ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ

Διατμηματικό Π.Μ.Σ
Μαθηματικά των Υπολογιστών και των Αποφάσεων

Συσταδοποίηση φασμάτων μάζας

Συγγραφέας:
Αριστοτέλης Κομποθρέκας

Επιβλέπων:
Ιωάννης Χατζηλυγερούδης

15 Οκτωβρίου 2010

ΠΕΡΙΛΗΨΗ

Η συσταδοποίηση (clustering) είναι μία από τις βασικές εργασίες εξόρυξης γνώσης από δεδομένα (Data Mining). Παρουσιάζονται οι κυριότεροι αλγόριθμοι συσταδοποίησης, και αναλύεται ο αλγόριθμος X-means. Ο X-means επιτρέπει την ομαδοποίηση των δεδομένων χωρίς να χρειάζεται να προσδιοριστεί ακριβώς ο αριθμός των συστάδων. Το WEKA είναι ένα λογισμικό μηχανικής μάθησης όπου περιλαμβάνει τον αλγόριθμο X-means. Η φασματομετρία μάζας είναι μία τεχνική για τον προσδιορισμό της σύστασης -φάσματος ενός χημικού δείγματος ή μορίου. Η συσταδοποίηση χρησιμοποιείται στη φασματομετρία μάζας για την ανάδειξη ομάδων όμοιων φασμάτων, όπου έτσι επιτυγχάνεται η καλύτερη κατανόηση του δείγματος αλλά επίσης και της προέλευσής του. Στην εργασία εφαρμόζεται ο αλγόριθμος X-means, μέσω του WEKA σε φάσματα μάζας χημικών ουσιών.

ΕΥΧΑΡΙΣΤΙΕΣ

Με την περάτωση της παρούσας διπλωματικής εργασίας θέλω να ευχαριστήσω τον κ. Βασίλειο Βουτσινά, Αναπληρωτή Καθηγητή για την υποστήριξη και τον κ. Μάκη Μινέσχο για την παραχώρηση των δεδομένων.

Περιεχόμενα

Κατάλογος Σχημάτων

ΠΡΟΛΟΓΟΣ

Η φασματομετρία μάζας αποτελεί μια τεχνική για τον ποιοτικό και ποσοτικό προσδιορισμό χημικών ενώσεων. Αποτελεί ένα σημανικό εργαλείο για την μελέτη της ποιότητας του περιβάλοντος μέσω της ανάλυσης φασμάτων μάζας σωματιδίων της ατμόσφαιρας.

Οι συσκευές που συλλέγουν δείγματα από την ατμόσφαιρα, παράγουν μια τεράστια ποσότητα δεδομένων, με αποτέλεσμα να μην είναι δυνατή η ανάλυση με το "χέρι". Γιαυτό το λόγο έχουν εφαρμοστεί διάφορες μέθοδοι ομαδοποίησης (clustering) των φασμάτων μάζας όπου η ποσοτική και ποιοτική ανάλυση γίνεται μέσω των αντιπροσώπων των συστάδων.

Στην παρούσα εργασία εφαρμόζεται για πρώτη φορά ο αλγόριθμος X-means σε φάσματα μάζας, όπου το πλεονέκτημά του έναντι των άλλων που έχουν εφαρμοστεί μέχρι τώρα, είναι ότι εντοπίζει αποτελεσματικά τις συστάδες χωρίς να είναι γνωστό εκ των προτέρων το πλήθος τους. Η ερμηνεία των συστάδων που προκύπτουν γίνεται μέσω ειδικού λογισμικού χημικής ανάλυσης.

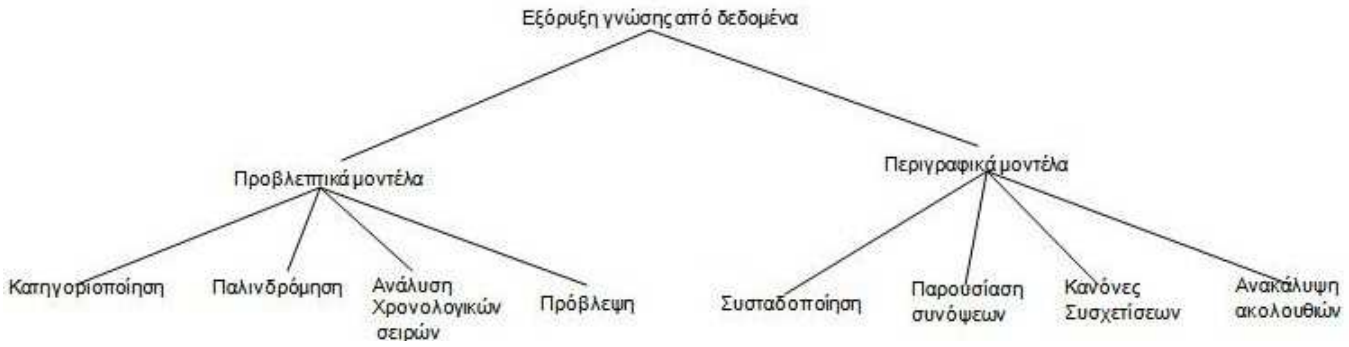
1 ΕΙΣΑΓΩΓΗ

Ο όγκος των δεδομένων που φυλάσσονται στα αρχεία και στις βάσεις δεδομένων αυξάνεται με έναν εκπληκτικό ρυθμό. Την ίδια στιγμή, οι χρήστες αυτών των δεδομένων επιζητούν από αυτά πιο εξειδικευμένες πληροφορίες. Απλές ερωτήσεις, που μπορούν να εκφραστούν σε μία δομημένη γλώσσα ερωτήσεων (SQL) δεν αρκούν για να υποστηρίξουν αυτές τις αυξανόμενες απαιτήσεις για πληροφορίες. Η εξόρυξη γνώσης από δεδομένα παρεμβαίνει προκειμένου να ικανοποιήσει αυτές τις ανάγκες. Η εξόρυξη γνώσης από δεδομένα (data mining), συχνά ορίζεται σαν η εύρεση πληροφοριών που είναι κρυμμένες σε μια βάση δεδομένων, ανακάλυψη καθοδηγούμενη από δεδομένα και συμπερασματική μάθηση. Η εξόρυξη γνώσης από δεδομένα περιλαμβάνει πολλούς διαφορετικούς αλγόριθμους για να εκπληρωθούν διαφορετικές εργασίες. Όλοι αυτοί οι αλγόριθμοι επιχειρούν να ταιριάξουν ένα μοντέλο στα δεδομένα. Οι αλγόριθμοι εξετάζουν τα δεδομένα και καθορίζουν ένα μοντέλο που να είναι το πλησιέστερο στα χαρακτηριστικά των δεδομένων που εξετάζονται. Οι αλγόριθμοι εξόρυξης γνώσης μπορεί να θεωρηθεί ότι αποτελούνται από τρία μέρη [1]:

- Μοντέλο: Ο σκοπός του αλγόριθμου είναι να ταιριάζει το μοντέλο στα δεδομένα.
- Προτίμηση: Πρέπει να χρησιμοποιούνται κάποια κριτήρια για να ταιριάξει ένα μοντέλο έναντι ενός άλλου.
- Αναζήτηση: Όλοι οι αλγόριθμοι απαιτούν μία τεχνική για να κάνουν αναζήτηση στα δεδομένα.

Το μοντέλο που δημιουργείται μπορεί να είναι είτε προβλεπτικό είτε περιγραφικό. Ένα προβλεπτικό μοντέλο (predictive model) κάνει μία πρόβλεψη για τις τιμές των δεδομένων, χρησιμοποιώντας γνωστά αποτελέσματα που έχει βρει από άλλα δεδομένα. Η μοντελοποίηση μπορεί να γίνει με βάση τη χρήση ιστορικών δεδομένων. Οι εργασίες εξόρυξης γνώσης από δεδομένα για το χτίσιμο ενός προβλεπτικού μοντέλου περιλαμβάνουν κατηγοριοποίηση, παλινδρόμηση, ανάλυση χρονολογικών σειρών και πρόβλεψη. Η πρόβλεψη μπορεί να χρησιμοποιηθεί επίσης για να υποδηλώσει ένα συγκεκριμένο τύπο λειτουργίας εξόρυξης γνώσης από δεδομένα.
Ένα περιγραφικό μοντέλο (descriptive model) αναγνωρίζει πρότυπα ή συσχετίσεις στα δεδομένα. Αντίθετα από το προβλεπτικό μοντέλο λειτουργεί σαν ένα μέσο που διερευνά τις ιδιότητες

των δεδομένων που εξετάζονται, όχι να προβλέψει νέες ιδιότητες. Η συσταδοποίηση, η παρουσίαση συνόψεων, οι κανόνες συσχετίσεων και η ανακάλυψη ακολουθιών συνήθως θεωρούνται σαν περιγραφικές εργασίες από τη φύση τους.



