρχικά στην αποθήκη δεδομένων (data warehouse). Οι τεχνικές της πολυδιάστατης ανάλυσης δεδομένων χρησιμοποιούνται για την ανάλυση τέτοιων δεδομένων που συλλέγονται στην αποθήκη δεδομένων για τις γενικές ιδιότητές του.

- ***Τηλεπικοινωνία***

Τα τηλεπικοινωνιακά δεδομένα που συλλέγονται, περιλαμβάνουν τον τύπο κλήσης, την τοποθεσία του καλούντος και του κληθέντος, τον χρόνο κλήσης, την διάρκεια κλήσης κ.α. Η πολυδιάστατη ανάλυση βοηθά στον προσδιορισμό και στην σύγκριση του φορτιού του συστήματος, κίνηση δεδομένων, κέρδος κ.α.

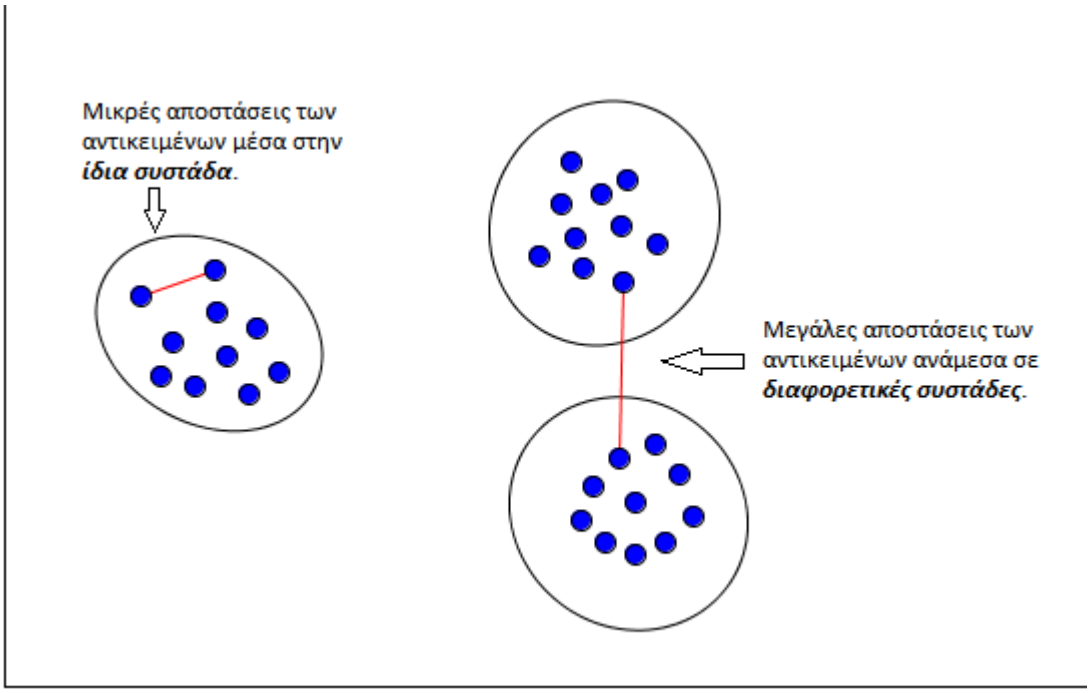
Το κυρίως πρόβλημα που αντιμετωπίστηκε από την βιομηχανία τηλεπικοινωνιών είναι οι παράνομες δραστηριότητες. Αυτές οι δραστηριότητες μπορεί να έχουν να κάνουν με σκόπιμες κλήσεις κατά την ώρα αιχμής, περιοδικές κλήσεις κ.α. με αποτέλεσμα να

επιδρούν αρνητικά στην επίδοση του δικτύου επικοινωνιών. Μέθοδοι όπως η συσταδοποίηση, συνεισφέρει στην ανίχνευση παράνομων προτύπων βελτιώνοντας την αποτελεσματικότητα των υπηρεσιών τηλεπικοινωνίας.

Κεφάλαιο 2ο: Συσταδοποίηση

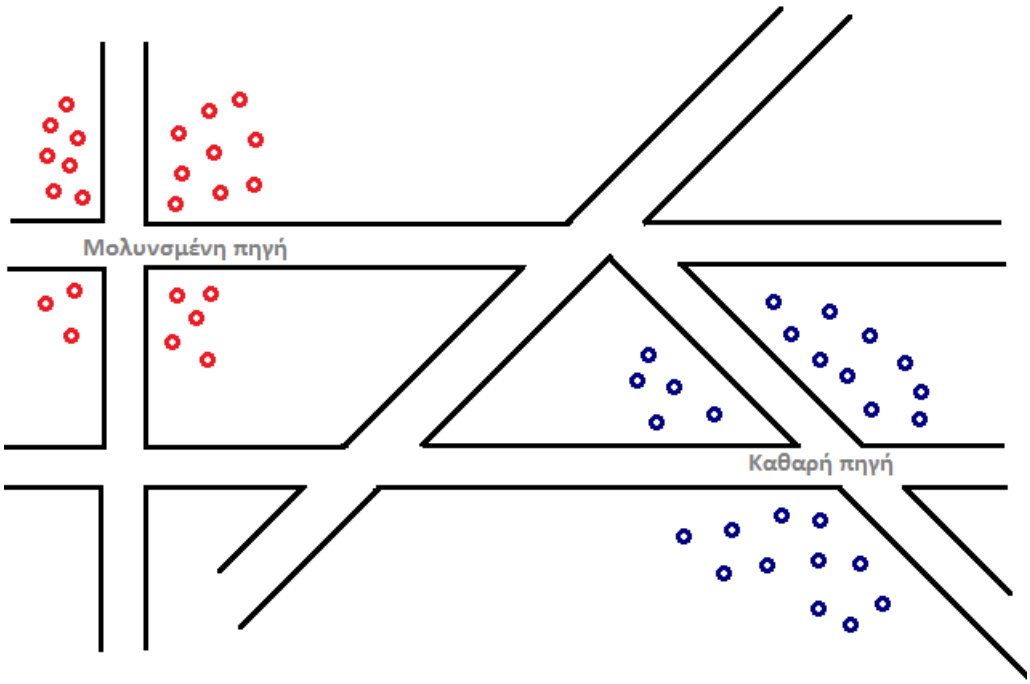
2.1. Εισαγωγή στην έννοια της συσταδοποίησης

Συσταδοποίηση (clustering) ή ομαδοποίηση είναι η διαδικασία εκείνη κατά την οποία ένα σύνολο από πρότυπα (παρατηρήσεις, δεδομένα ή διανύσματα χαρακτηριστικών), διαχωρίζονται σε ένα σύνολο από λογικές ομάδες (συστάδες - clusters). Η καταχώρηση δεδομένων στην ίδια ομάδα μεταφράζεται ως ομοιότητα των δεδομένων αυτών και αντίστροφα (αντικείμενα που ταξινομούνται σε διαφορετικές ομάδες είναι ανόμοια). Όσο μεγαλύτερη είναι η ομοιότητα (ή ομοιογένεια) μέσα σε μία ομάδα, τόσο μεγαλύτερη είναι η διαφορά μεταξύ των ομάδων και τόσο πιο επιτυχημένη η ομαδοποίηση. Η ομοιότητα ή μη, μεταξύ των δεδομένων, ουσιαστικά εξαρτάται από το συγκεκριμένο πρόβλημα που μελετάται, τη μορφή των δεδομένων και τις προς εξέταση μεταβλητές τους. Τα δεδομένα που εξετάζονται μπορούν να αναφερθούν με διαφορετικούς όρους: *πρότυπα*, *διανύσματα κ.α.*



Η συσταδοποίηση είναι μια διαδικασία που εντάσσεται γενικότερα στην μη επιβλεπόμενη μάθηση (unsupervised learning). Υπάρχει διαφορά μεταξύ επιβλεπόμενης μάθησης (supervised learning) και μη επιβλεπόμενης μάθησης (unsupervised learning). Στην επιβλεπόμενη μάθηση ή κατηγοριοποίηση ένα σύνολο από προ-ομαδοποιημένα στοιχεία είναι διαθέσιμο, και αυτό που μας ζητείται είναι να εντάξουμε ένα νέο στοιχείο σε κάποια από τις υπάρχουσες ομάδες. Συνήθως τα προ-ομαδοποιημένα στοιχεία χρησιμοποιούνται για να περιγράψουν τις διαφορετικές ομάδες – κλάσεις στις οποίες θα εντάξουμε νέα στοιχεία. Αντίθετα στην μη επιβλεπόμενη μάθηση ή συσταδοποίηση το πρόβλημα είναι να ομαδοποιήσουμε σε λογικές κλάσεις τα στοιχεία, χωρίς καμιά γνώση για προ-υπάρχουσες ομάδες και τον αριθμό των ομάδων που θα δημιουργηθούν όπως και το περιεχόμενό τους. Έτσι η συσταδοποίηση είναι απόλυτα οδηγημένη από τα δεδομένα (data driven) και παράγεται από αυτά.

Ένα διάσημο παράδειγμα της συσταδοποίησης, για να λύσει ένα πρόβλημα, πραγματοποιήθηκε το 1854 στο Soho του Λονδίνου και έγινε εξ ολοκλήρου χωρίς υπολογιστές[1]. Ο παθολόγος John Snow, που ασχολήθηκε με την επιδημία χολέρας, σημείωσε τις θέσεις των περιστατικών σε έναν χάρτη της πόλης. Μια πρόχειρη εικονογράφηση που υποδηλώνει τη διαδικασία παρουσιάζεται στο Σχ. 1.1.



Σχήμα 2.1: *Σημείωση των θέσεων περιστατικών χολέρας σε έναν χάρτη του Soho του Λονδίνου.*[6]

Παρατήρησε πως τα περιστατικά συγκεντρώνονταν γύρω από μερικές διασταυρώσεις δρόμων. Οι άνθρωποι που ζούσαν γύρω από τις περιοχές των πηγαδιών που είχαν μολυνθεί αρρώστησαν, ενώ οι άνθρωποι που ζούσαν γύρω από τις περιοχές των πηγαδιών που δεν είχαν μολυνθεί, δεν αρρώστησαν. Χωρίς την ικανότητα της ομαδοποίησης των δεδομένων, η αιτία της Χολέρας δεν θα είχε ανακαλυφθεί.

[6] http://en.wikipedia.org/wiki/1854_Broad_Stret_cholera_outbreak.

2.2. Ορισμός και στόχος της συσταδοποίησης

Δοθέντος ενός συνόλου δεδομένων - διανυσμάτων $X = \{x_1, x_2, x_3, ..., x_n\}$, ζητούνται m σύνολα - ομάδες $C_1, C_2, ..., C_m$, με m<<n έτσι ώστε:

$C_i \neq \varnothing, \quad \forall i = 1, 2, 3, ..., m$ και οι m ομάδες να αποτελούν διαμέριση του συνόλου X.

Ο ορισμός[7] αυτός αναφέρεται στην αυστηρή συσταδοποίηση διότι κάθε διάνυσμα ανήκει σε μία και μόνο ομάδα. Εναλλακτικά μπορεί να οριστεί η ασαφής συσταδοποίηση. Κάνοντας χρήση των ασαφών συνόλων μπορούμε να ορίσουμε m συναρτήσεις συμμετοχής $u_j : X \to [0,1]$, για $j = 1, 2, ..., m$. Οι συναρτήσεις $u(\cdot)$ ποσοτικοποιούν την βεβαιότητα που έχουμε για το αν κάποιο διάνυσμα n ανήκει στην ομάδα j.

Στόχος της συσταδοποίησης είναι μια ομάδα διανυσμάτων να διαχωριστεί σε ένα σύνολο ανομοιογενών συνόλων μεταξύ τους με βάση ενός συχνά υποκειμενικά επιλεγμένου μέτρου ομοιότητας, τέτοιο ώστε τα διανύσματα που ανήκουν στο ίδιο σύνολο να παρουσιάζουν όσο το δυνατόν μεγαλύτερη ομοιότητα από την ομοιότητα διανυσμάτων που ανήκουν σε διαφορετικά σύνολα. Ο συνδυασμός της επιλογής του αλγορίθμου, του κριτηρίου ομοιότητας και των χαρακτηριστικών τους μπορεί να επιφέρει διαφορετικά αποτελέσματα. Ωστόσο δεν υπάρχει απολύτως κανένας τρόπος να καθοριστεί ποιό κριτήριο είναι καταλληλότερο. Κάθε κριτήριο έχει τη δική του χρησιμότητα ανάλογα με την περίπτωση, παρόλο που κάποια έχουν πιο ευρεία χρησιμότητα από κάποια άλλα. Κατάλληλος τρόπος είναι εκείνος που τελικά θα επιλεγεί από το άτομο που εκτελεί τη διαδικασία, έτσι ώστε το αποτέλεσμα της συσταδοποίησης να ταιριάζει στις ανάγκες του.

[7] https://el.wikipedia.org/wiki/%CE%A3%CF%85%CF%83%CF%84%CE%B1%CE%B4%CE%BF%CF%80%CE%BF%CE%AF%CE%B7%CF%83%CE%B7

2.3. Βήματα συσταδοποίησης

Η διαδικασία της συσταδοποίησης ακολουθεί τα παρακάτω βασικά βήματα[8] :

1) Επιλογή χαρακτηριστικών γνωρισμάτων. Ο στόχος είναι να επιλεγούν τα καταλληλότερα χαρακτηριστικά (features selection) στα οποία πρόκειται να εφαρμοστεί η συσταδοποίηση ώστε να επιτυγχάνεται η βέλτιστη ομοιογένεια σε κάθε συστάδα. Επιπλέον, η διαδικασία της εξαγωγής χαρακτηριστικών (features extraction) χρησιμοποιεί έναν ή περισσότερους μετασχηματισμούς των χαρακτηριστικών εισόδου, για να παραχθούν νέα, που πιθανόν να παρουσιάζουν μεγαλύτερο ενδιαφέρον. Έτσι η προεπεξεργασία των δεδομένων - διανυσμάτων πριν την εφαρμογή της διαδικασίας συσταδοποίησης κρίνεται απαραίτητη.

2) Επιλογή αλγόριθμου συσταδοποίησης. Σε αυτό το στάδιο γίνεται η επιλογή ενός αλγορίθμου (αναλύονται παρακάτω) που θα οδηγήσει σε ένα καλό σχήμα συσταδοποίησης για ένα σύνολο δεδομένων. Για την επιλογή του αλγορίθμου χρειάζεται το *μέτρο γειτνίασης* και το *κριτήριο συσταδοποίησης* τα οποία ορίζουν απόλυτα τον αλγόριθμο, καθώς επίσης και η δυνατότητά του να καθορίσει ένα σχήμα συσταδοποίησης που να προσαρμόζεται στο συγκεκριμένο σύνολο δεδομένων.

[8] http://apothesis.teicm.gr/xmlui/bitstream/handle/123456789/830/vrionis.pdf?sequence=1

Επομένως το βήμα αυτό χρειάζεται τα εξής:

i. Το *μέτρο γειτνίασης (proximity measure)* αναφέρεται στην ομοιότητα δύο αντικειμένων (διανύσματα χαρακτηριστικών). Η επιλογή των χαρακτηριστικών πρέπει να γίνεται προσεκτικά ώστε η συμβολή τους να είναι ίση κατά τον υπολογισμό του μέτρου γειτνίασης και να μην υπερισχύει το ένα έναντι του άλλου.

ii. Το *κριτήριο συσταδοποίησης (clustering criterion)* εκφράζεται βάσει μιας συνάρτησης κόστους ή κάποιου άλλου τύπου κανόνων. Είναι σημαντικό να γνωρίζουμε τον τύπο των συστάδων που θα προκύψουν, για να διαλέξουμε το κατάλληλο κριτήριο που θα έχει ως αποτέλεσμα μία επιτυχημένη τμηματοποίηση η οποία θα ταιριάζει στο σύνολο δεδομένων.

3) Εγκυρότητα αποτελεσμάτων συσταδοποίησης. Σε αυτή τη φάση αξιολογούνται τα αποτελέσματα του αλγορίθμου συσταδοποίησης σύμφωνα με κατάλληλα κριτήρια ορθότητας συσταδοποίησης και τεχνικές. Παράδειγμα ενός τέτοιου κριτηρίου είναι η σύγκριση των αποτελεσμάτων της ανάλυσης με κάποια ήδη γνωστά αποτελέσματα ή η σύγκριση των αποτελεσμάτων δύο διαφορετικών συσταδοποιήσεων. Η ποιότητα της συσταδοποίησης εξαρτάται από την ομοιότητα (δηλαδή μεγάλη ομοιότητα εντός της συστάδας - μικρή ομοιότητα μεταξύ των συστάδων) και την μέθοδο υλοποίησης της συσταδοποίησης.

4) ***Ερμηνεία των αποτελεσμάτων***. Αποτελεί το τελευταίο στάδιο της διαδικασίας συσταδοποίησης, όπου οι αναλυτές καλούνται να εξάγουν γνώση από τις παραχθείσες συστάδες, συνδυάζοντας κι άλλα στοιχεία και αναλύσεις, με σκοπό το καλύτερο και εγκυρότερο αποτέλεσμα. Πρακτικά εξετάζεται αν οι ομάδες είναι αντιπροσωπευτικές σε

σχέση με τα δεδομένα που έπρεπε να ομαδοποιηθούν, αν τα δεδομένα τοποθετήθηκαν στις κατάλληλες ομάδες κ.α.

2.4. Μέθοδοι συσταδοποίησης

Το πρόβλημα της συσταδοποίησης έχει μελετηθεί εκτεταμένα και για το λόγο αυτό έχει προταθεί ένας μεγάλος αριθμός αλγορίθμων συσταδοποίησης. Σύμφωνα με τη μέθοδο που υιοθετείται για τον καθορισμό των συστάδων, οι αλγόριθμοι ταξινομούνται στους παρακάτω τύπους:

1. *Οι διαχωριστικές μέθοδοι (partitioning methods)* στις οποίες ένα σύνολο από m δεδομένα διαχωρίζεται σε k συστάδες, με k<<m και έπειτα βελτιστοποιούν το αποτέλεσμα. Το γενικό κριτήριο είναι η ελαχιστοποίηση κάποιων μέτρων ανομοιότητας μεταξύ των δειγμάτων μέσα σε κάθε μία από τις συστάδες, καθώς και η μεγιστοποίηση την ανομοιότητας μεταξύ των διαφορετικών συστάδων. Παράδειγμα αποτελεί ο αλγόριθμος *k - means* στον οποίο ο αριθμός των συστάδων δίνεται στην αρχή από των χρήστη και κάθε συστάδα έχει το δικό της κεντροειδές (centroid).

2. *Οι ιεραρχικές μέθοδοι (hierarchical methods)* στις οποίες δημιουργείται ένα δενδρόγραμμα με βάση τη διαδοχική σύνδεση μικρότερων συστάδων σε μεγαλύτερες ή τη διάσπαση μεγαλύτερων συστάδων σε μικρότερες. Μια τέτοια διάσπαση μπορεί να δημιουργηθεί με *bottom - up* ή με *top - down* τρόπο, δημιουργώντας συσσωρευτική ιεραρχική μέθοδο (*agglomerative hierarchical methods)* ή διαχωριστική ιεραρχική μέθοδο (*divisive hierarchical methods)*, αντίστοιχα. Και στις δύο περιπτώσεις χρειάζεται να έχει οριστεί μια συνάρτηση απόστασης μεταξύ των συστάδων. Συναρτήσεις

απόστασης είναι η single linkage, η complete linkage, η average linkage, η απόσταση των κεντροειδών (centroids distance) και η ward, τις οποίες θα δούμε αναλυτικά παρακάτω.

3. Οι μέθοδοι με βάση την πυκνότητα (density based methods) στις οποίες μια συστάδα είναι μια πυκνή περιοχή από σημεία η οποία διαχωρίζεται από άλλες περιοχές μεγάλης πυκνότητας με περιοχές χαμηλής πυκνότητας. Για κάθε παρατήρηση που ανήκει σε μια συστάδα, η γειτονιά της, η οποία είναι καθορισμένης διαμέτρου, πρέπει να περιλαμβάνει έναν ελάχιστο αριθμό παρατηρήσεων. Η συστάδα συνεχίζει να επεκτείνεται όσο η γειτονιά της διαθέτει την απαιτούμενη πυκνότητα. Οι μέθοδοι αυτές μπορούν να δημιουργήσουν συστάδες με μη κυρτά και περίπλοκα σχήματα. Επιπλέον έχουν την δυνατότητα να απομονώνουν τις εξαιρέσεις.

4. *Οι μέθοδοι βασιζόμενες σε πλέγμα (grid based methods)* οι οποίοι χρησιμοποιούνται κυρίως για την ανάλυση χωρικών δεδομένων. Το βασικό χαρακτηριστικό τους είναι ότι χωρίζουν το χώρο σε ένα πεπερασμένο αριθμό κελιών και στη συνέχεια κάνουν όλες τις διαδικασίες στον κβαντοποιημένο χώρο. Η αναζήτηση των συστάδων γίνεται στα κελιά του πλέγματος και όχι στις παρατηρήσεις. Επειδή ο αριθμός των κελιών είναι πολύ μικρότερος από τον αριθμό των παρατηρήσεων, οι μέθοδοι αυτές είναι σημαντικά ταχύτερες. Ένα σημαντικό ζήτημα είναι ο καθορισμός κελιών κατάλληλου μεγέθους.

2.5. Σημασία αριθμητικής απόστασης των δεδομένων

Όπως έχουμε ήδη αναφέρει, η μέθοδος της ανάλυσης συστάδων βασίζεται στην μέγιστη ομοιογένεια μεταξύ των παρατηρήσεων της ίδιας κλάσης, ή την μέγιστη ανομοιογένεια μεταξύ παρατηρήσεων διαφορετικών κλάσεων. Η ομοιογένεια (ή ανομοιογένεια), καθορίζεται από τις αποστάσεις μεταξύ των παρατηρήσεων, οι οποίες μπορεί να είναι

μονοδιάστατες ή πολυδιάστατες, με την κάθε διάσταση να αντιπροσωπεύει μία μεταβλητή. Υπάρχουν αρκετές μετρικές απόστασης, δίνοντας στον ερευνητή το περιθώριο να επιλέξει την κατάλληλη βάσει των δεδομένων του και του σκοπού της έρευνάς του.

2.5.1. Ορισμός μετρικής απόστασης

Έστω ότι ο χώρος είναι ένα σύνολο σημείων. Μια μετρική απόστασης σε αυτό το χώρο είναι μια συνάρτηση $d(x,y)$ που λαμβάνει δυο σημεία για είσοδο (x,y) και παράγει έναν πραγματικό αριθμό ικανοποιώντας τα εξής αξιώματα:

1. $d(x,y) \geq 0$ (δεν υπάρχουν αρνητικές αποστάσεις).
2. $d(x,y) = 0$ αν και μόνο αν $x = y$ (οι αποστάσεις είναι θετικές εκτός της απόστασης ενός σημείου από τον εαυτό του).
3. $d(x,y) = d(y,x)$ (η απόσταση είναι συμμετρική)
4. $d(x,y) \leq d(x,z) + d(z,y)$ (η τριγωνική ανισότητα)

Η τριγωνική ανισότητα είναι η πολυπλοκότερη συνθήκη. Διαισθητικά αναφέρει ότι η απόσταση από το x και y δεν είναι μικρότερη από την απόσταση μεταξύ του x και y αν αναγκασθούμε να περάσουμε από οποιοδήποτε ενδιάμεσο σημείο z. Το αξίωμα της τριγωνικής ανισότητας είναι αυτό που υπαγορεύει σε όλες τις μετρικές απόστασης να συμπεριφέρονται σαν η απόσταση να περιγράφει το μήκος της ελάχιστης διαδρομής από το ένα σημείο στο άλλο.

2.5.2. Απόσταση με αριθμητικά γνωρίσματα

Παραδείγματα που μπορούν να χρησιμοποιηθούν για τον υπολογισμό της απόστασης μεταξύ δύο διανυσμάτων $x_i = (x_{i1},...,x_{ir})$ και $x_j = (x_{j1},...,x_{jr})$, με αριθμητικά γνωρίσματα.

- Η **απόσταση Minkowski** είναι:

$$dist(x_i, x_j) = \left(|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + ... + |x_{ir} - x_{jr}|^h \right)^{\frac{1}{h}}$$

όπου h είναι θετικός ακέραιος και x_i, x_j διανύσματα χαρακτηριστικών ή δεδομένων .

Οι δύο παρακάτω μετρικές, *Euclidean* και *Manhattan,* είναι ειδικές περιπτώσεις της μετρικής *Minkowski*, για *h=2* και h=*1* αντίστοιχα.

- Αν $h = 2$, αυτή είναι η **Euclidean απόσταση:** κατάλληλη όταν τα δεδομένα σχηματίζουν απομονωμένες ομάδες.

$$dist(x_i, x_j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + ... + (x_{ir} - x_{jr})^2}.$$

- Αν $h = 1$, αυτή είναι η **Manhattan απόσταση:** μοιάζει πολύ με την ευκλείδεια απόσταση με τη διαφορά ότι αντί για τετραγωνικές αποκλίσεις χρησιμοποιούμε απόλυτες αποκλίσεις. Συνήθως λόγω της ομοιότητάς της με την ευκλείδεια απόσταση δίνει περίπου ίδια αποτελέσματα εκτός από την περίπτωση που υπάρχουν outliers, όπου επειδή τους δίνει μικρότερο βάρος (εξαιτίας της απόλυτης τιμής), μπορεί να οδηγήσει σε πιο ανθεκτικά αποτελέσματα.

$$dist\left(\mathrm{x}_i, \mathrm{x}_j\right) = \left|\mathrm{x}_{i1} - \mathrm{x}_{j1}\right| + \left|\mathrm{x}_{i2} - \mathrm{x}_{j2}\right| + \ldots + \left|\mathrm{x}_{ir} - \mathrm{x}_{jr}\right|.$$

Ονομάζεται Manhattan καθώς προσομοιάζει την απόσταση μεταξύ δύο σημείων στην περιοχή Manhattan της Νέας Υόρκης.

Η Ευκλείδεια απόσταση δεν επηρεάζεται από την προσθήκη νέων παρατηρήσεων. Ένα μειονέκτημα της Ευκλείδειας απόστασης είναι ότι η μεταβολή των μονάδων μέτρησης μιας μεταβλητής (π.χ μετατροπή μιας απόστασης από χιλιόμετρα σε μέτρα) επηρεάζει σημαντικά την απόσταση, και μπορεί να οδηγήσει σε αρκετά διαφορετικές συστάδες. Επίσης, οι μεταβλητές, οι οποίες παίρνουν μεγαλύτερες τιμές ή που παρουσιάζουν μεγάλες διαφορές τιμών μεταξύ των παρατηρήσεων, επηρεάζουν την απόσταση. Ένας τρόπος αντιμετώπισης αυτού του προβλήματος είναι η κανονικοποίηση των τιμών.
Στον υπολογισμό της Ευκλείδειας απόστασης όλα τα χαρακτηριστικά θεωρούνται ισότιμα. Στην περίπτωση που ο αναλυτής θέλει να δώσει βαρύτητα σε κάποια από αυτά, τους εκχωρεί συντελεστές βαρύτητας. Με τη χρήση συντελεστών βαρύτητας, η Ευκλείδεια απόσταση υπολογίζεται σύμφωνα με την Weighted Euclidean απόσταση.

- **Weighted Euclidean απόσταση**: Ένα βάρος συσχετίζεται με κάθε χαρακτηριστικό για να εκφράσει τη σημασία του σε σχέση με τα άλλα χαρακτηριστικά.

Όπου w_r ο συντελεστής βαρύτητας του γνωρίσματος r.

Άλλοι τύποι αποστάσεων είναι:

- **Squared Euclidean απόσταση**: η τυπική Ευκλείδεια απόσταση είναι τετραγωνική , προκειμένου να δίνεται σταδιακά μεγαλύτερη βαρύτητα σε ομάδες που είναι πιο απομακρυσμένες.

$$dist\left(x_i, x_j\right) = \left(x_{i1} - x_{j1}\right)^2 + \left(x_{i2} - x_{j2}\right)^2 + \ldots + \left(x_{ir} - x_{jr}\right)^2.$$

- **Απόσταση Chebychev:** Η απόσταση μέτρου είναι κατάλληλη σε περιπτώσεις όπου κάποιος θέλει να ορίσει δύο σημεία δεδομένων ως «διαφορετικά», αν είναι διαφορετικά για κάθε ένα από τα χαρακτηριστικά. Ισούται με την μέγιστη απόλυτη διαφορά των τιμών των μεταβλητών.

$$dist\left(x_i, x_j\right) = \max\left(\left|x_{i1} - x_{j1}\right|, \left|x_{i2} - x_{j2}\right|, \ldots, \left|x_{ir} - x_{jr}\right|\right).$$

Η απόσταση μπορεί να χρησιμοποιηθεί για να κατασκευαστεί πίνακας αποστάσεων σε κάθε στάδιο ανάλυσης. Ο πίνακας αυτός θα έχει μηδενικά στοιχεία στη διαγώνιο και την απόσταση μεταξύ του *i* στοιχείου (ή συστάδας) και του *j* στοιχείου (ή συστάδας) στη θέση (*i,j*).

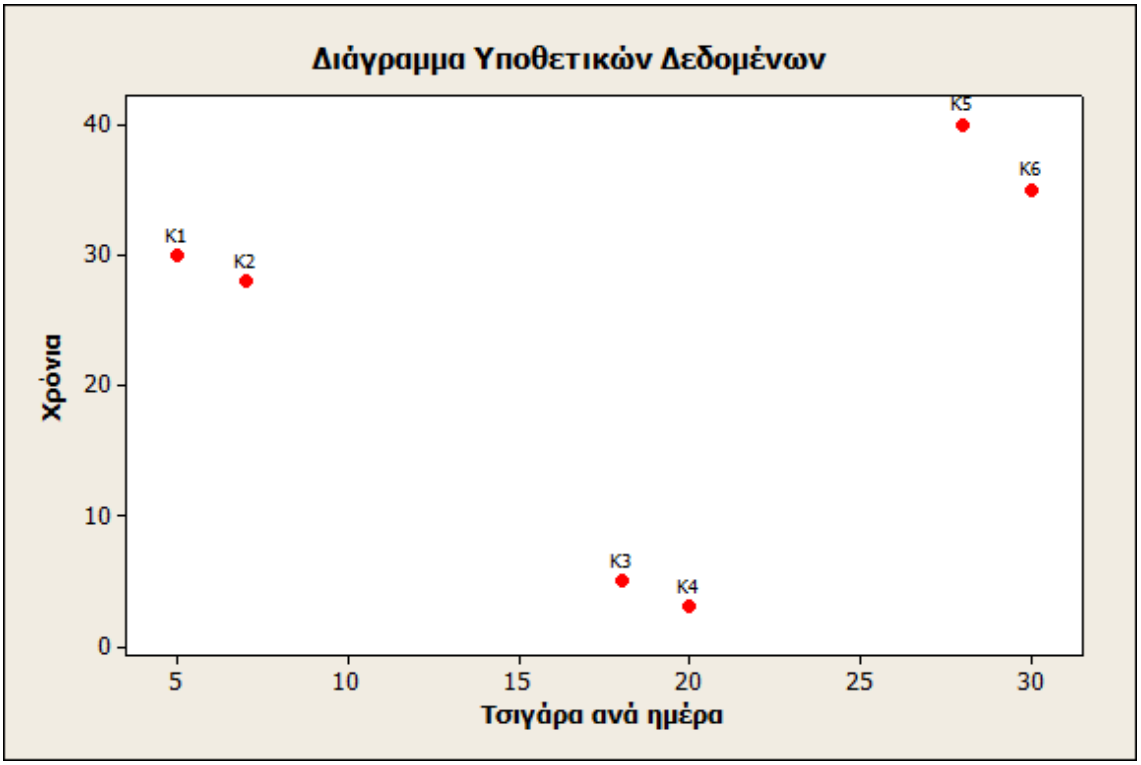
2.6. Γεωμετρική ερμηνεία της συσταδοποίησης

Η γεωμετρική έννοια τις ανάλυσης συστάδων είναι πολύ απλή στην κατανόησή της. Θεωρούμε τα υποθετικά δεδομένα του παρακάτω πίνακα, που περιέχει τα τσιγάρα ανά ημέρα και τα χρόνια που καπνίζουν 6 υποθετικοί καπνιστές.

Πίνακας με τα υποθετικά δεδομένα:

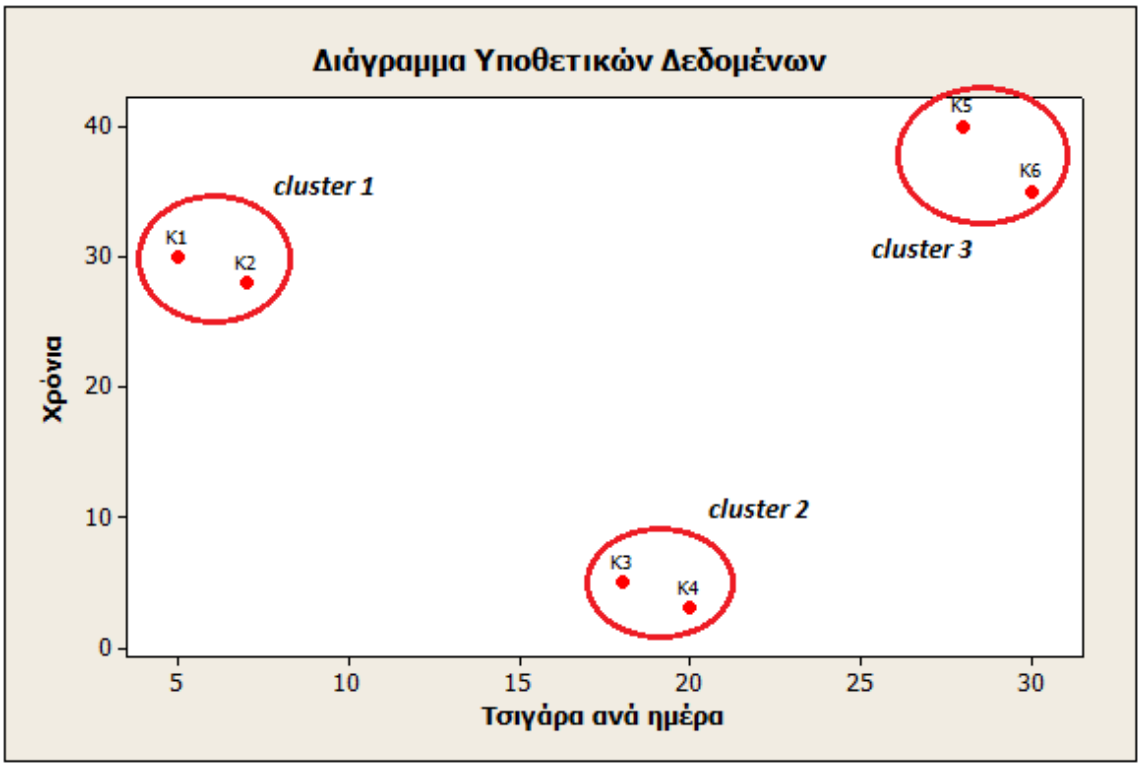
Καπνιστής	***Τσιγάρα ανά ημέρα***	***Χρόνια***
Κ1	5	30
Κ2	7	28
Κ3	18	5
Κ4	20	3
Κ5	28	40
Κ6	30	35

Όπως φαίνεται και στο παρακάτω διάγραμμα, κάθε παρατήρηση μπορεί να αναπαρασταθεί σαν ένα σημείο σ' ένα δισδιάστατο χώρο, αφού δυο είναι οι μεταβλητές που χρησιμοποιήσαμε (τσιγάρα ανά ημέρα, χρόνια). Γενικά κάθε παρατήρηση μπορεί να αναπαρασταθεί σ' ένα n-διάστατο χώρο, όπου n είναι ο αριθμός των μεταβλητών που χρησιμοποιούνται για να περιγράψουν τις παρατηρήσεις - διανύσματα.



Σχήμα 2.2.α: *Διάγραμμα υποθετικών δεδομένων*

Υποθέτουμε τώρα πως θέλουμε να σχηματίσουμε 3 συστάδες (cluster). Μια γρήγορα ματιά του πίνακα φανερώνει, ότι οι παρατηρήσεις Κ1 και Κ2 θα σχηματίσουν μια συστάδα, οι παρατηρήσεις Κ3 και Κ4 θα σχηματίσουν μια δεύτερη συστάδα και οι παρατηρήσεις Κ5 και Κ6 θα σχηματίσουν την τρίτη συστάδα, δηλαδή παρατηρήσεις με μικρή απόσταση ανήκουν στην ίδια ομάδα και παρατηρήσεις με μεγάλη απόσταση ανήκουν σε διαφορετικές ομάδες.