Σχήμα 1: Μοντέλα και εργασίες στην εξόρυξη γνώσης από δεδομένα

2 ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ ΑΠΟ ΔΕΔΟΜΕΝΑ-DATA MINING

2.1 Βασικές εργασίες εξόρυξης γνώσης από δεδομένα

2.1.1 Κατηγοριοποίηση

Η κατηγοριοποίηση (classification) απεικονίζει τα δεδομένα σε προκαθορισμένες ομάδες ή κατηγορίες - κλάσεις (classes). Αναφέρεται συχνά σαν εποπτευόμενη μάθηση (supervised learning), επειδή οι κατηγορίεσ-κλάσεις καθορίζονται πριν ακόμη εξεταστούν τα δεδομένα. Οι αλγόριθμοι κατηγοριοποίη-σης απαιτούν οι κατηγορίες να ορίζονται με βάση τις τιμές των γνωρισμάτων των δεδομένων. Συχνά περιγράφουν αυτές τις κατηγορίες κοιτάζοντας τα χαρακτηριστικά δεδομένων που είναι ήδη γνωστό ότι ανήκουν στις κατηγορίες. Η αναγνώριση προτύπου (pattern recognition) αποτελεί ένα είδος κατηγοριοποίησης, όπου ένα πρότυπο εισόδου κατηγοριοποιείται σε μία από διάφορες κατηγορίες, με βάση την εγγύτητα του ως προς αυτές τις προκαθορισμένες κατηγορίες[1].

2.1.2 Παλινδρόμηση

Η παλινδρόμηση (regression) χρησιμοποιείται για να απεικονιστεί ένα στοιχειώδες δεδομένο σε μία πραγματική μεταβλητή πρόβλεψης. Στην πραγματικότητα, η παλινδρόμηση περιλαμβάνει την εκμάθηση της συνάρτησης που κανει αυτή την απεικόνιση. Η παλινδρόμηση προϋποθέτει ότι τα σχετικά δεδομένα ταιριάζουν με μερικά γνωστά είδη συνάρτησης (π.χ. γραμμική, λογαριθμική κλπ.) και μετά καθορίζει την καλύτερη συνάρτηση αυτού του είδους που μοντελοποιεί τα δεδομένα που έχουν δοθεί. Ένα είδος ανάλυσης σφάλματος χρησιμοποιείται για να καθορίσει ποια συνάρτηση είναι η "καλύτερη".

2.1.3 Ανάλυση χρονοσειρών

Με την ανάλυση χρονολογικών σειρών ή χρονοσειρών (time series analysis), μελετάται η τιμή ενός γνωρίσματος καθώς μεταβάλλεται στο χρόνο. Οι τιμές συνήθως λαμβάνονται σε ίσα χρονικά διαστήματα (ημερήσια, εβδομαδιαία, ωριαία, κοκ.). Για να παρασταθούν οπτικά οι χρονοσειρές χρησιμοποιείται ένα διάγραμμα χρονοσειρών. Υπάρχουν τρεις βασικές λειτουργίες που πραγματοποιούνται στην ανάλυση χρονοσειρών. Στη μία περίπτωση, χρησιμοποιούνται μονάδες μέτρησης απόστασης για να καθορίζουν την ομοιότητα ανάμεσα σε διαφορετικές χρονοσειρές. Στη δεύτερη περίπτωση, εξετάζεται η δομή της χρονοσειράς για να καθορίσει (και ίσως να

κατηγοριοποίηση) τη συμπεριφορά της. Μία τρίτη εφαρμογή θα μπορούσε να είναι η χρήση διαγραμμάτων χρονοσειρών για την πρόβλεψη μελλοντικών τιμών.

2.1.4 Πρόβλεψη

Πολλές από τις πρακτικές εφαρμογές εξόρυξης γνώσης μπορούν να θεωρηθούν σαν πρόβλεψη μελλοντικών καταστάσεων με γνώση των προηγούμενων και των σημερινών δεδομένων. Η πρόβλεψη (prediction) μπορεί να θεωρηθεί σαν ένα είδος κατηγοριοποίησης. Η διαφορά είνα ότι ως πρόβλεψη θεωρείται περισσότερο το να δίνεται τιμή σε μία μελλοντική κατάσταση παρά σε μία τρέχουσα. Εδώ αναφερόμαστε σε ένα είδος εφαρμογής παρά σε μια προσέγγιση μοντελοποίησης. Οι εφαρμογές πρόβλεψης περιλαμβάνουν πρόγνωση πλημμυρών, αναγνώριση ομιλίας, μηχανική μάθηση και αναγνώριση προτύπου. Εάν και μπορούν να προβλεφθούν οι μελλοντικές τιμές με τεχνικές ανάλυσης χρονοσειρών ή παλινδρόμησης, μπορούν να χρησιμοποιηθούν επίσης και άλλες προσεγγίσεις.

2.1.5 Συσταδοποίηση

Η συσταδοποίηση (clustering) είναι παρόμοια με την κατηγοριοποίηση εκτός από το ότι οι συστάδεσ- ομάδες δεδομένων-δεν είναι προκαθορισμένς αλλά ορίζονται κυρίως από τα ίδια τα δεδομένα. Η συσταδοποίηση αναφέρεται εναλλακτικά και σαν μη εποπτευόμενη μάθηση ή τμηματοποίηση. 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 ομαδοποιούνται στις ίδες ομάδες.

2.1.6 Παρουσίαση συνόψεων

Η παρουσίαση συνόψεων(summarization) απεικονίζει τα δεδομένα σε υποσύνολά τους με συνοδευτικές απλές περιγραφές. Η σύνοψη των δεδομένων ονομάζεται επίσης και χαρακτηρισμός (characterization) ή γενίκευση (generalization) . Εξάγει ή παράγει αντιπροσωπευτικές πληροφορίες σχετικά με τις βάσεις δεδομένων. Αυτό γίνεται ανακτώντας, στην πραγματικότητα, τμήματα από τα δεδομένα. Εναλλακτικά, μπορούν να εξαχθούν από τα δεδομένα συνοπτικές πληροφορίες (όπως είναι ο μέσος όρος κάποου αριθμητικού γνωρίσματος). Εν ολίγοις, η παρουσίαση συνόψεων χαρακτηρίζει τα περιεχόμενα της βάσης δεδομένων.

2.1.7 Κανόνες συσχέτισης

Η ανάλυση συνδέσμων (link analysis), που εναλλακτικά αναφέρεται και σαν ανάλυση συγγένειας (affinity analysis) ή συσχέτιση (association), αναφέρεται στη διαδικασία εκείνη της εξόρυξης γνώσης που αποκαλύπτει συσχετίσεις μεταξύ των δεδομένων. Το καλύτερο παράδειγμα αυτού του είδους της εφαρμογής είναι ο προσδιορισμός κανόνων συσχετίσεων. Ένας κανόνας συσχέτισης (association rule) είναι ένα μοντέλο που αναγνωρίζει ειδικούς τύπους συσχέτισης μεταξύ δεδομένων. Αυτές οι συσχετίσεις συχνά χρησιμοποιούνται στις λιανικές πωλήσεις για να αναγνωριστούν προϊόντα που αγοράζονται μαζί. Συσχετίσεις χρησιμοποιούνται επίσης σε πολλές άλλες εφαρμογές.

2.1.8 Ανακάλυψη ακολουθιών

Η ακολουθιακή ανάλυση (sequential analysis) ή αλλιως ανακάλυψη ακολουθιών (sequence discovery) χρησιμοποιείται για να καθοριστούν σειριακά πρότυπα στα δεδομένα. Αυτά τα πρότυπα βασίζονται σε μία χρονική ακολουθία ενεργειών. Αυτά τα πρότυπα είναι παρόμοια με τις συσχετίσεις στο ότι συσχετίζονται τα δεδομένα (ή τα γεγονότα) που εξάγονται, με τη διαφορά ότι η συσχέτισή τους αυτή βασίζεται στο χρόνο. Αντίθετα με την ανάλυση καλαθιού αγορών , που προϋποθέτει να γνωρίζουμε ποια προϊόντα αγοράστηκαν ταυτόχρονα, στην ανακάλυψη ακολουθιών τα προϊόντα αγοράζονται με κάποια σειρά κατά τη διάρκεια μιας περιόδου.

2.2 Η Εξόρυξη ως στάδιο της ανακάλυψης γνώσης σε βάσεις δεδομένων

Οι όροι ανακάλυψη γνώσης σε βάσεις δεδομένων (Knowledge Discovery in Databases-KDD) και εξόρυξη γνώσης από δεδομένα (Data Mining) συχνά χρησιμοποιούνται εναλλακτικά για την ίδια έννοια. Στην πραγματικότητα, έχουν δοθεί πολλές διαφορετικές ονομασίες σε αυτήν τη διαδικασία ανακάλυψης χρήσιμων (κρυμμένων) προτύπων από τα δεδομένα: εξαγωγή γνώσης, ανακάλυψη πληροφοριών, εξερευνητική ανάλυση δεδομένων, συγκομιδή πληροφοριών, μη επιβλεπόμενη αναγνώριση προτύπου. Τα τελευταία χρόνια ο όρος KDD έχει χρησιμοποιηθεί για να εκφράσει μια διαδικασία που αποτελείται από πολλά βήματα, ένα από τα οποία είναι η εξόρυξη γνώσης από δεδομένα[2] [3].