Σχήμα 2.2.β: *παρατήρηση των 3ων συστάδων.*

Όπως παρατηρούμε, η ανάλυση συστάδων, ομαδοποιεί παρατηρήσεις, έτσι ώστε οι παρατηρήσεις σε κάθε ομάδα, να είναι όμοιες σε σχέση με τις μεταβλητές ομαδοποίησης.

Κεφάλαιο 3ο: Διαχωριστικές μέθοδοι συσταδοποίησης

3.1. Εισαγωγή

Στην διαχωριστική ομαδοποίηση, τα δεδομένα διαιρούνται σε κ διαμερίσεις, με κάθε διαμέριση να αντιπροσωπεύει μια ομάδα ή διαφορετικά ένα cluster. Αντίθετα με την

ιεραρχική ομαδοποίηση, ο αριθμός κ των clusters, μπορεί να καθοριστεί από την αρχή της διαδικασίας. Επειδή δεν είναι απαραίτητο να καθοριστεί ένας πίνακας αποστάσεων (ομοιοτήτων) μεταξύ των παρατηρήσεων που θέλουμε να ομαδοποιήσουμε και τα βασικά δεδομένα δεν χρειάζεται να αποθηκευτούν στον υπολογιστή κατά το τρέξιμο του αλγορίθμου, οι διαχωριστικές μέθοδοι μπορούν να εφαρμοστούν σε πολύ μεγάλο όγκο δεδομένων, σε σχέση με τις ιεραρχικές μεθόδους.

Οι περισσότεροι από τους μη ιεραρχικούς αλγορίθμους, διαφέρουν σε σχέση με:

i. Τη μέθοδο που χρησιμοποιήθηκε για την απόκτηση των αρχικών κέντρων (centroids) των clusters και
ii. τον κανόνα που χρησιμοποιήθηκε για την ανακατανομή των παρατηρήσεων.

Κάποιες από τις μεθόδους που χρησιμοποιούνται, για να πάρουμε τα αρχικά centroids είναι:

1. Επιλέγουμε τις k πρώτες παρατηρήσεις με μη ελλιπή δεδομένα, σαν κέντρα για τα αρχικά clusters.

2. Επιλέγουμε τυχαία, k μη ελλιπείς παρατηρήσεις σαν κέντρα για τα αρχικά clusters.
3. Επιλέγουμε την πρώτη μη ελλιπή παρατήρηση, σαν κεντροειδές για το πρώτο cluster. Η επιλογή του κεντροειδούς για το δεύτερο cluster γίνεται, έτσι ώστε η απόστασή του από το προηγούμενο κεντροειδές, να είναι μεγαλύτερη από μια καθορισμένη απόσταση που έχουμε ορίσει. Το τρίτο κεντροειδές επιλέγεται, έτσι ώστε η απόστασή του από τα προηγούμενα κεντροειδή που επιλέχθηκαν, να είναι μεγαλύτερη από την προκαθορισμένη απόσταση και συνεχίζουμε κατά αυτό τον τρόπο για την επιλογή των υπόλοιπων κεντροειδών.

Μόλις προσδιοριστούν τα centroids, σχηματίζονται τα αρχικά cluster. Στη συνέχεια αναθέτουμε κάθε μια από τις υπόλοιπες n-k παρατηρήσεις, στο cluster εκείνο, στο οποίο η παρατήρηση είναι πιο κοντά στο κεντροειδές του.

Όπως αναφέρθηκε και παραπάνω, οι διαχωριστικοί αλγόριθμοι διαφέρουν επίσης, στη διαδικασία που χρησιμοποιούν για την ανακατανομή των παρατηρήσεων σε k cluster. Κάποιοι από τους κανόνες ανακατανομής, είναι:

1. Υπολογίζουμε το νέο κεντροειδές του κάθε cluster ξεχωριστά και αναθέτουμε εκ νέου τις παρατηρήσεις, σ' εκείνο το cluster με το πλησιέστερο centroid. Μέχρι να ολοκληρωθεί η εκ νέου ανάθεση όλων των παρατηρήσεων στα clusters, τα κεντροειδή τους δεν αλλάζουν. Όταν όμως πραγματοποιηθεί η ανακατανομή των παρατηρήσεων, τα κεντροειδή υπολογίζονται ξανά. Αν η μεταβολή στα centroids των clusters, είναι μεγαλύτερη από ένα κριτήριο σύγκλισης που ορίστηκε, τα centroid επαναπροσδιορίζονται. Η διαδικασία ανακατανομής συνεχίζεται, μέχρι η μεταβολή των centroids να είναι μικρότερη, από το καθορισμένο κριτήριο σύγκλισης.

2. Υπολογίζουμε το κεντροειδές του κάθε cluster ξεχωριστά και αναθέτουμε εκ νέου τις παρατηρήσεις, σ' εκείνο το cluster με το πλησιέστερο centroid. Για την ανάθεση κάθε παρατήρησης, υπολογίζουμε εκ νέου το centroid του cluster στο οποίο θα τοποθετηθεί η παρατήρηση και του cluster από το οποίο αποδίδεται η παρατήρηση. Η εκ νέου ανάθεση συνεχίζεται ξανά, μέχρι η μεταβολή στα centroid των cluster, να γίνει μικρότερη από το καθορισμένο κριτήριο σύγκλισης.

3.2. Αλγόριθμος k - Means

Ο διαχωριστικός αλγόριθμος k-means είναι ένας από τους πιο απλούς και γνωστούς αλγόριθμους ομαδοποίησης που ανήκουν στην ευρύτερη κατηγορία των τεχνικών μάθησης χωρίς επίβλεψη. Ο όρος K-means χρησιμοποιήθηκε για πρώτη φορά από τον James MacQueen το 1967, για να περιγράψει τον αλγόριθμό του, ο οποίος ταξινομεί σε **k** προκαθορισμένες ομάδες (clusters), ένα σύνολο δεδομένων. Το βασικό πρόβλημα είναι ο αριθμός των ομάδων που θα δημιουργηθούν αρχικά, γιατί αυτός θα είναι και ο τελικός αριθμός ομάδων που θα εξάγει ο αλγόριθμος.

Ο αλγόριθμος αποτελείται από τα ακόλουθα *βήματα*:

- *Βήμα 1ο*: Επιλέγω k δεδομένα τα οποία αντιπροσωπεύουν τα κεντροειδή (centroids) των k ομάδων. Αυτά τα κεντροειδή πρέπει να επιλεγούν προσεκτικά, γιατί διαφορετικά αρχικά centroids δίνουν διαφορετικό αποτέλεσμα. Έτσι συχνά η καλύτερη επιλογή θεωρείται εκείνων που απέχουν όσο το δυνατόν πιο πολύ.
- *Βήμα 2ο*: Υπολογίζω τις αποστάσεις όλων των υπόλοιπων δεδομένων από τα k επιλεγμένα centroids.
- *Βήμα 3ο*: Τοποθετώ τα δεδομένα στην ομάδα με της οποίας το κεντροειδές είναι πιο κοντά.
- *Βήμα 4ο*: Υπολογίζω τα κέντρα των k νέων ομάδων με χρήση του μέσου όρου των σημείων τους.
- *Βήμα 5ο*: Πηγαίνω στο βήμα 2 και αν δε χρειάζεται να μετακινηθούν δεδομένα σε άλλες ομάδες σταματώ και το αποτέλεσμα είναι η ομαδοποίηση του συνόλου δεδομένων σε k clusters, αλλιώς πάω στο βήμα 3.

Αν και μπορεί να αποδειχθεί ότι ο αλγόριθμος πάντα τερματίζει, θα πρέπει να τονίσουμε ότι δεν καταφέρνει πάντα να βρίσκει τη βέλτιστη λύση. Ο αλγόριθμος επηρεάζεται σημαντικά από τα αρχικά centroids. Για το λόγο αυτό προτείνεται η εκτέλεσή του πολλές φορές, μέχρι να μειωθεί η επίδραση αυτή.

3.3. Αδυναμίες αλγορίθμου k-means

Παρά την ευρεία χρήση αυτού του αλγορίθμου, για ομαδοποίηση δεδομένων, παρουσιάζει και κάποιες αδυναμίες, όπως:

- ✓ Ο τρόπος με τον οποίο ορίζονται τα αρχικά centroids δεν είναι απόλυτα καθορισμένος. Ο πιο συνηθισμένος τρόπος, είναι η τυχαία τους επιλογή. Αυτός ο τρόπος εφαρμόζεται και στην παρούσα εργασία.

- ✓ Υπάρχει περίπτωση ένα cluster να μείνει χωρίς δεδομένα και έτσι να μην ανανεωθεί κάποιο centroid.

- ✓ Τα αποτελέσματα εξαρτώνται από το μέτρο απόστασης που θα χρησιμοποιηθεί.

- ✓ Ο αλγόριθμος δυσκολεύεται να αναγνωρίσει ομάδες με διαφορετικό σχηματισμό και μέγεθος. Το πρόβλημα αυτό μεγιστοποιείται κυρίως σε πολύ μεγάλα σύνολα δεδομένων.

- ✓ Το αποτέλεσμα εξαρτάται από τον καθορισμό του αρχικού πλήθους των ομάδων που θα δημιουργηθούν και τελικά. Ο αλγόριθμος δεν καταφέρνει να βρεί το βέλτιστο πλήθος από μόνος του, αλλά δίνεται αρχικά από το χρήστη.

3.4. Παράδειγμα υλοποίησης αλγορίθμου k - Means

Ζητείται να ομαδοποιηθούν τα στοιχεία Α(1,1), Β(2,1), Γ(4,3), Δ(5,4) σε δύο clusters. [9]

- ***Βήμα 1ο***:

Γίνεται τυχαία επιλογή των σημείων Α(1,1) και Β(2,1), που θα αποτελούν τα αρχικά κέντρα των clusters.

- ***Βήμα 2ο***:

Υπολογίζουμε τις αποστάσεις των υπόλοιπων σημείων από τα επιλεγμένα κέντρα και τα τοποθετούμε στο cluster εκείνο με τη μικρότερη απόσταση μεταξύ τους.

Υπολογισμός αποστάσεων με τετραγωνική ευκλείδεια απόσταση:

$$d^2\left((4,3),(1,1)\right)=(4-1)^2+(3-1)^2=9+4=\mathbf{13}$$

$$d^2\left((5,4),(1,1)\right)=(5-1)^2+(4-1)^2=16+9=25$$

$$d^2\left((4,3),(2,1)\right)=(4-2)^2+(3-1)^2=4+4=\mathbf{8}$$

$$d^2\left((5,4),(2,1)\right)=(5-2)^2+(4-1)^2=9+9=18$$

Μετά τη μέτρηση των αποστάσεων, προκύπτουν τα cluster:

Cluster 1 $\rightarrow A(1,1)$, *Cluster 2* $\rightarrow B(2,1),\ \Gamma(4,3),\ \Delta(5,4)$

όπως φαίνονται και στον παρακάτω πίνακα:

cluster	Α(1,1)	Β(2,1)	Γ(4,3)	Δ(5,4)
cluster 1	✓			
cluster 2		✓	✓	✓

[9] *http://www.teilar.gr/dbData/ProfAnn/profann-668f4316.pdf*

- ***Βήμα 3ο*:**

Υπολογίζουμε τις καινούριες συντεταγμένες των κέντρων των clusters βάσει των νέων δεδομένων:

$$d^2\left((2,1),(1,1)\right)=(2-1)^2+(1-1)^2=\mathbf{1}$$

$$d^2\left((4,3),(1,1)\right)=(4-1)^2+(3-1)^2=\mathbf{13}$$

$$d^2\left((5,4),(1,1)\right)=(5-1)^2+(4-1)^2=\mathbf{25}$$

$$d^2\left((2,1),\left(\frac{11}{3},\frac{8}{3}\right)\right)=\left(2-\frac{11}{3}\right)^2+\left(1-\frac{8}{3}\right)^2=\mathbf{5.56}$$

$$d^2\left((4,3),\left(\frac{11}{3},\frac{8}{3}\right)\right)=\left(4-\frac{11}{3}\right)^2+\left(3-\frac{8}{3}\right)^2=\mathbf{0.23}$$

$$d^2\left((5,4),\left(\frac{11}{3},\frac{8}{3}\right)\right)=\left(5-\frac{11}{3}\right)^2+\left(4-\frac{8}{3}\right)^2=\mathbf{3.56}$$

Μετά τη μέτρηση των αποστάσεων, προκύπτουν τα cluster:

Κέντρο Cluster 1: $(1,1)$

Κέντρο Cluster 2: $\left(\left(\frac{2+4+5}{3}\right),\left(\frac{1+3+4}{3}\right)\right)=\left(\frac{11}{3},\frac{8}{3}\right)$

- ***Βήμα 4ο*:**

Υπολογίζουμε τις αποστάσεις όλων των σημείων από τα νέα κέντρα όπως και στα παραπάνω βήματα και τα τοποθετούμε στο cluster εκείνο με του οποίου το κέντρο έχει τη μικρότερη απόσταση.

Παίρνουμε λοιπόν τα παρακάτω αποτελέσματα:

cluster	**A(1,1)**	**B(2,1)**	**Γ(4,3)**	**Δ(5,4)**
cluster 1	✓	✓		
cluster 2			✓	✓

- ***Βήμα 5ο:***

Υπολογίζουμε τις καινούριες συντεταγμένες των κέντρων των clusters, βάσει των νέων δεδομένων.

Κέντρο Cluster 1: $\left(\left(\frac{1+2}{2}\right),\left(\frac{1+1}{2}\right)\right)=(1.5,1)$

Κέντρο Cluster 2: $\left(\left(\frac{4+5}{2}\right),\left(\frac{3+4}{2}\right)\right)=(4.5,3.5)$

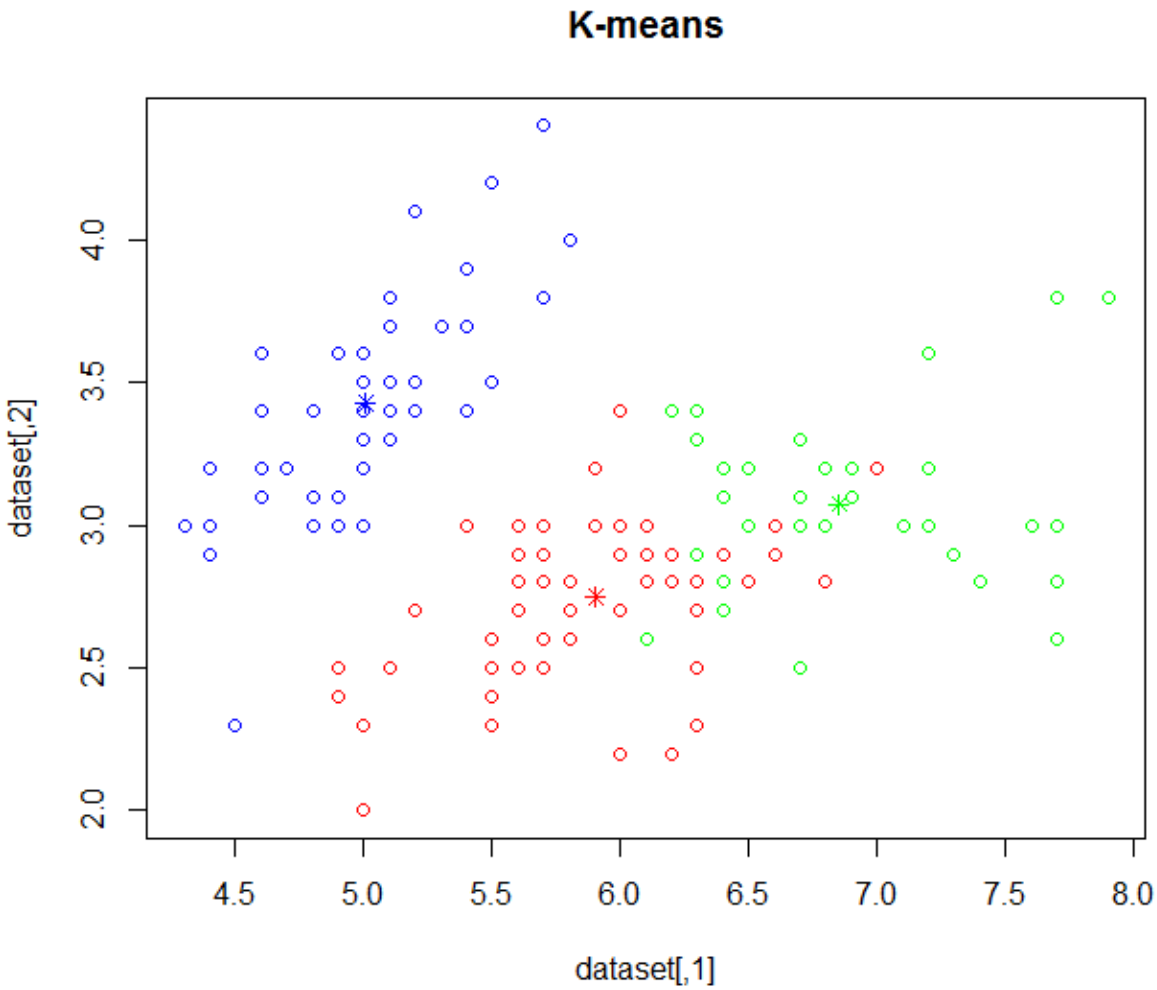
στη συνέχεια υπολογίζουμε τις αποστάσεις όλων των σημείων από τα νέα κέντρα και επιλέγουμε την ομάδα στην οποία ανήκει το κάθε στοιχείο και παίρνουμε τα αποτελέσματα του παρακάτω πίνακα:

cluster	**Α(1,1)**	**Β(2,1)**	**Γ(4,3)**	**Δ(5,4)**
cluster 1	✓	✓		
cluster 2			✓	✓

Παρατηρούμε ότι κανένα στοιχείο δεν άλλαξε cluster, άρα ο αλγόριθμος τερματίζει με τα παρακάτω τελικά αποτελέσματα:

ΣΤΟΙΧΕΙΑ	**Τελικά cluster**
Α(1,1)	**cluster 1**
Β(2,1)	**cluster 1**
Γ(4,3)	**cluster 2**

3.5. Παράδειγμα k - Means στην R



Σχήμα 3.1: *παρατήρηση των 3ων συστάδων και των κεντροειδών τους.*

Κώδικας στην R

```
cl <- kmeans(dataset, 3, algorithm="MacQueen")
plot(dataset, col = mycol[cl$cluster], main="K-means")
points(cl$centers, col = mycol, pch = 8)
```

Κεφάλαιο 4ο: Ιεραρχικές Μέθοδοι συσταδοποίησης

4.1. Εισαγωγή

Βασικό χαρακτηριστικό των ιεραρχικών μεθόδων και η κύρια διαφορά από τη μέθοδο k-means είναι ότι ο αριθμός των ομάδων δεν είναι γνωστός από την αρχή της μελέτης.
Στην ιεραρχική ομαδοποίηση τα στιγμιότυπα ομαδοποίησης δίνονται σε μορφή δενδρογράμματος. Στα δενδρογράμματα αυτά, επιλέγεται ένα επίπεδο που θα κλαδευτούν. Το σημείο που θα κλαδευτεί κάποιο δενδρόγραμμα δείχνει τον αριθμό των ομάδων που θα προκύψουν καθώς τα σημεία που περιέχει η κάθε ομάδα. Στις ιεραρχικές τεχνικές ομαδοποίησης ξεκινάμε είτε με διαδοχικές συγχωνεύσεις, είτε με διαδοχικές διαιρέσεις, γι' αυτό τις χωρίζουμε σε δύο κατηγορίες αντίστοιχα: την συσσωρευτική ιεραρχική συσταδοποίηση και την διαιρετική ιεραρχική συσταδοποίηση.