Ορισμός 1.
Η ανακάλυψη γνώσης σε βάσεις δεδομένων (KDD) είναι η διαδικασία εύρεσης χρήσιμων πληροφοριών και προτύπων στα δεδομένα.

Ορισμός 2
Η εξόρυξη γνώσης από δεδομένα είναι η χρήση αλγορίθμων για την εξαγωγή των πληροφοριών και προτύπων που παράγονται με τη διαδικασία KDD. Η KDD διαδικασία αποτελείται από τα επόμενα πέντε βήματα:

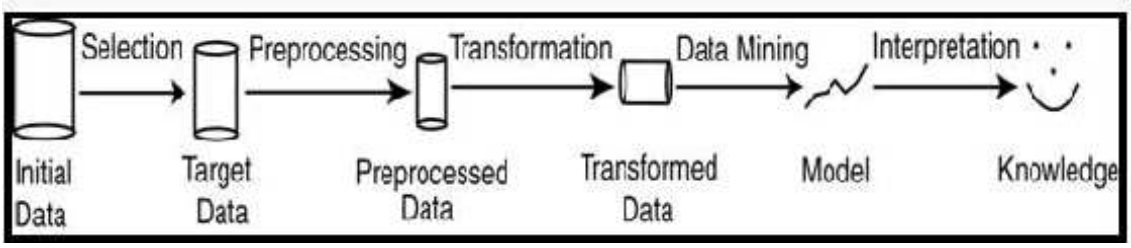
- Επιλογή: Τα δεδομένα που χρειάζονται για τη διαδικασία της ανακάλυψης γνώσης μπορούν να προέλθουν από πολλές διαφορετικές και ετερογενείς πηγές δεδομένων. Σε αυτό το πρώτο βήμα συλλέγονται δεδομένα από διάφορες βάσεις δεδομένων, αρχεία και μη ηλεκτρονικές πηγές.

- Προεπεξεργασία: Τα δεδομένα που πρόκειται να χρησιμοποιηθούν κατά την διαδικασία, ίσως να είναι λανθασμένα ή ελλιπή. Ίσως υπάρχουν ανώμαλα δεδομένα από πολλαπλές πηγές που περιλαμβάνουν διαφορετικούς τύπους δεδομένων και διαφορετικές μονάδες μέτρησης . Σε αυτό το βήμα μπορούν να πραγματοποιηθούν πολλές και διαφορετικές δραστηριότητες. Τα λανθασμένα δεδομένα μπορεί να διορθωθούν ή να αφαιρεθούν , ενώ τα ελλιπή δεδομένα πρέπει να συλλεχθούν ή να εκτιμηθούν .

- Μετασχηματισμός: Τα δεδομένα που προέρχονται από διαφορετικές πηγές χρειάζεται να μετατραπούν σε ένα κοινό σχήμα για την περαιτέρω επεξεργασία τους. Μερικά δεδομένα ίσως απαιτείται να κωδικοποιηθούν ή να μετασχηματιστούν σε πιο χρήσιμα σχήματα. Μπορεί να μειωθούν τα δεδομένα για να ελαττωθεί ο αριθμός των πιθανών τιμών των δεδομένων που θα ληφθούν υπόψη.

- Εξόρυξη γνώσης από δεδομένα: Με βάση το είδος της εξόρυξης που είναι να εκτελεστεί, σε αυτό το βήμα εφαρμόζονται αλγόριθμοι στα τροποποιημένα δεδομένα για να προκύψουν τα επιθυμητά αποτελέσματα.

- Ερμηνεία/αξιολόγηση: Είναι πολύ σημαντικό το πώς θα παρουσιαστούν στους χρήστες τα αποτελέσματα της εξόρυξης γνώσης, επειδή η χρησιμότητα ή μη των αποτελεσμάτων μπορεί να εξαρτάται ακριβώς από αυτήν την παρουσίαση. Σε αυτό το τελευταίο βήμα χρησιμοποιούνται διάφορες στρατηγικές οπτικοποίησης και γραφικές διαπαφές χρήστη(GUI).



Σχήμα 2: Διαδικασία εξόρυξης γνώσης από δεδομένα

2.3 Διάφορα θέματα εξόρυξης γνώσης από δεδομένα

Υπάρχουν πολλά σημαντικά θέματα υλοποίησης που σχετίζονται με την εξόρυξη γνώσης από δεδομένα[1]:

Ανθρώπινη αλληλεπίδραση: Αφού τα προβλήματα της εξόρυξης γνώσης από δεδομένα συνήθως δεν ορίζονται με ακρίβεια, μπορεί να είναι αναγκαία μια αλληλεπίδραση μεταξύ των ειδικών του πεδίου εφαρμογής με τους ειδικούς της συγκεκριμένης τεχνικής εξόρυξης γνώσης. Οι δεύτεροι χρησιμοποιούνται προκειμένου να μορφοποιήσουν τις ερωτήσεις και να βοηθήσουν στην ερμηνεία των αποτελεσμάτων. Οι πρώτοι είναι απαραίτητοι για να ταυτοποιήσουν τα δεδομένα εκπαιδευσης και να ορίσουν τα επιθυμητά αποτελέσματα.

Υπερπροσαρμογή: Όταν προκύπτει ένα μοντέλο που συσχετίζεται με μία δεδομένη κατάσταση μίας βάσης δεδομένων, είναι επιθυμητό αυτό το μοντέλο να ταιριάζει επίσης και σε μελλοντικές καταστάσεις της βάσης δεδομένων. Η υπερπροσαρμογή (overfitting) εμφανίζεται όταν το μοντέλο δεν ταιριάζει σε μελλοντικές καταστάσεις. Αυτό μπορεί να συμβαίνει εξαιτίας του μικρού μεγέθους των δεδομένων εκπαίδευσης. Έστω, για παράδειγμα , ένα μοντέλο κατηγοριοποίησης που κατατάσσει τους υπαλλήλους σε 'κοντούς ', 'μέτριους ' ή 'ψηλούς ', σε μια βάση δεδομένων που αφορά εργαζομένους. Εάν τα δεδομένα εκπαίδευσης είναι αρκετά λίγα, το μοντέλο ίσως λανθασμένα δείξει ότι το κάθε άτομο με ύψος κάτω από 1.80 είναι 'κοντό' επειδή στη βάση με τα δεδομένα εκπαίδευσης υπάρχει μόνο μια καταχώρηση για ύψος κάτω από 1.80. Σε αυτή την περίπτωση, πολλοί υπάλληλοι λανθασμένα θα καταχωρηθούν σαν 'κοντοί'. Η υπερπροσαρμογή μπορεί να εμφανιστεί και σε άλλες περιπτώσεις, ακόμα και όταν αλλάζουν τα δεδομένα.

Ερμηνεία των αποτελεσμάτων: Με τα σημερινά δεδομένα, τα αποτελέσματα από την εξόρυξη γνώσης πρέπει να ερμηνεύονται από ειδικούς του πεδίου, αλλιώς θα είναι χωρίς νόημα για το μέσο χρήστη.

Οπτικοποίηση των αποτελεσμάτων: Η οπτικοποίηση των αποτελεσμάτων των αλγορίθμων εξόρυξης γνώσης είναι χρήσιμη για να δούμε και να κατανοήσουμε ευκολότερα τα αποτελέσματα αυτά. Μεγάλα σύνολα δεδομένων: Τα ογκώδη σύνολα δεδομένων δημιουργούν προβλήματα όταν εφαρμόζονται αλγόριθμοι εξόρυξης γνώσης που έχουν σχεδιαστεί για

μικρά σύνολα δεδομένων. Πολλές εφαρμογές μοντελοποίησης αυξάνονται εκθετικά στον αριθμό των δεδομένων και γιάυτό το λόγο οι εφαρμογές αυτές είναι αναποτελεσματικές στα μεγαλύτερα σύνολα δεδομένων. Αποτελεσματικά εργαλεία για να αντιμετωπιστεί το πρόβλημα της κλιμάκωσης είναι η δειγματοληψία και ο παραλληλισμός.