Στην ***συσσωρευτική ιεραρχική συσταδοποίηση (agglomerative hierarchical clustering)*** γίνεται μια "bottom up" προσέγγιση που είναι και η πιο διαδεδομένη. Σε αυτή την περίπτωση χρησιμοποιείται ο αλγόριθμος που ξεκινά με κάθε μια από τις n παρατηρήσεις να θεωρείται σαν μια ξεχωριστή συστάδα (cluster). Οι παρατηρήσεις με τη μεγαλύτερη ομοιότητα ή διαφορετικά με τη μικρότερη ανομοιότητα δημιουργούν μια νέα συστάδα και τελικά καθώς μειώνεται η ομοιότητα μετά από διαδοχικές συγχωνεύσεις, με κάποιο κριτήριο ομοιότητας, ο αλγόριθμος καταλήγει σε μια μοναδική συστάδα, η οποία εμπεριέχει όλες τις n παρατηρήσεις. Η διαδικασία του αλγορίθμου μπορεί να αναπαρασταθεί με ένα δενδρόγραμμα που περιέχει n-1 επίπεδα και το καθένα αντιστοιχεί σε ένα βήμα του αλγόριθμου.

2. Στην ***διαιρετική ιεραρχική συσταδοποίηση (divisive hierarchical clustering)*** γίνεται μια "top down" προσέγγιση που είναι αντίστροφη ουσιαστικά από την συσσωρευτική και στην οποία ο αλγόριθμος δουλεύει αντίστροφα, δηλαδή ξεκινά με όλες τις παρατηρήσεις σε μια ενιαία ομάδα και η παρατήρηση που βρίσκεται πιο μακριά από τις υπόλοιπες αποχωρεί από την ομάδα και σχηματίζει μόνη της μια νέα, δημιουργώντας έτσι δυο υποομάδες. Η διάσπαση πραγματοποιείται με τέτοιο τρόπο, ώστε οι υποομάδες οι οποίες θα προκύψουν να έχουν τη μεγαλύτερη ανομοιότητα. Στη συνέχεια περνάει στη δεύτερη πιο μακριά παρατήρηση και την βγάζει είτε για να αποτελέσει μια νέα ομάδα μόνη της είτε για να προστεθεί στην πρώτη ομάδα και κατά αυτό τον τρόπο συνεχίζει μέχρι να μετακινηθούν όλες οι παρατηρήσεις της μελέτης. Η διαδικασία συνεχίζεται, μέχρι να δημιουργηθούν τόσες υποομάδες, όσες και οι αρχικές παρατηρήσεις. Η πολυπλοκότητα των διαιρετικών αλγορίθμων είναι μεγαλύτερη από αυτή των συσσωρευτικών, αφού η διάσπαση μιας συστάδας σε δυο μπορεί να γίνει με $2^{n-1}-1$ τρόπους. Η επιλογή της βέλτιστης διάσπασης είναι πρακτικά αδύνατη ακόμα για για μικρό n. Στην πράξη η διάσπαση γίνεται, αλλά όχι κατά το βέλτιστο τρόπο. Η όλη διαδικασία του αλγορίθμου μπορεί να αναπαρασταθεί, όπως και στους συσσωρευτικούς, με δενδρόγραμμα.

Δεδομένου ότι και στις δύο μεθόδους δημιουργούνται 1 έως n ομάδες, όπου n ο αριθμός των παρατηρήσεων, σκοπός των μεθόδων είναι να βρεθεί ο κατάλληλος αριθμός ομάδων στο διάστημα αυτό.

Το πρώτο βήμα στις ιεραρχικές μεθόδους είναι να σχηματιστεί ένας πίνακας αποστάσεων για όλα τα ζεύγη παρατηρήσεων, με τη βοήθεια κάποιου κριτηρίου ομοιότητας.

Εάν οι παρατηρήσεις που εξετάζονται είναι n, τότε δημιουργείται ένας πίνακας nxn

διαστάσεων. Κάθε εγγραφή του πίνακα είναι ένα μέτρο ανομοιότητας ή απόστασης μεταξύ των παρατηρήσεων. Ο πίνακας αποστάσεων έχει την ακόλουθη μορφή:

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \dots & \dots & \dots & 0 & \\ d(n,1) & \dots & \dots & d(n,\mathrm{n}-1) & 0 \end{bmatrix}$$

Όπου $d(x_1, x_2)$ είναι η απόσταση μεταξύ των x_1 και x_2. Εφόσον η απόσταση κάθε σημείου από τον εαυτό του είναι μηδενική $(d(x_1, x_2) = 0,\ i = 1,2,...,n)$ οι εγγραφές της διαγωνίου από επάνω και αριστερά προς κάτω και δεξιά έχουν μηδενικές τιμές.

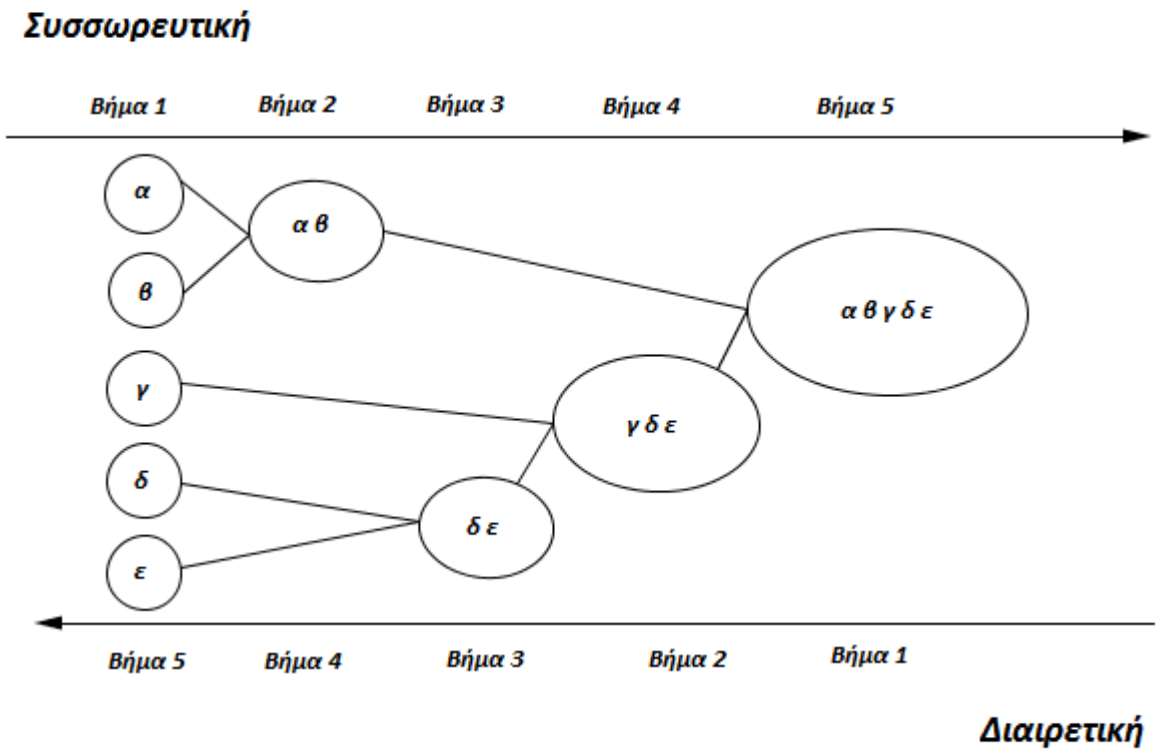
Ανάλογα με τον αλγόριθμο ομαδοποίησης σχηματίζονται κάποιες ομάδες παρατηρήσεων. Στη συνέχεια δημιουργείται ένας νέος πίνακας αποστάσεων και η διαδικασία επαναλαμβάνεται θεωρητικά μέχρι το επίπεδο όπου οι n παρατηρήσεις αποτελούν μια ή n ομάδες.

4.2. Δενδρόγραμμα (Dendrogram)

Τα αποτελέσματα των ιεραρχικών αυτών μεθόδων αναπαριστώνται σε ένα δισδιάστατο διάγραμμα που είναι γνωστό ως δενδρόγραμμα (dendrogram), επειδή μοιάζει με δένδρο, με όλα τα στοιχεία χωριστά στο ένα άκρο και μία συστάδα να περιέχει όλα τα στοιχεία στο άλλο. Όταν ένα ζεύγος παρατηρήσεων συγχωνευτεί ή μια ομάδα διασπαστεί, παρουσιάζεται ως ένας κλάδος στο δένδρο. Οι συσσωρευτικοί αλγόριθμοι ξεκινούν από τα φύλλα (n συστάδες) αυτού του δέντρου ενώ οι διαιρετικοί από τη ρίζα (μια συστάδα). Εάν κόψουμε αυτό το δέντρο σε κάποιο επίπεδο θα μας δώσει συσταδοποίηση

συγκεκριμένης ακρίβειας.

Η επιλογή του αριθμού των ομάδων γίνεται συνήθως με εμπειρικά κριτήρια, αν και έχουν αναπτυχθεί και ορισμένα στατιστικά. Η μελέτη του δενδρογράμματος αποτελεί το βασικό εργαλείο για να προσδιοριστεί ο αριθμός των ομάδων που θα προκύψουν από την ανάλυση.



- Βλέπουμε πως στο ένα άκρο υπάρχουν όλες οι παρατηρήσεις {α,β,γ,δ,ε}, όπου η κάθε μια αποτελεί μια ομάδα και στο άλλο όλες οι παρατηρήσεις μαζί σε μια ομάδα.
- Μια συστάδα σε ένα επίπεδο προκύπτει από την ένωση των συστάδων ή τη διάσπαση συστάδας σε προηγούμενο επίπεδο.
- Σε κάθε επίπεδο του δενδρογράμματος περιγράφεται ένας συγκεκριμένος τρόπος του διαμερισμού των παρατηρήσεων σε συστάδες.

4.3. Πλεονεκτήματα – Μειονεκτήματα των Ιεραρχικών μεθόδων

Όπως κάθε εφαρμόσιμη μέθοδος, έτσι και οι ιεραρχικές έχουν τα πλεονεκτήματά τους και τα μειονεκτήματά τους.

Τα βασικά *πλεονεκτήματα* των ιεραρχικών μεθόδων είναι τα ακόλουθα:

- ✓ Παρουσιάζουν πολύ καλή προσαρμοστικότητα.
- ✓ Δεν απαιτούν να είναι γνωστός ο αριθμός των ομάδων εκ των προτέρων.
- ✓ Έχουν ευκολία χειρισμού κάθε τύπου μέτρου απόστασης ή ομοιότητας.
- ✓ Δημιουργούν πολλά επίπεδα με διαφορετική ομαδοποίηση το καθένα, επιτρέποντας στο χρήστη να επιλέξει εκείνος το επίπεδο που επιθυμεί.

Τα κύρια *μειονεκτήματα* των ιεραρχικών μεθόδων είναι τα εξής:

- ✓ Οι ομάδες που δημιουργούνται στα αρχικά βήματα δεν μπορούν να μεταβληθούν στη συνέχεια. Από τη στιγμή που δυο αντικείμενα τότε εντάχθηκαν στην ίδια ομάδα, δεν υπάρχει δυνατότητα να διαχωριστούν αργότερα και να ενταχθούν σε διαφορετικές ομάδες.
- ✓ Δημιουργούν πολλές φορές μερικές ομάδες με πολλές παρατηρήσεις και αφήνουν κάποιες παρατηρήσεις να αποτελούν μόνες τους μία ομάδα.
- ✓ Χρειάζεται να ελέγξουν πολλές αποστάσεις, και για το λόγο αυτό καθυστερούν όταν χρειάζεται να επεξεργαστούν μεγάλο αριθμό δεδομένων. Το υπολογιστικό κόστος είναι μεγάλο, τουλάχιστον $O(n^2)$, όπου n το πλήθος των δεδομένων.

4.4. Βήματα συγχώνευσης της Συσσωρευτικής Ιεραρχικής Συσταδοποίησης

Στη μέθοδο αυτή πραγματοποιείται μια σειρά συγχωνεύσεων, οι οποίες είναι μη αναστρέψιμες. Όταν ένας συσσωρευτικός αλγόριθμος, έχει τοποθετήσει δύο δεδομένα στην ίδια ομάδα, αυτά δεν μπορούν στη συνέχεια να εμφανιστούν σε διαφορετικές ομάδες.

- *Βήματα συγχώνευσης της συσσωρευτικής ιεραρχικής ομαδοποίησης*

1. Αρχίζουμε με *N* συστάδες, με την κάθε μία να περιέχει μόνο ένα δεδομένο και ένα *NxN* συμμετρικό πίνακα με τις αποστάσεις μεταξύ των συστάδων.

2. Βρίσκουμε στον πίνακα το ζεύγος των *U* και *V* συστάδων που έχουν την μικρότερη απόσταση μεταξύ τους όπως φαίνεται στον παρακάτω πίνακα.

	1	*2*	*3*	*…*	*U*	*…*	*V*	*…*	*N*
1									
2									
3									
…									
U									
…									
V									
…									
N									

3. Ενώνουμε τις συστάδες *U* και *V* σε μια συστάδα, την οποία ονομάζουμε *UV*. Ανανεώνουμε τον πίνακα αποστάσεων διαγράφοντας τις γραμμές και τις στήλες που

αντιστοιχούν στις *U* και *V* και προσθέτουμε μια γραμμή και μια στήλη με τις αποστάσεις της *UV* από τις υπόλοιπες συστάδες, όπως φαίνεται στον πίνακα που ακολουθεί.

	1	*2*	*3*	*…*	*UV*	*…*	*N*
1							
2							
3							
…							
UV							
…							
N							

4. Επαναλαμβάνουμε τα βήματα 2 και 3, (*N*-1) φορές, μέχρι να υπάρχει μόνο μια συστάδα. Καταγράφουμε τις συστάδες που δημιουργήθηκαν κατά τη διάρκεια της διαδικασίας και το επίπεδο (απόσταση) στο οποίο δημιουργήθηκε η κάθε μία.

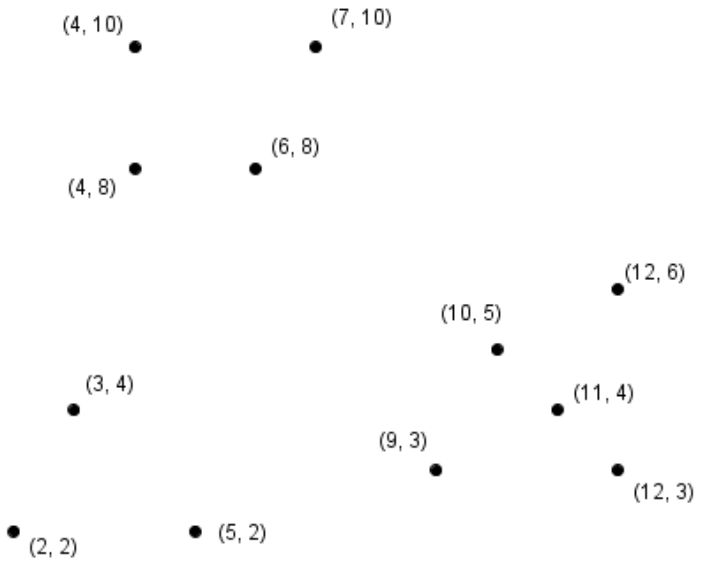
4.5. Αλγόριθμος Ιεραρχικής Συσταδοποίησης

Κάθε ιεραρχικός αλγόριθμος συσταδοποίησης δουλεύει ως εξής:

Αρχίζουμε με κάθε σημείο να αποτελεί τη δική του ομάδα. Όσο προχωρούμε, δημιουργούνται μεγαλύτερες ομάδες συνδυάζοντας δύο μικρότερες ομάδες. Έχουμε να αποφασίσουμε εκ των προτέρων:

1. Πως θα συμβολίζονται οι ομάδες
2. Πως θα επιλέγουμε ποιές δύο ομάδες θα συγχωνευθούν
3. Πότε θα σταματήσουμε να συνδιάζουμε ομάδες

Για αρχή, υποθέτουμε ότι ο χώρος είναι Ευκλείδειος. Αυτό μας επιτρέπει να αντιπροσωπεύεται μία ομάδα με το κεντροειδές ή το μέσο σημείο στην ομάδα.
Σημειώστε ότι σε ομάδα ενός σημείου, αυτό το σημείο είναι το κεντροειδές, οπότε μπορούμε να αρχικοποιήσουμε τις ομάδες ευθέως. Έπειτα, μπορούμε να χρησιμοποιήσουμε τον κανόνα συγχώνευσης ότι επιλέγουμε τις δύο ομάδες με τη μικρότερη απόσταση και η απόσταση μεταξύ δύο ομάδων είναι η Ευκλείδεια απόσταση μεταξύ των κεντροειδών τους. Υπάρχουν και άλλοι τρόποι για να ορίσουμε την απόσταση μεταξύ ομάδων και επίσης μπορούμε να επιλέξουμε το καλύτερο ζεύγος ομάδων με κριτήρια πέρα από τη μεταξύ τους απόσταση.
[2] Παράδειγμα:

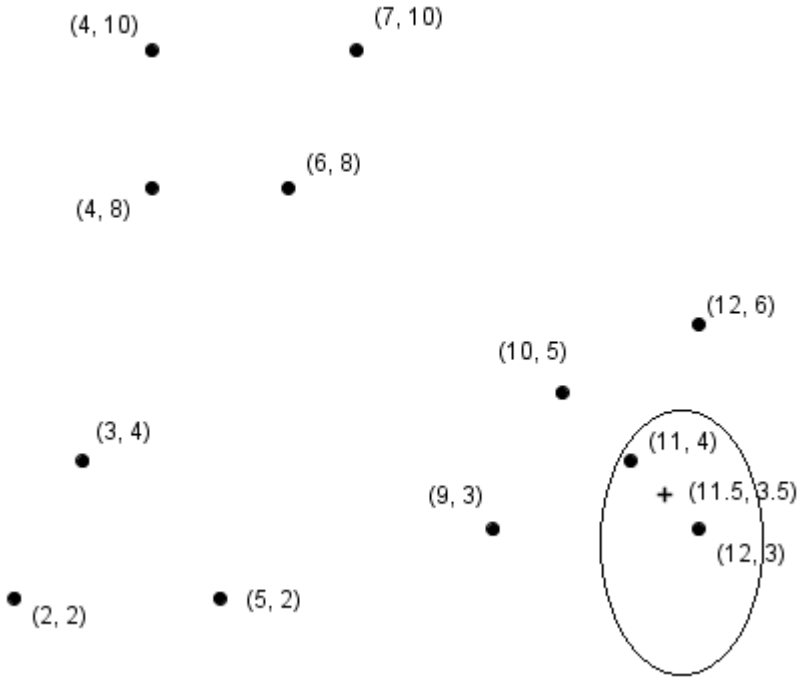


Σχήμα 4.1: *Δώδεκα σημεία που θα ομαδοποιηθούν ιεραρχικά.*

[3] A. Rajaraman, Jaffrey D. Ullman, "Εξόρυξη από μεγάλα Σύνολα Δεδομένων", ελληνική έκδοση, εκδόσεις Νέων Τεχνολογιών.

Ας δούμε πως η ιεραρχική ομαδοποίηση θα δούλευε στα δεδομένα του Σχήματος 1. Αυτά τα σημεία βρίσκονται σε ένα δισδιάστατο Ευκλείδειο χώρο και κάθε σημείο συμβολίζεται με τις συντεταγμένες του (x,y). Αρχικά, κάθε σημείο είναι σε μία ομάδα μόνο του και αποτελεί το κεντροειδές αυτής της ομάδας. Μεταξύ όλων των ζευγών σημείων, δύο είναι τα πλησιέστερα: (10,5) και (11,4) και (12,3). Κάθε ένα βρίσκεται σε απόσταση $\sqrt{2}$. Ας λύνουμε τις ισοπαλίες αυθαίρετα και ας αποφασίσουμε να συνδυάσουμε το (11,4) με το (12,3). Το αποτέλεσμα φαίνεται στο Σχήμα 2, που περιλαμβάνει το κεντροειδές της νέας ομάδας που είναι το (11,5, 3,5).

Στο επόμενο βήμα, το (10,5) συνδυάζεται με τη νέα ομάδα, καθώς είναι κοντά στο (11,4). Αλλά ο κανόνας μας για τις αποστάσεις απαιτεί να συγκρίνουμε μόνο κεντροείδη και η απόσταση από το (10,5) στο κεντροειδές της νέας ομάδας είναι $1,5\sqrt{2}$, που είναι λίγο μεγαλύτερο από το 2.



Σχήμα 4.2: *Συνδυάζοντας τα δύο πρώτα σημεία σε μία ομάδα.*

Συνεπώς, τώρα οι δύο πλησιέστερες ομάδες είναι αυτές των σημείων (4,8) και (4,10). Τις συνδυάζουμε σε μία ομάδα με κεντροειδές το (4,9).
Σε αυτό το σημείο, τα δύο πλησιέστερα κεντροειδή είναι τα (10,5) και (11,5, 4,10), οπότε συνδυάζουμε αυτές τις δύο ομάδες. Το αποτέλεσμα είναι μία ομάδα με τρία σημεία, (10,5), (11,4) και (12,3). Το κεντροειδές αυτής της ομάδας τυγχάνει να είναι ένα από τα σημεία της, αλλά αυτό είναι συμπτωματικό. Η κατάσταση των ομάδων απεικονίζεται στο Σχήμα 3.

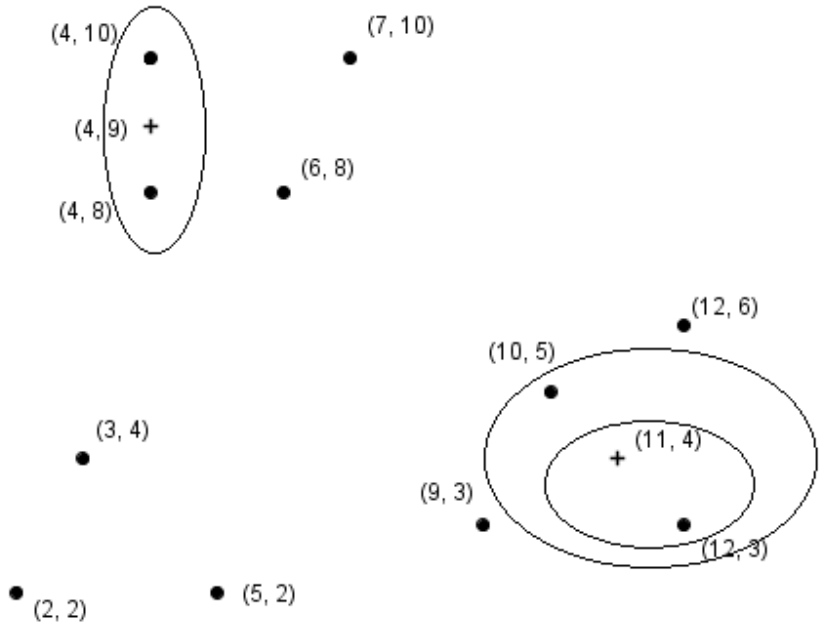
Τώρα υπάρχουν διάφορα ζεύγη κεντροειδών σε απόσταση $\sqrt{5}$ και αυτά είναι τα πλησιέστερα κεντροειδή. Δείχνουμε στο Σχήμα 4 το αποτέλεσμα της επιλογής τριών από αυτά:

1. Το (6,8) συνδυάζεται με την ομάδα δύο στοιχείων που έχει ως κεντροειδές το (4,9).

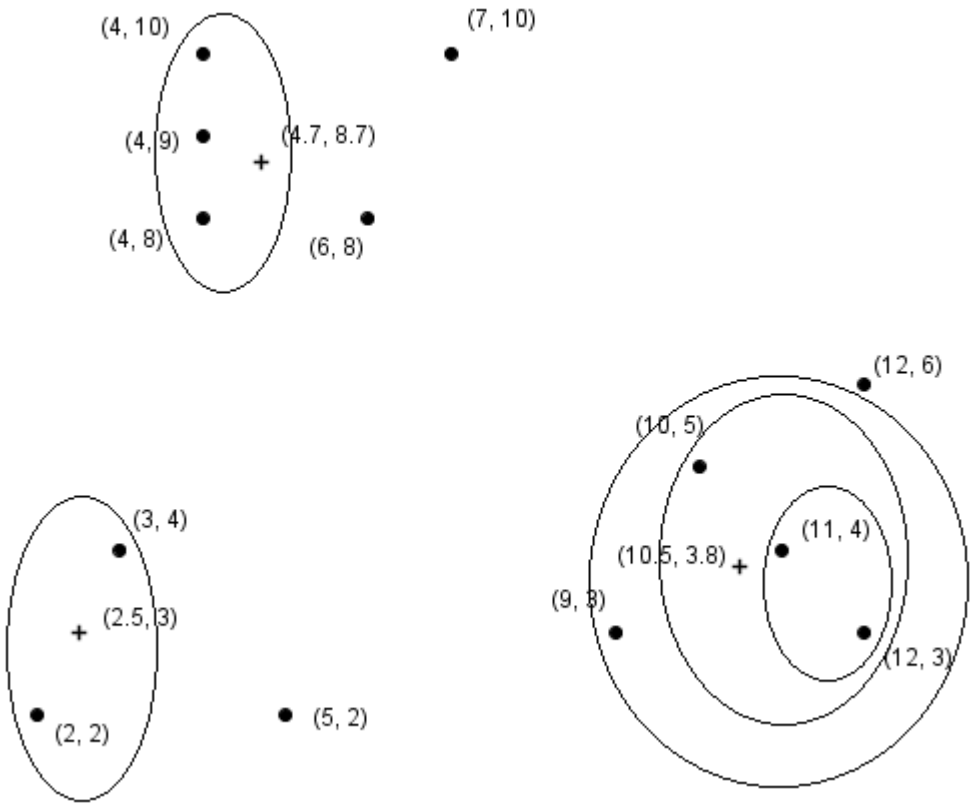
2. Το (2,2) συνδυάζεται με το (3,4).

3. Το (9,3) συνδυάζεται με την ομάδα τριών στοιχείων με κεντροειδές το (11,4).

Μπορούμε να συνεχίζουμε να συνδυάζουμε ομάδες περαιτέρω. Στη συνέχεια, θα συζητήσουμε εναλλακτικούς τρόπους τερματισμού.



Σχήμα 4.3: *Η ομαδοποίηση μετά από δύο επιπλέον βήματα.*



Σχήμα 4.4: *Τρία ακόμη βήματα της ιεραρχικής ομαδοποίησης.*

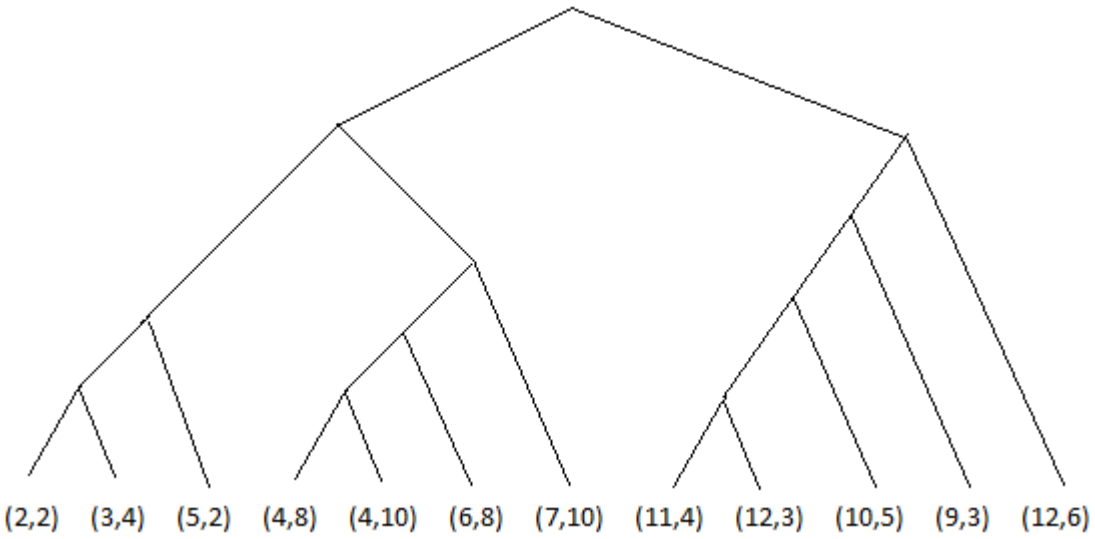
Υπάρχουν διάφορες προσεγγίσεις που μπορούμε να χρησιμοποιήσουμε σχετικά με τον τερματισμό της διαδικασίας ομαδοποίησης.

1. Μπορεί να μας έχει λεχθεί, ή να έχουμε γνώση, σχετικά με το πόσες ομάδες υπάρχουν στα δεδομένα. Για παράδειγμα, αν μας έχουν πει ότι τα δεδομένα των σκύλων έχουν προέλθει από τις ράτσες Τσιουάουα, Ντατσχάουντ και Μπιγκλ, τότε γνωρίζουμε να σταματήσουμε όταν έχουν μείνει τρεις ομάδες.

2. Μπορούμε να σταματήσουμε τις συγχωνεύσεις όταν, σε κάποιο σημείο, ο καλύτερος συνδυασμός από τις υπάρχουσες ομάδες παράγει μία ομάδα που είναι ανεπαρκή. Για να δώσουμε ένα παράδειγμα, θα μπορούσαμε να επιμείνουμε ότι κάθε ομάδα έχει μέση

απόσταση μεταξύ του κεντροειδούς της και των σημείων της που δεν ξεπερνά κάποιο όριο. Αυτή η προσέγγιση έχει νόημα αν έχουμε λόγο να πιστεύουμε ότι δεν εκτείνεται κάποια ομάδα πάρα πολύ στο χώρο.

3. Μπορούμε να συνεχίζουμε την ομαδοποίηση μέχρι να μείνει μόνο μία ομάδα. Όμως, δεν έχει νόημα να επιστρέψουμε μία ομάδα που αποτελείται από όλα τα αντικείμενα. Αντιθέτως, επιστρέφουμε το δένδρο που αναπαριστά τον τρόπο συνδυασμού όλων των σημείων. Αυτή η μορφή απάντησης έχει μεγάλη σημασία σε κάποιες εφαρμογές, όπως όταν τα σημεία είναι γονιδιωτά διαφόρων ειδών και το μέτρο απόστασης αντανακλά τη διαφορά μεταξύ γονιδιωμάτων. Τότε το δένδρο συμβολίζει την εξέλιξη αυτών των ειδών, δηλαδή, την πιθανή σειρά με την οποία δύο είδη προήλθαν από κοινό πρόγονο.



Σχήμα 4.5: *Δένδρο που δείχνει την πλήρη ομαδοποίηση των σημείων του σχήματος 1.2.*

4.6. Μέθοδοι σύνδεσης της Συσσωρευτικής Ιεραρχικής Συσταδοποίησης

Οι κυριότερες *μέθοδοι σύνδεσης* (linkage methods), ανάμεσα σε δυο cluster, στις συσσωρευτικές ιεραρχικές μεθόδους είναι οι εξής (Everitt et al., 2011· Johnson & Wichern, 2007):

- ***Κοντινότερος γείτονας ή απλή σύνδεση*** (*single linkage method*) η οποία χρησιμοποιεί την ελάχιστη απόσταση, από όλες τις πιθανές, μεταξύ ενός στοιχείου της *U* συστάδας και ενός στοιχείου της *V* συστάδας. Με απλούστερα λόγια, η απόσταση των συστάδων είναι η απόσταση μεταξύ των δύο πλησιέστερων σημείων τους.

$$d_{UV} = \min\left(d_{ij}\right), \text{ με } i \in U \text{ και } j \in V,$$

όπου d_{UV} είναι η απόσταση ανάμεσα στα δυο clusters *U* και *V* και d_{ij} είναι η απόσταση ανάμεσα στα αντικείμενα i και j δυο στοιχείων τους αντίστοιχα. Η d_{ij} μπορεί να είναι η ευκλείδεια απόσταση ή κάποια άλλη επιθυμητή που ανταποκρίνεται καλύτερα στα δεδομένα μας, όπως *euclidean*, *minkowski* και άλλες.

Ένα βασικό μειονέκτημα της απλής σύνδεσης, είναι ότι συνδέει συστάδες που έχουν δυο κοντινά σημεία και πολλά άλλα σημεία που βρίσκονται σε μεγάλες αποστάσεις. Κύριο πρόβλημα επίσης που μπορεί να προκύψει, είναι η δημιουργία μιας επιμήκους συστάδας στην οποία θα προστίθενται διαρκώς νέα σημεία. Πλεονέκτημα της απλής σύνδεσης είναι ότι μπορεί να εντοπίσει μη ελλειψοειδείς συστάδες.

- ***Μακρινότερος γείτονας ή πλήρης σύνδεση*** (*complete linkage method*) η οποία χρησιμοποιεί τη μέγιστη απόσταση από όλες τις πιθανές μεταξύ ενός στοιχείου της *U* συστάδας και ενός στοιχείου της *V* συστάδας.

$$d_{UV} = \max\left(d_{ij}\right), \text{ με } i \in U \text{ και } j \in V.$$

Στη μέθοδο αυτή η λογική υπολογισμού της απόστασης των συστάδων είναι αντίστροφη

από αυτή της απλής σύνδεσης. Πλεονέκτημά της, σε σχέση με την απλή σύνδεση, είναι πως αποφεύγεται η δημιουργία επιμηκών συστάδων. Αντιθέτως, η πλήρης σύνδεση τείνει να δημιουργήσει συμπαγείς και σφαιρικές συστάδες με συγκρίσιμη διάμετρο. Η μέθοδος της πλήρης σύνδεσης ενδείκνυται, όταν γνωρίζουμε ότι δεδομένα της ίδιας συστάδας είναι δυνατόν να βρίσκονται σε μεγάλες αποστάσεις μεταξύ τους. Ένα μειονέκτημα της μεθόδου είναι η ευαισθησία της στην ύπαρξη δεδομένων με ακραίες τιμές. Εάν υπάρχει ένα δεδομένο με ακραίες τιμές σε μια συστάδα, τότε δύσκολα αυτή η συστάδα θα συγχωνευθεί με κάποια άλλη.