Υψηλές διαστάσεις: Το σχήμα μίας συμβατικής βάσης δεδομένων μπορεί να αποτελείται από πολλά διαφορετικά γνωρίσματα. Το πρόβλημα εδώ είναι ότι ίσως δεν χρειάζονται όλα τα γνωρίσματα για να λυθεί ένα συγκεκριμένο πρόβλημα εξόρυξης γνώσης. Στην πράξη αν χρησιμοποιήσουμε κάποια γνωρίσματα μπορεί να εμποδίσουμε τη σωστή ολοκλήρωση μιας εργασίας. Η χρήση άλλων γνωρισμάτων μπορεί απλά να αυξήσει τη συνολική πολυπλοκότητα και να μειώσει την απόδοση ενός αλγορίθμου. Αυτό το πρόβλημα μερικές φορές αναφέρεται σαν η κατάρα των υψηλών διαστάσεων (dimensionality curse) , εννοώντας ότι υπάρχουν πολλά γνωρίσματα (διαστάσεις) που εμπλέκονται και είναι δύσκολο να καθοριστεί ποια γνωρίσματα πρέπει να χρησιμοποιηθούν. Μία λύση στο πρόβλημα των υψηλών διαστάσεων είναι να μειθούν τα γνωρίσματα, κάτι που αναφέρεται ως μείωση των υψηλών διαστάσεων (dimensionality reduction). Όμως δεν είναι πάντα εύκολο να προσδιοριστούν τα γνωρίσματα που δεν χρειάζονται.

Δεδομένα πολυμέσων: Οι περισσότεροι από τους αλγόριθμους που έχουν προταθεί κατά καιρούς στοχεύουν στα παραδοσιακά είδη δεδομένων(αριθμητικά, χαρακτήρες, κείμενο, κ.λπ). Η χρήση των δεδομένων πολυμέσων , σαν και αυτά που βρίσουκε στις γεωγραφικές βάσεις δεδομένων, περιπλέκει ή καθιστά ακτάλληλους πολλούς από τους αλγόριθμους αυτούς.

Ελλιπή δεδομένα: Κατά τη διάρκεια της φάσης της προεπεξεργασίας στη διαδικασία KDD , τα δεδομένα που λείπουν μπορούν να συπληρωθούν με κατέκτίμηση τιμές. Αυτή η προσέγγιση, καθώς και άλλες προσεγγίσεις που αντιμετωπίζουν το πρόβλημα των ελλιπών δεδομένων, ενδεχομένως οδηγούν σε λανθασμένα αποτελέσματα κατά την εξόρυξη γνώσης από δεδομένα.

Άσχετα δεδομένα: Μερικά γνωρίσματα στη βάση δεδομένων ίσως να μην έχουν ενδιαφέρον όσον αφορά στη συγκεκριμένη εργασία εξόρυξης γνώσης που πραγματοποιείται.

Δεδομένα με θόρυβο: Μερικές τιμές των γνωρισμάτων μπορεί να είναι άκυρες ή λανθασμένες. Αυτές οι τιμές συνήθως διορθώνονται πρίν τρέξουμε την εφαρμογή της εξόρυξης γνώσης από δεδομένα.

Δεδομένα που αλλάζουν: Οι βάσεις δεδομένων δε μπορεί να θεωρηθούν ότι είναι στατικές. Όμως οι περισσότεροι αλγόριθμοι εξόρυξης γνώσης υποθέτουν ότι η βάση δεδομένων είναι στατική. Αυτό απαιτεί ο αλγόριθμος να ξανατρέχει από την αρχή κάθε φορά που αλλάζει η βάση δεδομένων.

Ολοκλήρωση: Η διαδικασία KDD σήμερα δεν αποτελεί μέρος των συνηθησμένων εργασιών επεξεργασίας των δεδομένων. Οι απαιτήσεις της KDD μπορεί να αντιμετωπίζονται σαν ιδιαίτερες, ασυνήθιστες, ή σαν απαιτήσεις της "μιας φοράς". Οι απαιτήσεις αυτές γίνονται άρα αναποτελεσματικές και όχι αρκετά γενικές για να χρησιμοποιούνται σε συνεχή βάση. Φυσικά ένας επιθυμητός στόχος είναι η ενσωμάτωση των λειτουργιών της εξόρυξης γνώσης σε παραδοσιακά συστήματα διαχείρησης βάσεων δεδομένων.

Εφαρμογή: Αποτελεί πρόκληση το να προσδιοριστεί η ενδεικνυόμενη χρήση για μια πληροφορία που προήλθε από τη λειτουργία της εξόρυξης γνώσης. Πράγματι, η αποτελεσματική ερμηνεία των αποτελεσμάτων θεωρείται μερικές φορές, από τα στελέχη μίας επιχείρησης, πιο δύσκολο έργο από το τρέξιμο ενός αλγορίθμου. Επειδή τα δεδομένα είναι πληροφορίες που δεν ήταν γνωστές στο παρελθόν, οι τεχνικές των επιχειρήσεων πρέπει να τροποποιηθούν για να καθορίσουν τον τρόπο με τον οποίο θα χρησιμοποιήσουν τις κρυμμένες πληροφορίες.

3 ΣΥΣΤΑΔΟΠΟΙΗΣΗ

3.1 Γενικά

Συσταδοποίηση (clustering) είναι η οργάνωση μιας συλλογής προτύπων σε ομάδες, και επιτυγχάνεται βρίσκοντας ομοιότητες μεταξύ των προτύπων βάσει των χαρακτηριστικών που υπάρχουν σε αυτά. Οι ομάδες αυτές ονομάζονται συστάδες (clusters). Τα πρότυπα που βρίσκονται σε μια συστάδα είναι πιο όμοια μεταξύ τους, από άλλα πρότυπα που βρίσκονται σε άλλη συστάδα. Είναι σημαντικό να καταλάβουμε τη διαφορά μεταξύ συσταδοποίησης και κατηγοριοποίησης. Στην κατηγοριοποίηση διαθέτουμε μια συλλογή προκατηγοριοποιημένων προτύπων, όπου το πρόβλημα είναι να κατηγοριοποιήσουμε πρότυπα που δεν έχουν κατηγοριοποιηθεί. Τα πρότυπα που έχουν κατηγοριοποιηθεί χρησιμοποιούνται ως πρότυπα εκπαίδευσης για την περιγραφή των κλάσεων κατηγοριοποίησης όπου σε αυτές θα αντιστοιχηθούν τα υπόλοιπα πρότυπα. Στη συσταδοποίηση, το πρόβλημα είναι η ομαδοποίηση ενός συνόλου προτύπων, σε ομογενείς συστάδες που έχουν κάποιο νόημα. Η συσταδοποίηση έχει χρησιμοποιηθεί σε πολλά πεδία εφαρμογών, συμπεριλαμβανομένων της βιολογίας, ιατρικής, ανθρωπολογίας, μάρκετινγκ και οικονομίας. Εφαρμογές της συσταδοποίησης περιλαμβάνουν την ταξινομία φυτών και ζώων, την κατηγοριοποίηση βάσει ασθένειας, την επεξεργασίας, την αναγνώριση προτύπων, και την ανάκτηση κειμένων. Ένα από τα πρώτα πεδία στα οποία χρησιμο-ποιήθηκε η συσταδοποίηση ήταν η βιολογική ταξινομία. Πρόσφατες χρήσεις της συσταδοποίησης περιλαμβάνουν την εξέταση των δεδομένων των αρχείων λειτουργίας του Web (Web logs) για τον εντοπισμό προτύπων σχετικά με τον τρόπο χρήσης του δικτύου.

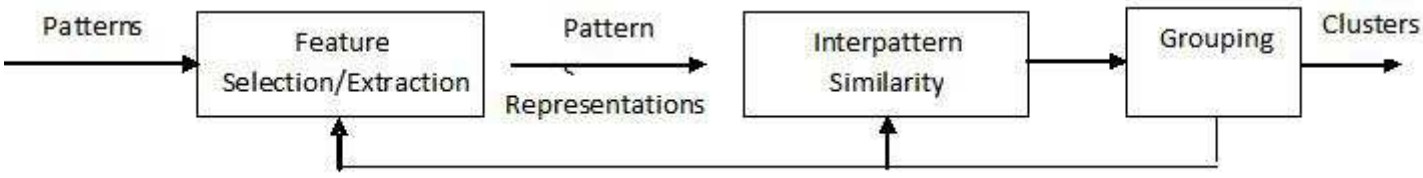
3.2 Εργασίες Συσταδοποίησης

Μια τυπική εργασία συσταδοποίησης περιλαμβάνει τα ακόλουθα βήματα [4]:

1. Αναπαράσταση των προτύπων
2. Ορισμός ενός μέτρου εγγύτητας κατάλληλο για τα δεδομένα του πεδίου εφαρμογής
3. Συσταδοποίηση ή ομαδοποίηση
4. Αναπαράσταση των συστάδων (αν χρειάζεται)

5. Αξιολόγηση του αποτελέσματος (αν χρειαζεται)

Το παρακάτω σχήμα (σχήμα 3) δείχνει μια τυπική ακολουθία των πρώτων τριών βημάτων, όπου φαίνεται ότι το αποτέλεσμα της ομαδοποίησης μπορεί να δεχτεί και περαιτέρω επεξεργασία.



Σχήμα 3: Βήματα της συσταδοποίησης

Ως αναπαράσταση των προτύπων αναφέρεται στον αριθμό των τάξεων/κλάσεων, των αριθμό των διαθέσιμων προτύπων, και τον αριθμό, τον τύπο και την κλίμακα των χαρακτηριστικών που θα είναι διαθέσιμα στον αλγόριθμο συσταδοποίησης.