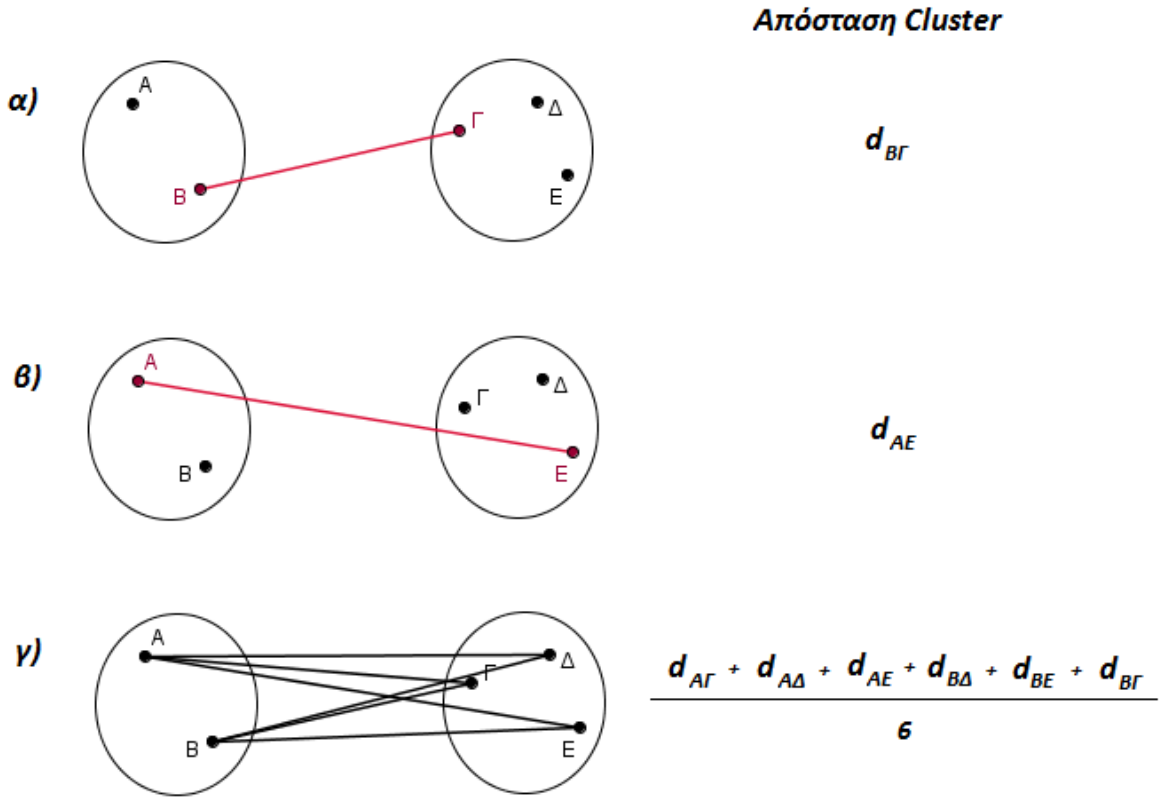
- ***Σύνδεση μέσου όρου*** (*average linkage method*) σύμφωνα με την οποία, η απόσταση συστάδων είναι ίση με τη μέση απόσταση όλων των ζευγών δεδομένων, όπου το πρώτο δεδομένο ανήκει στην πρώτη συστάδα και το δεύτερο δεδομένο ανήκει στη δεύτερη συστάδα. Πρόκειται δηλαδή για τη μέση απόσταση μεταξύ των δεδομένων των συστάδων. Ο μαθηματικός ορισμός δίνεται από την παρακάτω σχέση:

$$d_{UV} = \frac{1}{n_U n_V} \sum_{i \in U} \sum_{j \in V} d_{ij} \ ,$$

όπου U , V είναι οι δυο συστάδες, d_{ij} είναι η απόσταση μεταξύ των δεδομένων *i, j* και n_U , n_V είναι το πλήθος των παρατηρήσεων στις συστάδες *U* και *V* αντίστοιχα.

Η σύνδεση μέσου όρου αποτελεί ενδιάμεση λύση ανάμεσα στην ύπαρξη δεδομένων με ακραίες τιμές της μεθόδου πλήρης σύνδεσης, και των επιμηκών συστάδων της απλής σύνδεσης. Λόγω του υπολογισμού της μέσης απόστασης μεταξύ των ζευγών συστάδων, δεν δημιουργούνται μεγάλες ομάδες. Επιπλέον, αποφεύγεται η ύπαρξη παρατηρήσεων με ακραίες τιμές. Θεωρείται πως έχει μεγάλο υπολογιστικό κόστος, καθώς υπολογίζει τις αποστάσεις όλων των ζευγών. Μειονέκτημά της επίσης είναι η τάση της να διασπά μεγάλες ομάδες, οι οποίες μπορεί να είναι δυνατό να δημιουργηθούν.

Οι τρείς παραπάνω μέθοδοι παρουσιάζονται στο σχήμα που ακολουθεί:



Σχήμα 4.6: *Απόσταση μεταξύ των ομάδων για α)* ***απλή σύνδεση****, β)* ***πλήρης σύνδεση****, γ)* ***σύνδεση μέσου όρου****.*

➢ ***Σύνδεση των κεντροειδών*** (*centroids method*) η οποία υπολογίζει αποστάσεις μεταξύ των κέντρων των συστάδων.

$$d_{UV} = d_{k_U k_V},$$

όπου k_U, k_V είναι τα κέντρα των συστάδων U, V αντίστοιχα.

Για αυτή τη μέθοδο προτείνεται η χρήση της τετραγωνικής ευκλείδειας απόστασης. Έχει το πλεονέκτημα ότι δεν επηρεάζεται σημαντικά από την ύπαρξη παρατηρήσεων με ακραίες τιμές.

➢ ***Μέθοδος του Ward*** η οποία διαφέρει σημαντικά από τις προηγούμενες μεθόδους, καθώς δεν υπολογίζει όπως οι προηγούμενες κάποια απόσταση μεταξύ των

συστάδων. Η βασική ιδέα είναι είναι η μεγιστοποίηση της ομοιογένειας στο εσωτερικό των συστάδων. Το μέτρο που εφαρμόζεται είναι το άθροισμα του τετραγωνικού σφάλματος και στόχος της μεθόδου είναι η ελαχιστοποίηση του.
Το τετραγωνικό σφάλμα δίνεται από τη σχέση:

$$\mathrm{E} = \sum_{x \in C_i} \left(x - k_i \right)^2$$

όπου k_i είναι το κεντροειδές της συστάδας C_i. Η μέθοδος για να συνενώσει δυο συστάδες, επιλέγει το ζεύγος, το οποίο όταν ενωθεί θα μας δώσει τη συστάδα με το ελάχιστο τετραγωνικό σφάλμα. Γενικά αυτή η μέθοδος θεωρείται ως πολύ αποτελεσματική όμως τείνει να δημιουργεί συστάδες μικρού μεγέθους. Το προτεινόμενο μέτρο απόστασης για την μέθοδο αυτή φαίνεται πως είναι η τετραγωνική ευκλείδεια απόσταση. Πρόκειται για το ιεραρχικό ανάλογο του K-means.

Πρέπει να σημειωθεί ότι ανάλογα με τη μέθοδο της ένωσης των ομάδων δημιουργούνται διαφορετικά cluster.

4.7. Εφαρμογή Απλής Σύνδεσης

Παρακάτω παρουσιάζεται ένα παράδειγμα[10] με συνολικά πέντε παρατηρήσεις έτσι ώστε ακολουθώντας την μέθοδο της απλής σύνδεσης να βρούμε τις συστάδες που θα δημιουργηθούν, χρησιμοποιώντας ως μέτρο την τετραγωνική Ευκλείδεια απόσταση.

[10] *https://repository.kallipos.gr/bitstream/11419/2065/1/02_chapter_06.pdf*

Πίνακας παρατηρήσεων

ΤΡΟΦΕΣ	*ΕΝΕΡΓΕΙΑ*	*ΛΙΠΗ*
1:ΠΕΠΟΝΙ	33	0.3
2:ΣΤΑΦΥΛΙΑ	69	1
3:ΜΠΑΝΑΝΕΣ	85	1.1
4:ΠΟΡΤΟΚΑΛΙ	49	0.2
5:ΜΗΛΑ	58	0.6

Υπολογίζουμε την απόσταση κάθε παρατήρησης από όλες τις άλλες, χρησιμοποιώντας τον τύπο της τετραγωνικής Ευκλείδειας απόστασης και τοποθετούμε τα αποτελέσματα στο πίνακα αποστάσεων.

Πίνακας αποστάσεων

	1:ΠΕΠΟΝΙ	2:ΣΤΑΦΥΛΙΑ	3:ΜΠΑΝΑΝΕΣ	4:ΠΟΡΤΟΚΑΛΙ	5:ΜΗΛΑ
1:ΠΕΠΟΝΙ	0				
2:ΣΤΑΦΥΛΙΑ	1296.49	0			
3:ΜΠΑΝΑΝΕΣ	2704.64	256.01	0		
4:ΠΟΡΤΟΚΑΛΙ	256.01	400.64	1296.81	0	
5:ΜΗΛΑ	625.09	121.16	729.25	**81.16**	0

Θεωρούμε αρχικά κάθε παρατήρηση ως μια αυτοτελή ομάδα. Η διαδικασία της ομαδοποίησης αρχίζει με την συγχώνευση των δύο κοντινότερων παρατηρήσεων, δηλαδή των ομάδων με τη μικρότερη απόσταση. Αφού $\min(d_{ik}) = d_{54} = 81.16$, τα αντικείμενα 5 και 4, θα συγχωνευτούν για να σχηματίσουν την συστάδα (45). Το επόμενο επίπεδο ομαδοποίησης θα χρειαστεί τις αποστάσεις μεταξύ της συστάδας (45) και των παρατηρήσεων 1,2,3. Οι αποστάσεις των κοντινότερων γειτόνων είναι :

$$d_{(45)1} = \min\{d_{41}, d_{51}\} = \min\{256.01, 625.09\} = 256.01$$

$$d_{(45)2} = \min\{d_{42}, d_{52}\} = \min\{400.64, 121.16\} = 121.16$$

$$d_{(45)3} = \min\{d_{43}, d_{53}\} = \min\{1296.81, 729.25\} = 729.25$$

Έπειτα διαγράφοντας τις στήλες και τις γραμμές του αρχικού πίνακα αποστάσεων, που αντιστοιχούν στα αντικείμενα 4 και 5 και προσθέτοντας μια γραμμή και στήλη για την συστάδα (45), παίρνουμε τον καινούργιο πίνακα αποστάσεων :

	45	1:ΠΕΠΟΝΙ	2:ΣΤΑΦΥΛΙΑ	3:ΜΠΑΝΑΝΕΣ
45	0			
1:ΠΕΠΟΝΙ	256.01	0		
2:ΣΤΑΦΥΛΙΑ	**121.16**	1296.49	0	
3:ΜΠΑΝΑΝΕΣ	729.25	2704.64	256.01	0

Η μικρότερη απόσταση μεταξύ των ζευγαριών των ομάδων είναι τώρα $d_{(45)2} = 121.16$ και σε αυτό το βήμα συγχωνεύονται οι συστάδες 2 και (45) για να εξάγουμε την επόμενη ομάδα (245). Υπολογίζοντας :

$$d_{(245)1} = \min\left\{d_{(45)1}, d_{12}\right\} = \min\{256.01, 1296.49\} = 256.01$$

$$d_{(245)3} = \min\left\{d_{(45)3}, d_{23}\right\} = \min\{729.25, 256.01\} = 256.01$$

επαναλαμβάνοντας τα παραπάνω βήματα προκύπτει ο νέος πίνακας αποστάσεων:

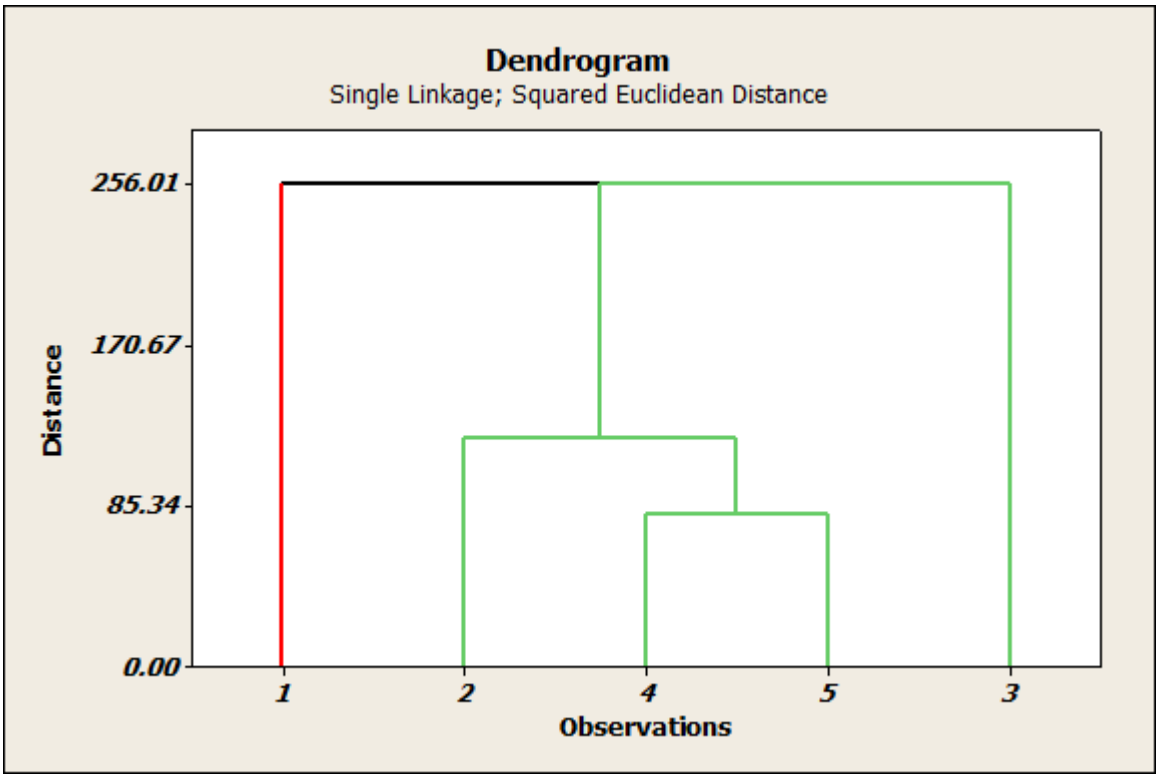
	245	1:ΠΕΠΟΝΙ	3:ΜΠΑΝΑΝΕΣ
245	0		
1:ΠΕΠΟΝΙ	256.01	0	
3:ΜΠΑΝΑΝΕΣ	**256.01**	2704.64	0

Η ελάχιστη απόσταση τώρα των κοντινότερων γειτόνων μεταξύ των υπάρχουσων ζευγαριών είναι $d_{3(245)} = 256.01$ και συγχωνεύονται οι συστάδες (245) και 3 για να πάρουμε την συστάδα (2345). Σε αυτό το σημείο προκύπτουν δύο ευδιάκριτες συστάδες, οι (2354) και η 3. Ο τελικός πίνακας αποστάσεων γίνεται :

	2345	1:ΠΕΠΟΝΙ
2345	0	
1:ΠΕΠΟΝΙ	**256.01**	0

Παρατηρούμε πως υπολογίζαμε διαδοχικά τον πίκακα αποστάσεων όσο υπήρχαν διαθέσιμες συστάδες για ομαδοποίηση και σταματήσαμε όταν όλες οι συστάδες συγχωνεύθηκαν σε μια.

Το δενδρόγραμμα για το παράδειγμα που μελετήσαμε φαίνεται παρακάτω.



Σχήμα 4.7: συγχώνευση *συστάδων με απλή σύνδεση.*

4.8. Εφαρμογή πλήρης σύνδεσης

Θεωρούμε ξανά τον πίνακα αποστάσεων του παραπάνω παραδείγματος:

	1:ΠΕΠΟΝΙ	2:ΣΤΑΦΥΛΙΑ	3:ΜΠΑΝΑΝΕΣ	4:ΠΟΡΤΟΚΑΛΙ	5:ΜΗΛΑ
1:ΠΕΠΟΝΙ	0				
2:ΣΤΑΦΥΛΙΑ	1296.49	0			
3:ΜΠΑΝΑΝΕΣ	2704.64	256.01	0		
4:ΠΟΡΤΟΚΑΛΙ	256.01	400.64	1296.81	0	
5:ΜΗΛΑ	625.09	121.16	729.25	**81.16**	0

Στο πρώτο στάδιο τα αντικείμενα 4 και 5 θα συγχωνευτούν αφού είναι τα πιο όμοια (μικρότερη απόσταση). Αυτό θα δώσει την ομάδα (45). Στο δεύτερο στάδιο τώρα φαίνεται η αλλαγή από την απλή σύνδεση αφού υπολογίζουμε την μέγιστη απόσταση μεταξύ των υπολοίπων αντικειμένων 1,2,3. Άρα έχουμε :

$$d_{(45)1} = \max\{d_{41}, d_{51}\} = \max\{256.01, 625.09\} = 625.09$$

$$d_{(45)2} = \max\{d_{42}, d_{52}\} = \max\{400.64, 121.16\} = 400.64$$

$$d_{(45)3} = \max\{d_{43}, d_{53}\} = \max\{1296.81, 729.25\} = 1296.81$$

και ο τροποποιημένος πίνακας αποστάσεων θα γίνει:

	45	1:ΠΕΠΟΝΙ	2:ΣΤΑΦΥΛΙΑ	3:ΜΠΑΝΑΝΕΣ
45	0			
1:ΠΕΠΟΝΙ	625.09	0		
2:ΣΤΑΦΥΛΙΑ	400.64	1296.49	0	
3:ΜΠΑΝΑΝΕΣ	1296.81	2704.64	**256.01**	0

Η επόμενη συγχώνευση θα γίνει μεταξύ των αμέσως πιο ομοίων ομάδων 2 και 3, συγκροτώντας την ομάδα (23). Έπειτα έχουμε :

$$d_{(23)(45)} = \max\left\{d_{2(45)}, d_{3(45)}\right\} = \max\{400.64, 1296.81\} = 1296.81$$

$$d_{(23)1} = \max\{d_{21}, d_{31}\} = \max\{1296.49, 2704.64\} = 2704.64$$

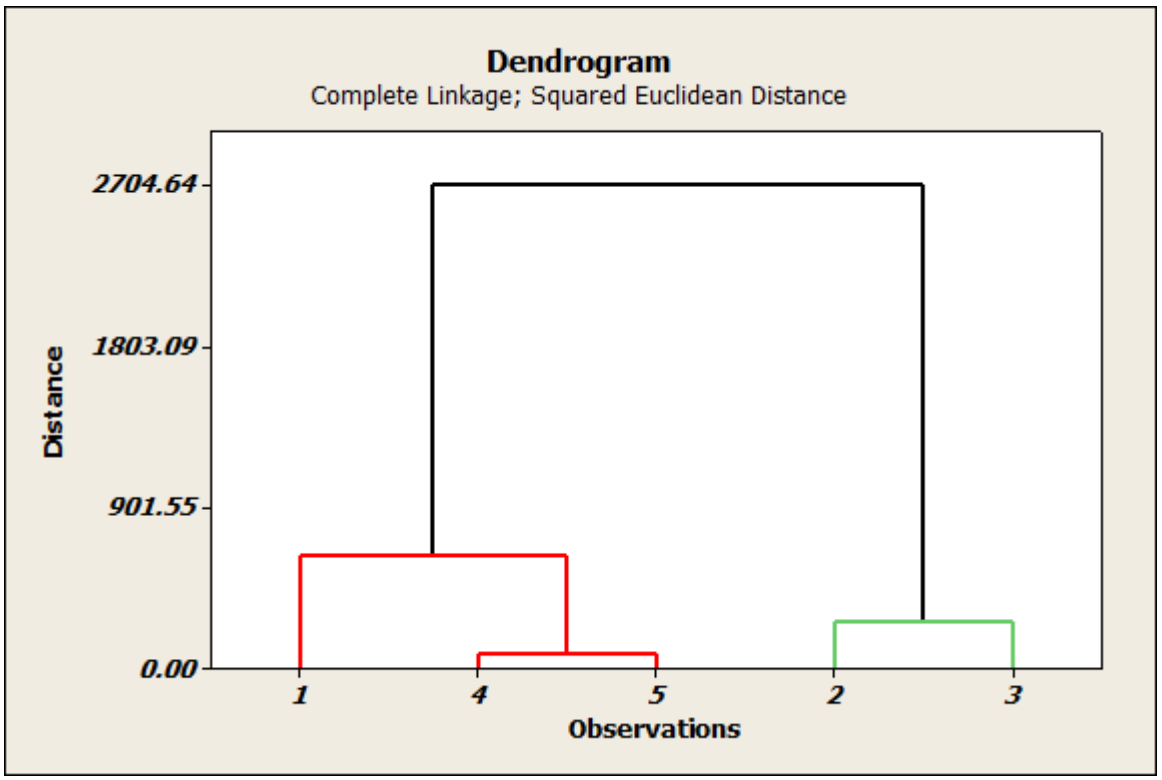
	23	45	1:ΠΕΠΟΝΙ
23	0		
45	1296.81	0	
1:ΠΕΠΟΝΙ	2704.64	**625.09**	0

Η επόμενη συγχώνευση θα παράξει την ομάδα (145). Στο τελευταίο στάδιο, θα συγχωνευτούν επομένως οι ομάδες (23) και (145) σε μια νέα ομάδα, την (12345), αφού:

$$d_{(145)(23)} = \max\left\{d_{1(23)}, d_{(45)(23)}\right\} = \max\{2704.64, 1296.81\} = 2704.64$$

	23	145
23	0	
145	**2704.64**	0

Το δενδρόγραμμα για το παράδειγμα που μελετήσαμε φαίνεται παρακάτω.



Σχήμα 4.8: συγχώνευση *συστάδων με πλήρη σύνδεση.*

4.9. Εφαρμογή πλήρης σύνδεσης στην R

> x <- mtcars["Honda Civic",]

> y <- mtcars["Camaro Z28",]

> dist(rbind(x, y))

Honda Civic

Camaro Z28 *335.8883*

> z <- mtcars["Pontiac Firebird",]

> dist(rbind(y, z))

Camaro Z28

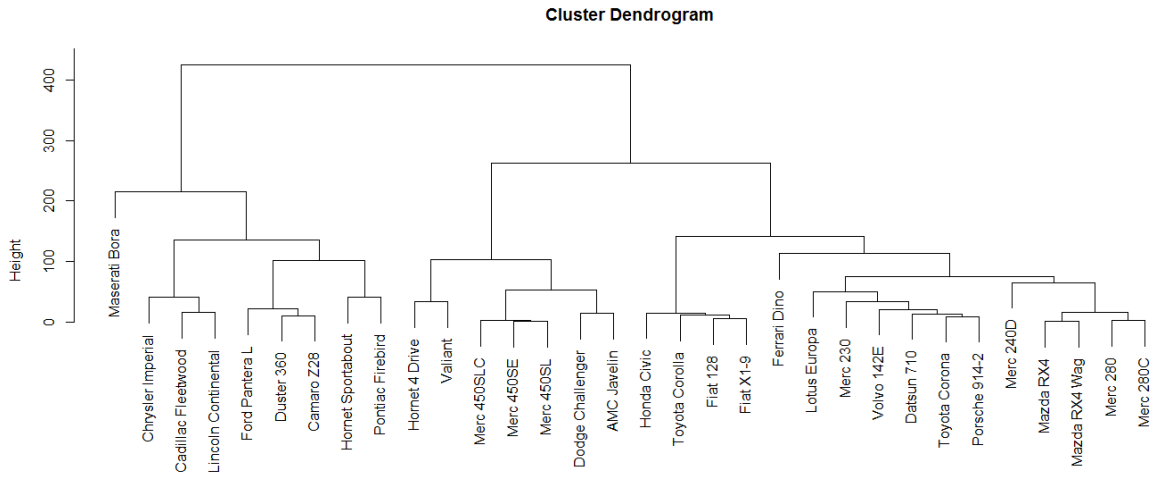
Pontiac Firebird *86.26658*

> dist(as.matrix(mtcars))

> d <- dist(as.matrix(mtcars))

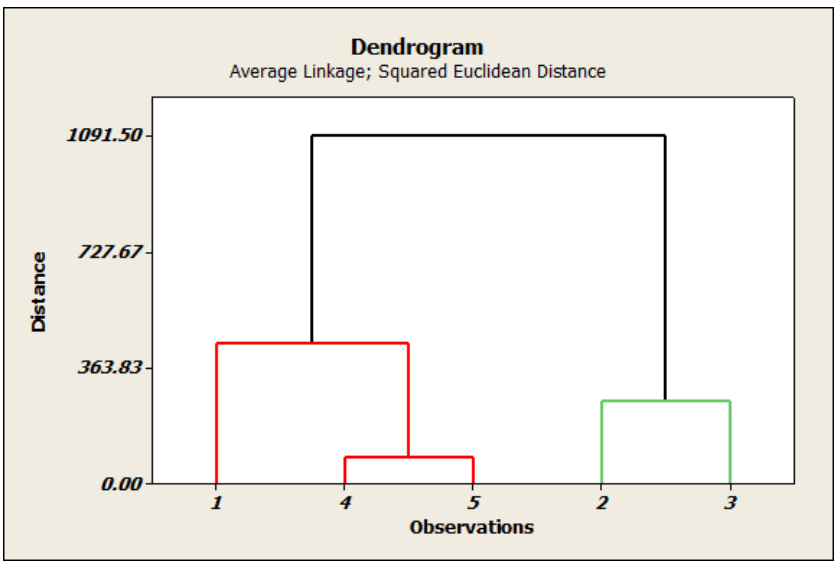
> hc <- hclust(d)

> plot(hc)



Σχήμα 4.9: δενδρόγραμμα πλήρης σύνδεσης.[11]

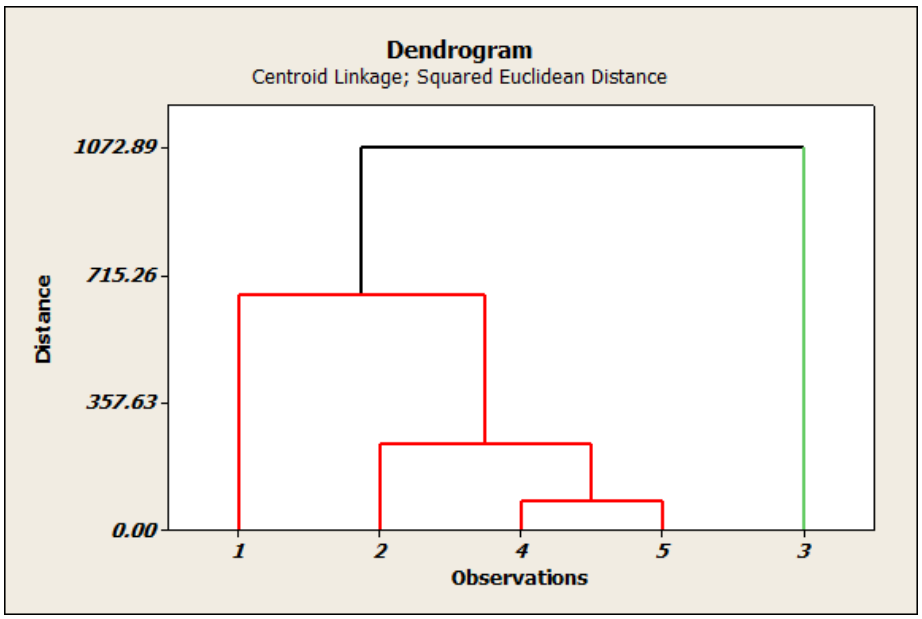
4.10. Εφαρμογή μέσου όρου



Σχήμα 4.10: δενδρόγραμμα μέσου όρου.

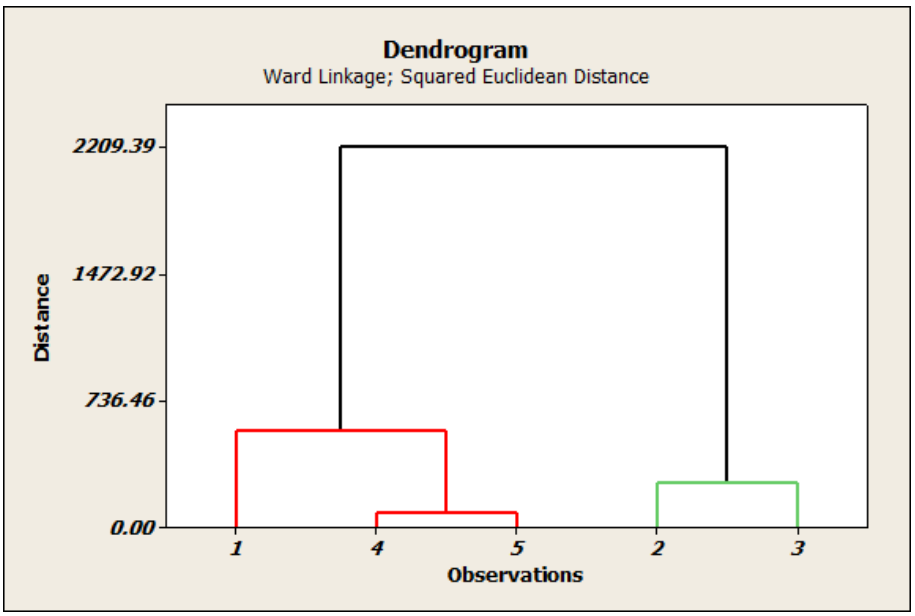
[11] **http://www.r-tutor.com/gpu-computing/clustering/hierarchical-cluster-analysis**

4.11. Εφαρμογή κεντροειδών



Σχήμα 4.11: δενδρόγραμμα κεντροειδών.

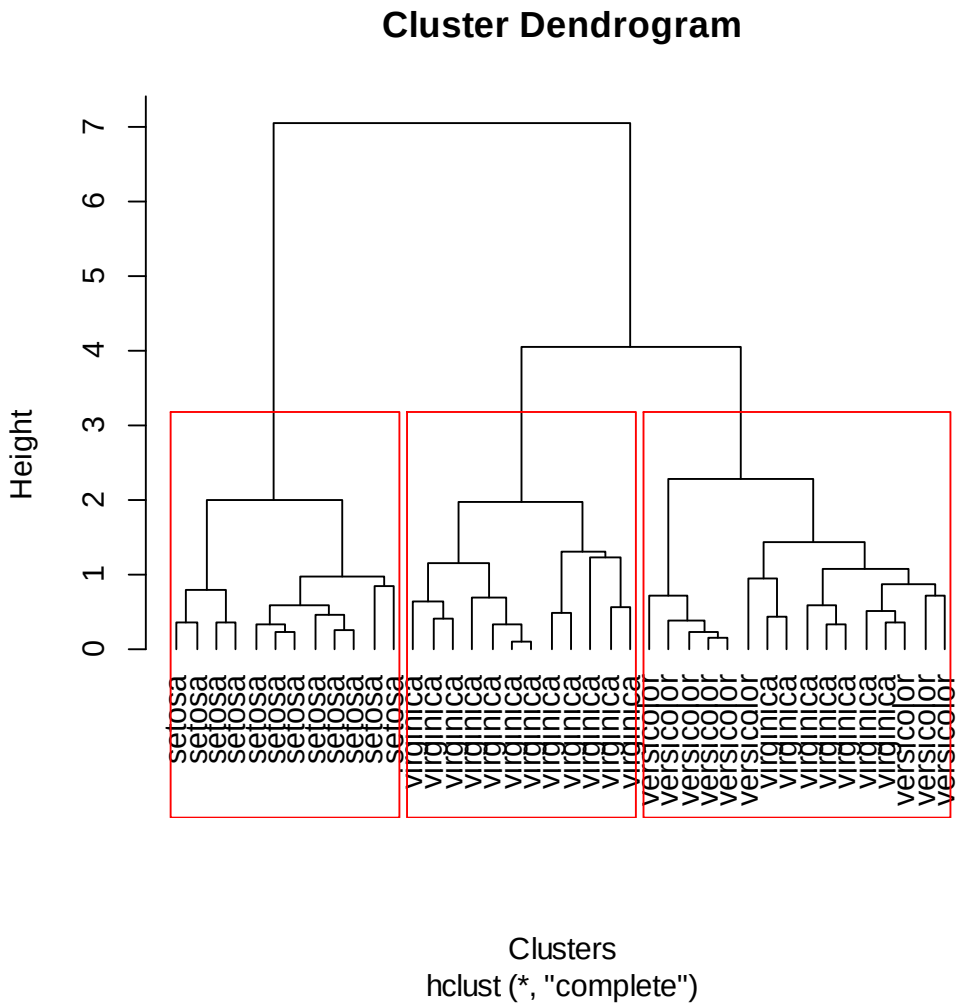
4.12. Εφαρμογή ward



Σχήμα 4.12: δενδρόγραμμα ward.

4.13. Παραδείγματα

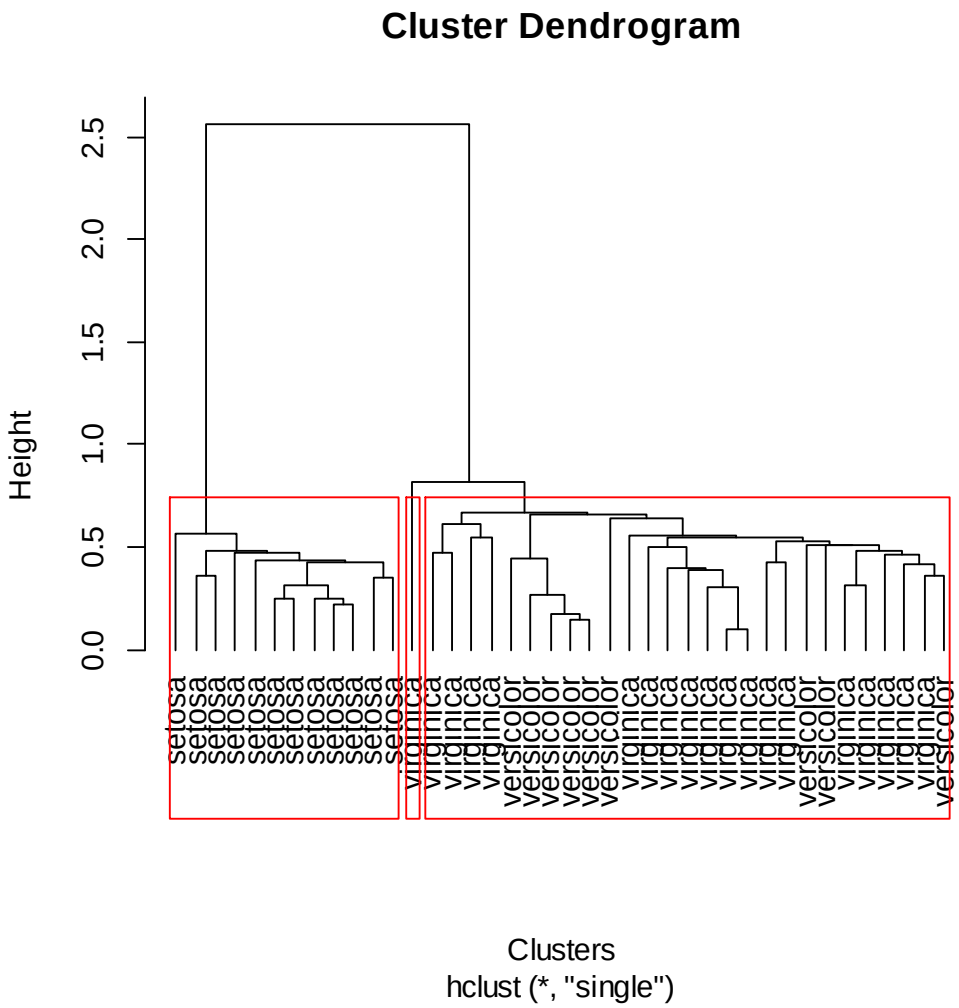
1. Ιεραρχική συσταδοποίηση πάνω στο σύνολο δεδομένων iris. [8]



Σχήμα 4.13: Ιεραρχική συσταδοποίηση με χρήση κριτηρίου πλήρης σύνδεσης (complete linkage).

```
> data(iris)
> set.seed(500)
> idx <- sample(1:dim(iris)[1], 40)
> irisSample <- iris[idx,]
> irisSample$Species <- NULL
```

```
> hc <- hclust(dist(irisSample), method="single")
> plot(hc, hang = -1, labels=iris$Species[idx], xlab="Clusters")
> rect.hclust(hc, 3,)
> hc <- hclust(dist(irisSample), method="complete")
> plot(hc, hang = -1, labels=iris$Species[idx], xlab="Clusters")
> rect.hclust(hc, 3,)
```



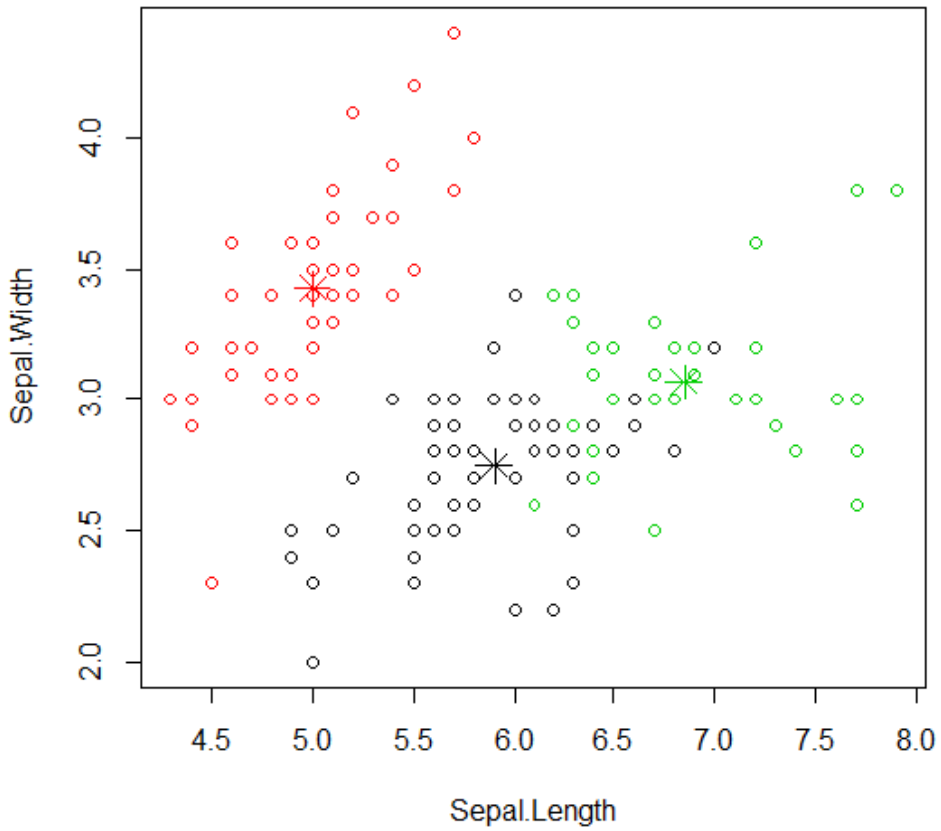
Σχήμα 4.14: Ιεραρχική συσταδοποίηση με χρήση κριτηρίου απλής σύνδεσης (single linkage).