Η επιλογή χαρακτηριστικών είναι η διαδικασία όπου προσδιορίζεται το πιο αποτελεσματικό υποσύνολο του συνόλου των χαρακτηριστικών για να χρησιμοποιηθεί στην ομαδοποίηση. Η εξαγωγή δεδομένων είναι η χρήση ενός ή περισσότερων μετασχηματισμών πάνω στα χαρακτηριστικά για την παραγωγή νέων βασικών χαρακτηριστικών. Και οι δύο αυτές μέθοδοι μπορούν να χρησιμοποιηθούν για να αποκτηθεί το κατάλληλο σύνολο χαρακτηριστικών που θα χρησιμοποιηθεί στην ομαδοποίηση.

Η εγγύτητα προτύπων μετράται συνήθως με μία συνάρτηση απόστασης που όρίζεται σε ζεύγη προτύπων. Υπάρχει μία ποικιλία σε μέτρα απόστασης. Ένα απλό μέτρο απόστασης όπως η Ευκλείδια απόσταση μπορεί να χρησιμοποιηθεί για να αντικατοπτρίσει την ανομοιότητα μεταξύ δύο προτύπων, ενώ άλλα μέτρα ομοιότητας μπορούν να χρησιμοποιηθούν για να χαρακτηρίσουν την εννοιολογική ομοιότητα μετάξύ προτύπων.

Το βήμα της ομαδοποίησης μπορεί να πραγματοποιηθεί με πολλούς τρόπους. Η ομαδοποίηση που προκύπτει μπορεί να είναι μία αυστηρή διαμέριση των δεδομένων σε ομάδες ή ασαφής, όπου κάθε πρότυπο έχει ένα βαθμό συμμετοχής σε κάθε συστάδα.Οι αλγόριθμοι ιεραρχικής συσταδοποίησης παράγουν μια σειρά από εμφωλευμένες διαμερίσεις που βασίζονται σε ένα κριτήριο συγχώνευσης ή διαχωρισμού των συστάδων, που βασίζεται στην ομοιότητα. Οι διαμεριστικοί αλγόριθμοι συσταδο-ποίησης, προσδιορίζουν τη διαμέριση που βελτιστοποιεί ένα κριτήριο συσταδοποίησης. Άλλες τεχνικές για τη διαδικασία της ομαδοποίησης περιλαμβάνουν πιθανότητες

και γραφοθεωρητικές μεθόδους συσταδοποίησης.

Η αναπαράσταση των συστάδων είναι η διαδικασία της εξαγωγής μιας συμπαγούς αναπαράστασης ενός συνόλου δεδομένων. Η αναπαράσταση αυτή θα πρέπει να είναι απλή με την προοπτική να χρησιμοποιηθούν τα δεδομένα για περαιτέρω ανάλυση, ή να είναι εύκολα κατανοητά από τους ανθρώπους. Μια συμπαγής περιγραφή της κάθε συστάδας, συχνά με πρότυπες συστάδες ή με πρότυπα αντιπροσώπους όπως το centroid .

Η ανάλυση για την ισχύ των συστάδων, είναι μια εκτίμηση για το αποτέλεσμα της διαδικασίας συσταδοποίησης. Εκτιμήσεις για την εγγυρότητα γίνονται για να προσδιοριστεί αν το αποτέλεσμα, οι συστάδες, έχουν κάποιο νόημα. Μια δομή συστάδων είναι έγκυρη αν δεν έχει προκύψει κατά τύχει ή δεν είναι τεχνητό προϊόν του αλγορίθμου. Για τον έλεγχο-εκτίμηση της εγκυρότητας χρησιμοποιούνται στατιστικές μέθοδοι, και έλεγχοι υποθέσεων.

3.3 Ορισμός Συστάδας

Όταν αναφερόμαστε στη συσταδοποίηση πρέπει να ορίσουμε και τη συστάδα. Έχουν προταθεί πολλοί ορισμοί, εδώ δίνεται ορισμός των Θεοδωρίδη και Κουτρουμπά [5].
Έστω X ένα σύνολο δεδομένων, με $X = \{x_i, i = 1, ..., N\}$. Έστω η διαμέριση $\Re$, του X σε μ σύνολα, $C_j, j = 1, ..., m$. Αυτά τα σύνολα λέγονται συστάδες και πρέπει να πληρούν τις ακόλουθες τρεις συνθήκες:

- $C_i \neq \emptyset, i = 1, ..., m$
- $\bigcup_{i=1}^{m} C_i = X$
- $C_{ij} = \emptyset, i \neq j, i, j = 1, ..., m$

Είναι σημαντικό να πούμε οτι τα πρότυπα-αντικείμενα (διανύσματα) που περιέχονται σε μια συστάδα C_i είναι περισσότερα όμοια μεταξύ τους και λιγότερο όμοια με τα αυτά που βρίσκονται σε άλλες συστάδες. Η ομοιότητα αυτή είναι θεμελιώδης έννοια στον ορισμό μιας συστάδας, και ένα μέτρο για την ομοιότητα μεταξύ δύο προτύπων με τα ίδια χαρακτηριστικά είναι βασικό στις περισσότερες διαδικασίες συσταδοποίησης . Λόγω της πληθώρας των τύπων και των χαρακτηριστικών και της κλίμακας, το μέτρο απόστασης πρέπει να επιλεγεί προσεκτικά. Συνηθίζεται να υπολογίζεται η ανομοιότητα μεταξύ των προτύπων χρησιμοποιώντας ένα μέτρο απόστασης που ορίζεται στο χώρο των χαρακτηριστικών. Η πιο δημοφιλής μετρική είναι η Ευκλείδεια απόσταση.

3.3.1 Μέτρα ομοιότητας και ανομοιότητας

Τα μέτρα ομοιότητας χρησιμοποιούνται για να βρουν όμοια ζευγάρια προτύπων που περιέχονται στο X. Έστω $s(i, j)$ ο συντελεστής ομοιότητας. Αν τα πρότυπα i και j είναι παρόμοια, τότε ο $s(i, j)$ γίνεται μεγαλύτερος. Διαφορετικά, ο $s(i, j)$ γίνεται μικρότερος. Για όλα τα πρότυπα i και j, ένα μέτρο πρέπει να ικανοποιεί τις ακόλουθες συνθήκες:

- $0 \leq s(i, j) \leq 1$
- $s(i, i) = 1$
- $s(i, j) = s(j, i)$

Τα μέτρα ανομοιότητας χρησιμοποιούνται για να βρουν ανόμοια ζευγάρια προτύπων στο X. Ο συντελεστής ανομοιότητας, $d(i, j)$, είναι μικρός όταν τα πρότυπα i και j είναι όμοια, διαφορετικά, ο συντελεστής $d(i, j)$ μεγαλώνει. Όπως και τα μέτρα ομοιότητας, έτσι και τα μέτρα ανομοιότητας πρέπει να ικανοποιούν τις ακόλουθες συνθήκες:

- $0 \leq d(i, j) \leq 1$
- $d(i, i) = d$
- $s(i, j) = s(j, i)$

Οι περισσότεροι αλγόριθμοι συσταδοποίησης χρησιμοποιούν τα μέτρα ανομοιότητας για να ενώσουν, ή να διαχωρίσουν τα πρότυπα. Τα μέτρα που χρησιμοποιούνται πιο συχνά στην πράξη είναι δύο:

- Ευκλείδεια Απόσταση

Η Ευκλείδεια απόσταση μεταξύ των σημείων $x = (x_1, x_2, ..., x_n)$ και $y = (y_1, y_2, ..., y_n)$ δίνεται από:

$$d_1(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

όπου x_i και y_i είναι οι ι-οστες συντεταγμένες των x και y αντίστοιχα, και x και y είναι πρότυπα του X.

Η Ευκλείδεια απόσταση δουλεύει καλά όταν μια βάση δεδομένων έχει ¨συμπαγείσ' ή ¨απομονωμένεσ' συστάδες. Το μειονέκτημα στην απευθείας χρήση των μετρικών Μινκοωσκι είναι η τάση των χαρακτηριστικών με τη μεγαλύτερη κλίμακα να κυριαρχούν των άλλων. Η λύση σε αυτό το πρόβλημα περιλαμβάνει κανονικοποίηση των συνεχών χαρακτηριστικών σε ένα κοινό εύρος ή διασπορά.(Η γραμμική συσχέτιση μεταξύ των χαρακτηριστικών μπορεί να παραμορφώσει τα μέτρα απόστασης. Η παράμόρφωση μπορεί να μετριάσει με ένα μετασχηματισμό στα δεδομένα ή χρησιμοποιώντας την τετραγωνική απόσταση Mahalanobis)

- Απόσταση Manhattan

Η απόσταση Manhattan μεταξύ των προτύπων $x = (x_1, x_2, ..., x_n)$ και $y = (y_1, y_2, ..., y_n)$ δίνεται από:

$d_2(x, y) = \sum_{i=1}^{n} |x_i - y_i|$
όπου x_i και y_i είναι οι ι-οστες συντεταγμένες των x και y αντίστοιχα, και x και y είναι πρότυπα του X.

3.3.2 Απόσταση Συστάδων

Πολλοί αλγόριθμοι συσταδοποίησης απαιτούν τον ορισμό της απόστασης μεταξύ συστάδων (αντί της απόστασης μεταξύ των στοιχείων των συστάδων)[6]. Ο ορισμός αυτός δεν είναι εύκολη υπόθεση καθώς υπάρχουν πολλές ερμηνείες σχετικά με την απόσταση μεταξύ δύο συστάδων. Έστω λοιπόν C_i και C_j δύο συστάδες και έστω $|C_i|$ και $|C_j|$ ο αριθμός των προτύπων που περιέχει η κάθε συστάδα. Επίσης $d(C_i, C_j)$ είναι το μέτρο ανομοιότητας των συστάδων C_i και C_j αντίστοιχα και $d(i, j)$ είναι το μέτρο ανομοιότητας μεταξύ δύο προτύπων, όπου το i πρότυπο ανήκει στη συστάδα C_i και το j στη συστάδα C_j.

- Unweighted pair group Method using Arithmetic Averages (UPGMA)

Καλείται επίσης Group Average Method
Η UPGMA εισήχθει από τους Sokal kai Michener to 1958 [7] [8]. Οι απόσταση $|C_i|$ μεταξύ των συστάδων C_i ΚΑΙ C_j ως ο μέσος όρος όλων των ανομοιοτήτων $d(i, j)$. Αυτό σημαίνει ότι:
$\delta_1(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{i \in C_i, j \in C_j} d(i, j)$

- Απόσταση απλού συνδέσμου (Single Link-SLINK)

Στην SLINK που παρουσίασε ο Florek και άλλοι [8], η απόσταση μεταξύ δύο συστάδων ορίζεται ως η μικρότερη απόσταση μεταξύ ενός προτύπου της μιας συστάδας και ενός στοιχείου της άλλης. Έτσι, $\delta_2(C_i, C_j) = \min_{i \in C_i, j \in C_j} d(i, j)$

- Απόσταση πλήρους συνδέσμου (Complete Link-CLINK)

Η απόσταση CLINK είναι ακριβώς το αντίθετο της SLINK. Η CLINK ορίζεται ως:
$\delta_3(C_i, C_j) = \max_{i \in C_i, j \in C_j} d(i, j)$

- Απόσταση των κέντρων βάρους

Αν οι συστάδες έχουν αντιπροσωπευτικά κέντρα βάρους, τότε η απόσταση μεταξύ των συστάδων ορίζεται ως η απόσταση μεταξύ των κέντρων βάρους. Έτσι αν K_i και K_j είναι τα κέντρα βάρους των συστάδων C_i και C_j αντίστοιχα, τότε
$\delta_4(C_i, C_j) = d(K_i, K_j)$

3.4 Τεχνικές Συσταδοποίησης

Με τη βοήθεια της ιεραρχίας του σχήματος 4 [4], μπορούν να περιγραφούν διαφορετικές προσεγγίσεις στη συσταδοποίηση δεδομένων. Στο πρώτο επίπεδο υπάρχει μια διάκριση μεταξύ ιεραρχικών και διαμεριστικών προσεγγίσεων. Οι ιεραρχικές μέθοδοι παράγουν μία εμφωλευμένη σειρά διαμερίσεων, ενώ οι διαμεριστικές μέθοδοι παράγουν μόνο μία. Υπάρχουν όμως σημαντικά θέματα που επηρεάζουν τις διαφορετικές προσεγγίσεις άσχετα με τη θέση τους στην ταξινομία.

3.4.1 Συσσωρευτικοι-Διαιρετικοί

Αυτή η πλευρά σχετίζεται με τη δομή του αλγορίθμου και τη λειτουργία. Μια συσσωρευτική προσέγγιση ξεκινά με κάθε πρότυπο σε μια διαφορετική συστάδα και διαδοχικά συγχωνεύει συστάδες μέχρι κάποιο κριτήριο τερματισμού να ικανοποιηθεί. Μια διαιρετική μέθοδος αρχίζει με όλα τα πρότυπα σε μια μοναδική συστάδα, και εκτελεί διαχωρισμούς μέχρι κάποιο κριτήριο.

3.4.2 Μονοθετικοί-Πολυθετικοί

Αυτό σχετίζεται με την ακολουθιακή ή ταυτόχρονη χρήση των χαρακτηριστικών στη διαδικασία της συσταδοποίησης. Οι περισσότεροι αλγόριθμοι είναι πολυθετικοί: αυτό σημαίνει ότι όλα τα χαρακτηριστικά εισάγονται στον υπολογισμό των αποστάσεων μεταξύ των προτύπων, και οι αποφάσεις βασίζονται σε αυτές τις αποστάσεις.

3.4.3 Ασαφή και μη ασαφή

Ένας μη ασαφή αλγόριθμος συσταδοποίησης τοποθετεί κάθε πρότυπο σε μία και μοναδική συστάδα. Ένας ασαφής όμως αλγόριθμος θέτει βαθμούς συμμετοχής σε διάφορες συστάδες στο κάθε πρότυπο. Μια σαφής συσταδοποίηση μπορεί να γίνει σαφής αν τοποθετησουμε κάθε πρότυπο μόνο στη συστάδα που έχει το μεγαλύτερο βαθμό συμμετοχής.

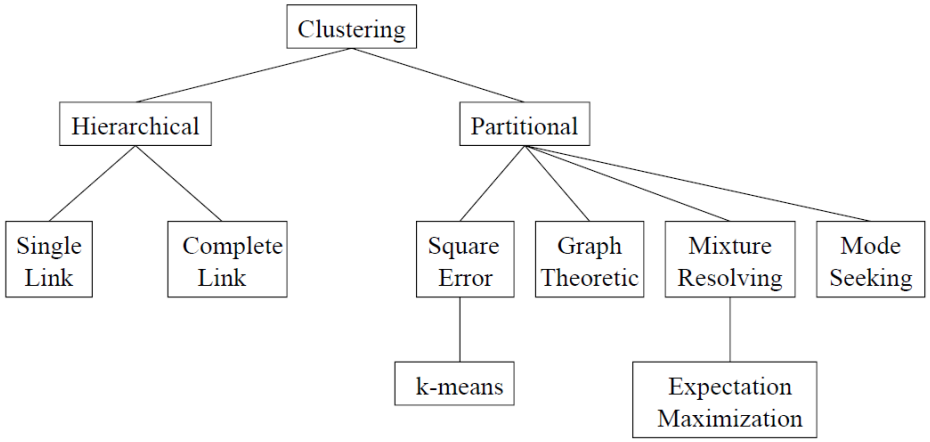
3.4.4 Ντετερμινιστικοί-Στοχαστικοί

Αυτό σχετίζεται πιο πολύ με τις διαμεριστικές προσεγγίσεις όπου σχεδιάζονται έτσι ώστε να βελτιστοποιείται η συνάρτηση του τετραγωνικού σφάλματος. Η βελτιστοποίηση αυτή μπορεί

να πραγματοποιηθεί χρησιμοποιώντας παραδοσιακές τεχνικές ή μέσω τυχαίας αναζήτησης του χώρου καταστάσεων που αποτελείται από όλες τις πιθανές ετικέτες.

3.4.5 Αυξητικοί-Μη αυξητικοί

Αυτό το θέμα προκύπτει όταν το σύνολο των προτύπων προς συσταδοποίηση είναι τεράστιο, και περιορισμοί στο χρόνο εκτέλεσης και στο χώρο της μνήμης επιρεάζουν την αρχιτεκτονική του αλγορίθμου.



Σχήμα 4: Μία ταξινόμηση των αλγορίθμων συσταδοποίησης

4 ΑΛΓΟΡΙΘΜΟΙ ΣΥΣΤΑΔΟΠΟΙΗΣΗΣ

4.1 Ιεραρχικοί αλγόριθμοι συσταδοποίησης

Οι αλγόριθμοι ιεραρχικής συσταδοποίησης οργανώνουν τα δεδομένα σε μια ιεραρχική δομή σύμφωνα με ένα πίνακα εγγύτητας. Τα αποτελέσματα ενός ιεραρχικού αλγορίθμου παρουσιάζονται συνήθως με ένα δυαδικό δέντρο ή ένα δεντρόγραμμα. Ο κόμβος ρίζα του δεντρογράμματος αναπαριστά όλο το σύνολο των δεδομένων και κάθε κόμβος φύλλο αναπαριστά ένα πρότυποδεδομένο. Οι ενδιάμεσοι κόμβοι, περιγράφουν πόσο κοντά είναι μεταξύ τους τα πρότυπα, και το ύψος του δεντρογράμματος εκφράζει συνήθως την απόσταση ενός ζευγαριού προτύπων ή συστάδων. Το τελικό αποτέλεσμα της συσταδοποίησης μπορεί να αποκτηθεί κόβοντας το δεντρόγραμμα σε διαφορετικά επίπεδα. Αυτή η αναπαράσταση παρέχει πληροφοριακή περιγραφή και απεικόνιση για τη δομή της συσταδοποίησης των δεδομένων, ειδικά όταν υπάρχουν πραγματικές ιεραρχικές σχέσεις στα δεδομένα, όπως δεδομένα από εξελικτική έρευνα σε διαφορετικά είδη οργανισμών.