2. Συσταδοποίηση k-means πάνω στο σύνολο δεδομένων iris.[8]

```
> iris_new <- iris
> iris_new$Species <- NULL
> kc <- kmeans(iris_new, 3)
> table(iris$Species, kc$cluster)
```

	1	2	3
setosa	0	50	0
versicolor	48	0	2
virginica	14	0	36

```
> plot(iris_new[c("Sepal.Length", "Sepal.Width")], col=kc$cluster)
> points(kc$centers[,c("Sepal.Length", "Sepal.Width")], col=1:3, pch=8, cex=2)
```



Σχήμα 4.15: απεικόνηση συσταδοποίησης, με k-means, k=3.

3.

Name	*Weight.kilos*	*Height.cms*
Beefy	11.31	33.79
Benny	9.34	34.38
Bertie	10.79	40.86
Biffy	11.04	37.07
Billy	9.74	33.77
Champ	2.94	22.98
Charger	2.99	16.21
Chalie	2.66	22.38
Chewy	2.32	19.68
Chechee	2.82	20.11
Chico	2.34	18.78
Chief	3.12	20.92
Laddy	29.57	61.69
Larry	29.64	59.03
Lassie	28.59	62.98
Lemmy	33.03	60.69
Loco	32.83	60.26
Loulou	31.23	61.34

Case Processing Summary[a]

Cases					
Valid		Missing		Total	
N	Percent	N	Percent	N	Percent
18	100,0	0	,0	18	100,0

a. Single Linkage

Proximity Matrix

	Squared Euclidean Distance																	
Case	1:Beefy	2:Benny	3:Bertie	4:Biffy	5:Billy	6:Champ	7:Charger	8:Chalie	9:Chewy	10:Chechee	11:Chico	12:Chief	13:Laddy	14:Larry	15:Lassie	16:Lemmy	17:Loco	18:Loulou
1:Beefy	,000	4,229	50,255	10,831	2,465	186,913	378,279	205,011	279,912	259,222	305,761	232,713	1111,84	973,046	1150,65	1195,37	1163,77	1155,81
2:Benny	4,229	,000	44,093	10,126	,532	170,920	370,471	188,622	265,370	246,143	292,360	219,860	1155,09	1019,71	1188,52	1253,43	1221,55	1206,01
3:Bertie	50,255	44,093	,000	14,427	51,371	381,317	668,463	407,607	520,333	494,083	558,929	456,433	786,577	685,471	806,134	887,847	862,122	837,224
4:Biffy	10,831	10,126	14,427	,000	12,580	264,138	499,942	286,020	378,451	355,210	410,214	323,549	949,505	828,202	979,331	1041,46	1012,58	996,669
5:Billy	2,465	,532	51,371	12,580	,000	162,664	353,916	179,859	253,585	234,482	279,460	208,947	1172,76	1034,08	1208,55	1267,11	1234,87	1221,93
6:Champ	186,913	170,920	381,317	264,138	162,664	,000	45,835	,438	11,274	8,251	18,000	4,276	2207,62	2012,49	2257,92	2327,45	2283,21	2271,81
7:Charger	378,279	370,471	668,463	499,942	353,916	45,835	,000	38,178	12,490	15,239	7,027	22,201	2774,93	2543,77	2842,79	2880,87	2830,83	2834,21
8:Chalie	205,011	188,622	407,607	286,020	179,859	,438	38,178	,000	7,406	5,179	13,062	2,343	2269,42	2071,14	2320,72	2389,99	2345,12	2334,13
9:Chewy	279,912	265,370	520,333	378,451	253,585	11,274	12,490	7,406	,000	,435	,810	2,178	2507,40	2294,80	2565,00	2624,92	2577,60	2571,34
10:Chechee	259,222	246,143	494,083	355,210	234,482	8,251	15,239	5,179	,435	,000	1,999	,746	2444,46	2234,08	2501,93	2559,38	2512,62	2507,04
11:Chico	305,761	292,360	558,929	410,214	279,460	18,000	7,027	13,062	,810	1,999	,000	5,188	2582,74	2365,35	2642,70	2698,32	2650,23	2645,99
12:Chief	232,713	219,860	456,433	323,549	208,947	4,276	22,201	2,343	2,178	,746	5,188	,000	2361,80	2155,68	2417,76	2476,26	2430,32	2423,95
13:Laddy	1111,84	1155,09	786,577	949,505	1172,76	2207,62	2774,93	2269,42	2507,40	2444,46	2582,74	2361,80	,000	7,081	2,625	12,972	12,673	2,878
14:Larry	973,046	1019,71	685,471	828,202	1034,08	2012,49	2543,77	2071,14	2294,80	2234,08	2365,35	2155,68	7,081	,000	16,705	14,248	11,689	7,864
15:Lassie	1150,65	1188,52	806,134	979,331	1208,55	2257,92	2842,79	2320,72	2565,00	2501,93	2642,70	2417,76	2,625	16,705	,000	24,958	25,376	9,659
16:Lemmy	1195,37	1253,43	887,847	1041,46	1267,11	2327,45	2880,87	2389,99	2624,92	2559,38	2698,32	2476,26	12,972	14,248	24,958	,000	,225	3,663
17:Loco	1163,77	1221,55	862,122	1012,58	1234,87	2283,21	2830,83	2345,12	2577,60	2512,62	2650,23	2430,32	12,673	11,689	25,376	,225	,000	3,726
18:Loulou	1155,81	1206,01	837,224	996,669	1221,93	2271,81	2834,21	2334,13	2571,34	2507,04	2645,99	2423,95	2,878	7,864	9,659	3,663	3,726	,000

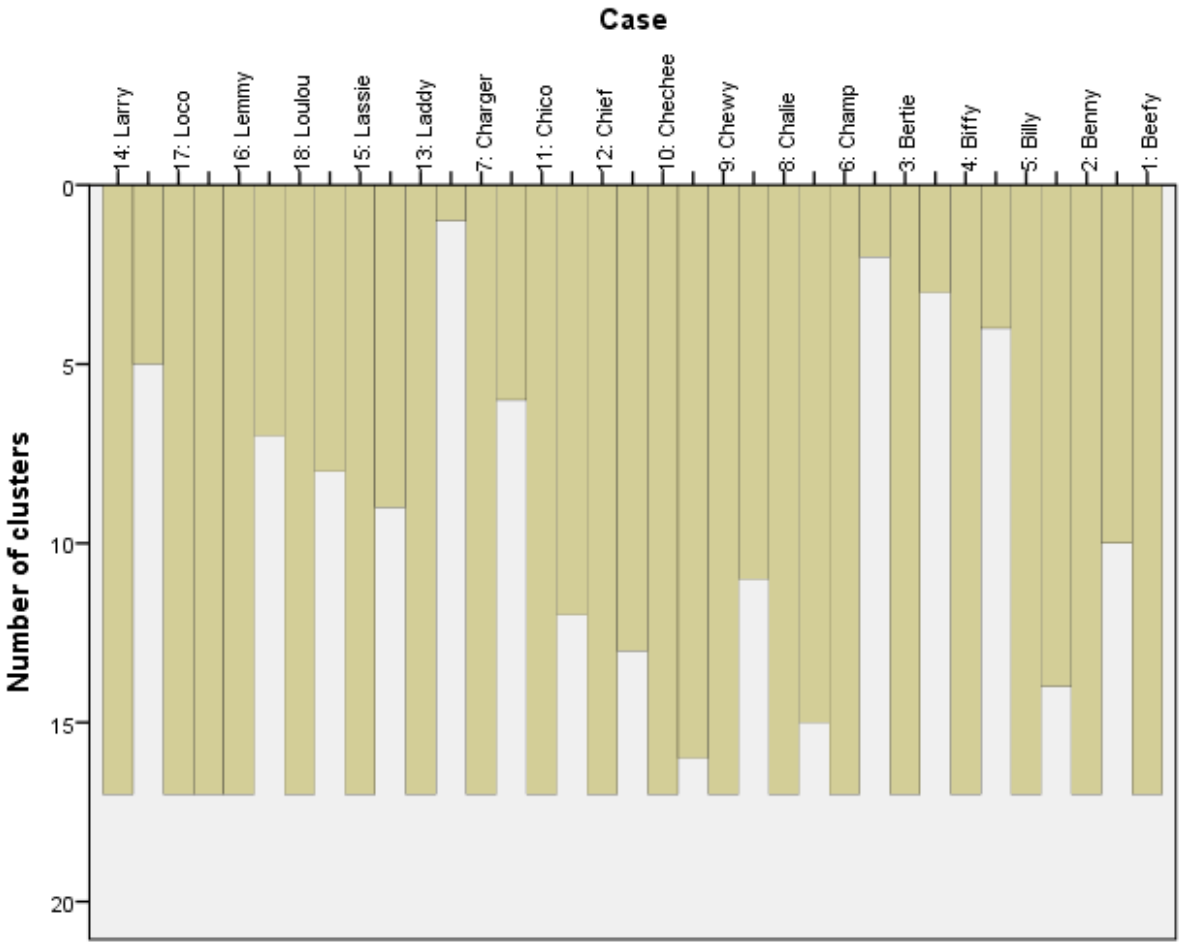
This is a dissimilarity matrix

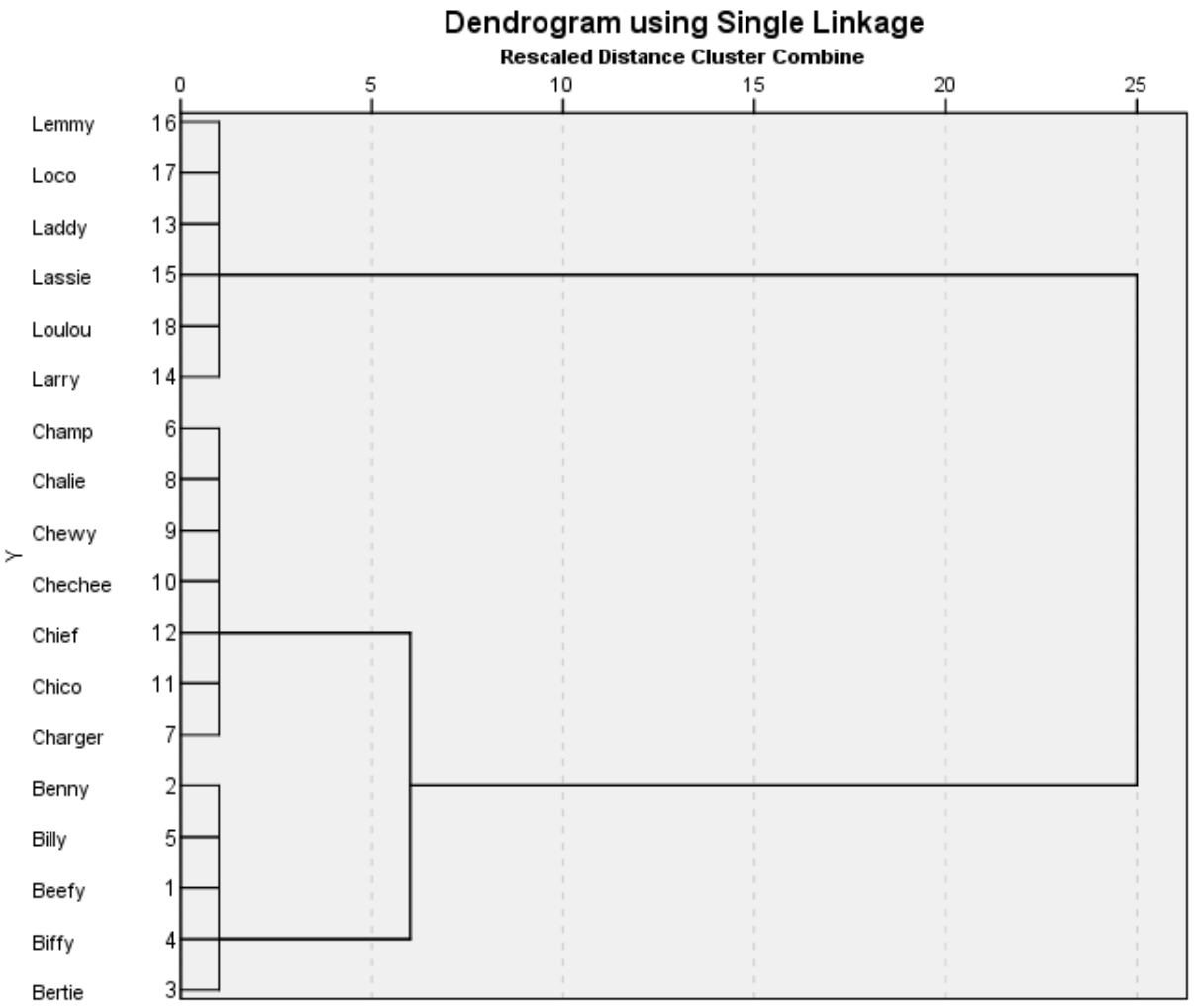
Παραπάνω δίνεται ο πίνακας αποστάσεων που υπολογίστηκε με την **Squared Euclidean distance.** Περιλαμβάνει τα τετράγωνα των αποστάσεων και είναι συμμετρικός ως προς τη διαγώνιο. Για παράδειγμα η απόσταση μεταξύ των σκύλων, Beefy και Benny, υπολογίστηκε ως εξής:

$$d\left(Beefy, Benny\right) = \left(11.31 - 9.34\right)^2 + \left(33.79 - 34.38\right)^2 = 3.8809 + 0.3481 = 4.229$$

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	16	17	,225	0	0	11
2	9	10	,435	0	0	5
3	6	8	,438	0	0	7
4	2	5	,532	0	0	8
5	9	12	,746	2	0	6
6	9	11	,810	5	0	7
7	6	9	2,343	3	6	12
8	1	2	2,465	0	4	14
9	13	15	2,625	0	0	10
10	13	18	2,878	9	0	11
11	13	16	3,663	10	1	13
12	6	7	7,027	7	0	16
13	13	14	7,081	11	0	17
14	1	4	10,126	8	0	15
15	1	3	14,427	14	0	16
16	1	6	162,664	15	12	17
17	1	13	685,471	16	13	0


Case
14: Larry
17: Loco
16: Lemmy
18: Loulou
15: Lassie
13: Laddy
7: Charger
11: Chico
12: Chief
10: Chechee
9: Chewy
8: Chalie
6: Champ
3: Bertie
4: Biffy
5: Billy
2: Benny
1: Beefy
Number of clusters
0
5
10
15
20


Dendrogram using Single Linkage
Rescaled Distance Cluster Combine
0
5
10
15
20
25
Y
Lemmy 16
Loco 17
Laddy 13
Lassie 15
Loulou 18
Larry 14
Champ 6
Chalie 8
Chewy 9
Chechee 10
Chief 12
Chico 11
Charger 7
Benny 2
Billy 5
Beefy 1
Biffy 4
Bertie 3

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] J. Han, M. Kamber, "*Data Mining: Concepts and Techniques*", Morgan Kaufmann Publishers, 2nd edition.

[2] J. Han, M. Kamber, J. Pei, "*Data Mining: Concepts and Techniques*", Morgan Kaufmann Publishers, 3rd edition.

[3] A. Rajaraman, Jaffrey D. Ullman, "*Εξόρυξη από μεγάλα Σύνολα Δεδομένων*", ελληνική έκδοση, εκδόσεις Νέων Τεχνολογιών.

[4] Bing Liu, "Web Data Mining", Exploring Hyperlinks, Contents, and Usage Data, Springer.

[5] I.H. Witten & E. Frank, "*Data Mining: Practical Machine Learning Tools and Techniques*", 2nd edition, Elsevier.

[6] J. Leskovec, A. Rajaraman, Jaffrey D. Ullman,"Mining of Massive Datasets ", Cambridge University Press.

[7] Ε. Παρασύρη, "*Εξόρυξη γνώσης και δεδομένων. Πλεονεκτήματα και μειονεκτήματα σε μια επιχείρηση*", πτυχιακή εργασία.

[8] Β.Σ. Βερύκιος, Β. Καγκλής, Η.Κ. Σταυρόπουλος, "*Η επιστήμη των δεδομένων μέσα από τη γλώσσα R*", Ελληνικά Ακαδημαϊκά Ηλεκτρονικά Συγγράμματα και Βοηθήματα.

Για τη δημιουργία των σχημάτων, χρησιμοποιήθηκαν: *minitab, spss, R, geogebra και excel.*