Οι ιεραρχικοί αλγόριθμοι ταξινομούνται σε συσσωρευτικούς και διαιρετικούς. Η συσσωρευτική συσταδοποίηση ξεκινά με N συστάδες και κάθε μία από αυτές περιέχει ακριβώς ένα πρότυπο. Μια σειρά από διαδοχικές συγχωνεύσεις οδηγεί τελικά όλα τα πρότυπα στην ίδια συστάδα. Οι διαιρετικοί αλγόριθμοι λειτουργούν με τον αντίθετο τρόπο. Στην αρχή όλα τα πρότυπα ανήκουν στην ίδια συστάδα, και με μια διαδικασία διαχωρισμού όλα τα πρότυπα καταλήγουν να αποτελούν μια συστάδα. Για μια συστάδα με N αντικείμενα υπάρχουν $2^{N-1}-1$ πιθανές υποδιαιρέσεις της συστάδας σε δύο υποσυστάδες, το οποίο είναι πολύ απαιτεί πολλούς υπολογισμούς [16]. Επομένως οι διαιρετικοί αλγόριθμοι δεν χρησιμοποιούνται συχνά στην πράξη. Δύο γνωστοί διαιρετικοί αλγόριθμοι είναι οι MONA και DIANA και περιγράφονται στο [17].

Επειδή υπάρχουν διαφορετικοί ορισμοί για την απόσταση δύο συστάδων, υπάρχουν πολλοί διαφορετικοί συσσωρευτικοί αλγόριθμοι. Οι πιο απλοί και πιο γνωστοί αλγόριθμοι περιλαμβάνουν τεχνικές απλού συνδέσμου[54] και πλήρους συνδέσμου[55]. Στη μέθοδο απλού συνδέσμου, η απόσταση μεταξύ δύο συστάδων προσδιορίζεται από τα πιο κοντινά πρότυπα μεταξύ των δύο συστάδων , γιαυτό καλείται επίσης και μέθοδος του πλησιέστερου γείτονα. Η μέθοδος του πλήρους συνδέσμου χρησιμοποιεί τη μεγαλύτερη απόσταση μεταξύ ενός ζευγαριού προτύπων για να ορίσει την απόσταση των συστάδων. Πιο περίπλοκοι συσσωρευτικοί αλγόριθμοι συσταδοποίησης, περιλαμβάνουν group average linkage, median linkage, centroid linkage, και τη μέθοδο του Ward. Οι μέθοδοι απλού συνδέσμου, πλήρους συνδέσμου και μέσου συνδέσμου

λαμβάνουν υπόψη όλα τα σημεία δύο συστάδων για να υπολογίσουν την απόσταση μεταξύ τους, και καλούνται γραφοθεωρητικές μέθοδοι. Οι υπόλοιπες μέθοδοι καλούνται γεωμετρικές επειδή χρησιμοποιούν τα γεωμετρικά κέντρα για να αναπαράστίσουν τις συστάδες και να προσδιορίσουν τις αποστάσεις μεταξύ τους. Χαρακτηριστικά και ιδιότητες αυτών των μεθόδων συνοψίζονται στο [16].

Οι ιεραρχικοί αλγόριθμοι είναι ευαίσθητοι στο θόρυβο και στις ακραίες τιμές. Όταν ένα πρότυπο καταχωρηθεί σε μια συστάδα, δεν λαμβάνεται πάλι υπόψη, το οποίο σημαίνει ότι οι ιεραρχικοί αλγόριθμοι δεν είναι ικανοί να διορθώσουν προηγούμενα λάθη στην ταξινόμηση. Η υπολογιστική πολυπλοκότητα των ιεραρχικών αλγορίθμων είναι $O(N^2)$. Ένα ακόμα μειονέκτημα των ιεραρχικών αλγορίθμων είναι η τάση να σχηματίζουν σφαιρικές συστάδες. Τα τελευταία χρόνια, η ανάγκη χειρισμού πολύ μεγάλων συνόλων δεδομένων στην εξόρυξη γνώσης από δεδομένα και σε άλλους τομείς, είχε σαν αποτέλεσμα να εμφανιστούν νέες τεχνικές ιεραρχικής συσταδοποίησης με αρκετα βελτιωμένη επίδοση. Τυπικά παραδείγματα είναι οι CURE[18], ROCK[19], CHAMELEON[20] kai BIRCH[21].

Ο αλγόριθμος BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) κατασκευάστηκε για τη συσταδοποίηση μεγάλου πλήθους μετρικών δεδομένων. Ο αλγόριθμος θεωρεί ότι μπορεί να υπάρχει περιορισμένη κύρια μνήμη και επιτυγχάνει γραμμικό χρόνο I/O απαιτώντας μία μόνο σάρωση της βάσης. Ο αλγόριθμος είναι αυξητικός και χρησιμοποιεί και μία τεχνική για την αντιμετώπιση του προβλήματος των ακραίων σημείων. Τα σημεία που εντοπίζονται σε 'αραιοκατοικημένες περιοχές ' απομακρύνονται. Η βασική ιδέα του αλγορίθμου είναι η κατασκευή ενός δέντρου που διατηρεί όλη την πληροφορία για τη διενέργεια της συσταδοποίησης. Η συσταδοποίηση εκτελείται στο ίδιο το δέντρο και οι ετικέτες των κόμβων του δέντρου περιέχουν την απαραίτητη πληροφορία για τον υπολογισμό των αποστάσεων. ένα κύριο χαρακτηριστικό του αλγορίθμου BIRCH είναι η χρήση του χαρακτηριστικού συσταδοποίησης (clustering feature), μιας τριάδας που περιέχει πληροφορία για τη συστάδα. Το χαρακτηριστικό συσταδοποίησης μιας συστάδας αποτελεί μια περίληψη της πληροφορίας της συστάδας. Ο αλγόριθμος BIRCH εφαρμόζεται μόνο σε αριθμητικά δεδομένα.

Οι Guha, Rastogi και Shim παρατήρησαν ότι με τους ιεραρχικούς αλγόριθμους που βασίζονταν στο κέντρο βάρους, δεν ήταν δυνατό να δημιουργηθούν τυχαία σχήματα συστάδων. Έτσι δημιούργησαν τον ιεραρχικό αλγόριθμο CURE , για να ερευνήσουν πιο εξεζητημένα σχήματα συστάδων. Το κρίσιμο χαρακτηριστικό του CURE είναι η χρήση πιο καλά διασκορπισμένων σημείων για να αντιπροσωπεύουν κάθε συστάδα, το οποίο δίνει τη δυνατότητα να βρεθούν και

άλλα σχήματα συστάδων εκτός από υπερσφαίρες και αποφεύγεται επίσης το φαινόμενο της αλυσίδας της μεθόδου ελάχιστου συνδέσμου [16]. Ο αλγόριθμος CURE αποτελείται από ένα ιεραρχικό και ένα διαμεριστικό τμήμα. Αρχικά, από κάθε συστάδα επιλέγεται ένα σταθερό σύνολο σημείων. Αυτά τα καλά διαχωρισμένα σημεία συρρικνώνονται στη συνέχεια κοντά στο κέντρο βάρους της συστάδας εφαρμόζοντας έναν παράγοντα συρρίκνωσης α. Όταν το α ισούται με 1 όλα τα σημεία αυτά αναπαριστούν μία συστάδα καλύτερα από ότι θα την αναπαριστούσε ένα μόνο σημείο (όπως για παράδειγμα, το κέντρο βάρους ή το medoid). Χρησιμοποιώντας πολλαπλά αντιπροσωπευτικά σημεία, μπορούν να αναπαρασταθούν καλύτερα συστάδες ασυνήθιστου σχήματος (όχι οι απλές σφαίρες). Επιπλέον ο αλγόριθμος CURE χρησιμοποιεί έναν ιεραρχικό αλγόριθμο συσταδοποίησης. Σε κάθε βήμα του συσσωρευτικού αλγορίθμου, επιλέγονται για συγχώνευση οι συστάδες με τα πλησιέστερα ζεύγη αντιπροσωπευτικών συνόλων των δύο συστάδων. Ο Guha κ.α. πρότειναν επίσης έναν ακόμα συσσωρευτικό ιεραρχικό αλγόριθμο, τον ROCK, για να ομαδοποιήσουν δεδομένα με ποιοτικά γνωρίσματα[19].

4.2 Διαμεριστικοί αλγόριθμοι συσταδοποίησης

Σε αντίθεση με τους ιεραρχικούς αλγόριθμους, οι οποίοι πετυχαίνουν διάφορετικά επίπεδα συστάδων με επαναληπτικές συγχωνεύσεις ή διαμερίσεις, οι διαμεριστικοί δημιουργούν τις συστάδες σε ένα βήμα, χωρίς ιεραρχική δομή. Δεδομένου ότι η έξοδος αποτελείται από ένα μόνο σύνολο συστάδων , ο χρήστης θα πρέπει να δώσει ως είσοδο το επιθυμητό πλήθος συστάδων. Επιπλέον χρησιμοποιούνται μέτρα ποιότητας (μετρικές, συναρτήσεις κριτηρίων) για τον προσδιορισμό της καταλληλότητας των προτεινόμενων λύσεων. Ένα τέτοιο μέτρο ποιότητας θα μπορούσε να είναι η μέση απόσταση μεταξύ των συστάδων ή κάποια άλλη μετρική. Η λύση που τελικά χρησιμοποιείται είναι αυτή που έχει την καλύτερη τιμή για τη συνάρτηση κριτηρίου. Ένα συνηθισμένο μέτρο είναι μία μετρική τετραγωνικού σφάλματος, η οποία μετράει την τετραγωνική απόσταση των σημείων της συστάδας από το κέντρο της συστάδας.
Η πολυπλοκότητα των διαμεριστικών αλγορίθμων μπορεί να εκτοξευτεί υψηλά λόγω του πλήθους των πιθανών λύσεων[45]. Ακόμα και για μικρού μεγέθους προβλήματα συσταδοποίησησ(ομαδοποίηση 30 προτύπων σε 3 ομάδες), ο αριθμός των πιθανών διαμερίσεων είναι $2 * 10^4$. Γιαυτό το λόγο αναπτύχθηκαν προσεγγιστικοί αλγόριθμοι για τον εντοπισμό προσσεγιστικών λύσεων. Το κριτήριο τετραγωνικού σφάλματος ορίζεται ως:

$J(G, M) = \sum_{i=1}^{K} \sum_{j=1}^{N} g_{ij} \left\| x_j - m_i \right\|^2$

όπου Γ: ένας πίνακας διαμέρισης με στοιχεία
g_{ij}

όπου
$g_{ij} = 1$, αν $x_j \in clusteri$, διαφορετικά $g_{ij} = 0$με $\sum_{i=1}^{K} g_{ij} = 1 \forall j$
Μ Είναι ο πίνακας των μέσων των συστάδων
$M = [m_1, ..., m_K]$, m_i είναι ο δειγματικός μέσος για την ιοστή συστάδα.
$(1/N_i) \sum_{j=1}^{N} g_{ij} \xi_j$
N_i είναι ο αριθμός των προτύπων στην ιοστή συστάδα.

Ο αλγόριθμος K-means είναι ο πιο γνωστός αλγόριθμος συσταδοποίησης τετραγωνικού σφάλματος[56],[57].

1) Αρχικοποιείται μία τυχαία K-διαμέριση ή μία διαμέριση που βασίζεται σε προγενέστερη γνώ-

ση. Υπολογίζεται ο πίνακας $M = [m_1, ..., m_K]$.

2) Κάθε πρότυπο τοποθετείται στη πλησιέστερη συστάδα C_w, έτσι ώστε
$x_j \in C_w$, αν $\|x_j - m_w\| < \|x_j - m_i\|$ για $j = 1, ..., N, i \neq w$, και $i = 1, ..., K$.

3)Υπολογίζεται πάλι ο πίνακας M βάση της τελευταίας διαμέρισης.

4) Επαναλαμβάνονται τα βήματα 2 και 3 μέχρις ότου να μην υπάρχει αλλαγή σε κάποια συστάδα.

Ο αλγόριθμος K-means είναι πολύ απλός και μπορεί εύκολα να υλοποιηθεί για να λύσει πολλά πρακτικά προβλήματα. Έχει καλά αποτελέσματα σε συμπαγής και σφαιρικές συστάδες. Η πολυπλοκότητα χρόνου του K-means είναι O(NKd). Επειδή το K και το d είναι συνήθως πολύ μικρότερα του N , ο αλγόριθμος K-means μπορεί να χρησιμοποιηθεί για να τη συσταδοποίηση πολύ μεγάλων βάσεων δεδομένων. Παράλληλες τεχνικές για τον K-means έχουν αναπτυχθεί όπου μπορούν να επιταχύνουν αρκετά τον αλγόριθμο[67]. Ο K-means έχει όμως αρκετά μειονεκτήματα, με αποτέλεσμα να αναπτυχθούν πολλές παραλλαγές του, με σκοπό να ξεπεραστούν τα εμπόδια.
Δεν υπάρχει αποδοτική και καθολική μέθοδος για τον προσδιορισμό των αρχικών διαμερίσεων και του αριθμού των συστάδων K. Υπάρχουν αλγόριθμοι, παραλλαγές του K-means, που εξετάζουν τρόπους βελτίωσης των πιθανοτήτων για την εύρεση του ολικού βέλτιστου. Συνήθως αυτό επιτυγχάνεται επιλέγοντας προσεκτικά τις αρχικές συστάδες και τους μέσους. Μια άλλη παραλλαγή, ο αλγόριθμος ISODATA[24] επιτρέπει στις συστάδες να διασπώνται και να συγχωνεύονται. Στην περίπτωση αυτή εξετάζεται η στατιστική διακύμανση (variance) της συστάδας και, αν αυτή είναι πολύ μεγάλη, η συστάδα διασπάται. Κατάνάλογο τρόπο, αν η απόσταση μεταξύ των κέντρων βάρους δύο συστάδων είναι μικρότερη κάποιου προκαθορισμένου κατωφλίου, οι συστάδες συγχωνεύονται.
Ο αλγόριθμος K-means βρίσκει ένα τοπικό βέλτιστο και μπορεί στην πραγματικότητα να χάσει το ολικό βέλτιστο. Ο K-means δε δουλεύει με μη αριθμητικά δεδομένα επειδή ο μέσος θα πρέπει να οριστεί στον τύπο του γνωρίσματος. Επιπλέον, ο K-means παράγει συστάδες κυρτού σχήματος. Επίσης δεν χειρίζεται καλά τα ακραία σημεία. Η λειτουργία διαχωρισμού του αλ-

γορίθμου ISODATA απαλείφει την δυνατότητα απομακρυσμένων συστάδων, ενώ ο αλγόριθμος PAM[17] αναπαριστά κάθε συστάδα μέσω ενός medoid. Η χρήση Medoid ενδείκνυται για την αντιμετώπιση του προβλήματος των απομονομένων σημείων. Ο K-modes, μια παραλλαγή του K-means, χειρίζεται μη αριθμητικά δεδομένα. Ο αλγόριθμος αυτός χρησιμοποιεί τις επικρατούσες τιμές(modes) αντί τους μέσους. Οι τυπικές τιμές για το K κυμαίνονται μεταξύ του 2 και του 10. Για την επέκταση του K-means σε κατηγορικά δεδομένα έχουν προταθεί διαφορετικά μέτρα ανομοιότητας[58],[26]. Παρόλο που ο αλγόριθμος K-means παράγει συχνά καλά αποτελέσματα, δεν είναι αποδοτικός από άποψη χρόνου και δεν έχει καλή κλιμάκωση. Το πλήθος των υπολογισμών για τις αποστάσεις μπορεί να μειωθεί, αν μεταξύ των επαναλήψεων αποθηκεύουμε τις πληροφορίες που σχετίζονται με τις αποστάσεις.[βιβλιογραφία] Τελευταίες εξελίξεις και βελτιώσεις του K-means και άλλων αλγορίθμων τετραγωνικού σφάλματος με τις εφαρμογές τους μπορούν να βρεθούν στα [27],[28],[29],[30],[31],[32].

4.3 Συσταδοποίηση που βασίζεται σε πυκνότητα πιθανότητας

Από την πλευρά των πιθανοτήτων, τα πρότυπα-δεδομένα δημιουργούνται σύμφωνα με κάποιες κατανομές πιθανοτήτων. Μπορούν να παραχθούν από διαφορετικούς τύπους συναρτήσεων πυκνότητας (π.χ. πολυμεταβλητής Gauss ή κατανομή t), ή από τις ίδιες οικογένειες αλλά με διαφορετικές παραμέτρους. Αν οι κατανομές είναι γνωστές τότε, ο εντοπισμός των συστάδων ενός δοσμένου συνόλου προτύπων ισοδυναμεί με την εκτίμηση των παραμέτρων των υποκείμενων μοντέλων. Αν υποθέσουμε ότι η εκ των προτέρων πιθανότητα $P(C_i)$ για τη συστάδα $C_i, i = 1, ..., k$ και η υπό συνθήκη πυκνότητα πιθανότητας $p(x|C_i, j_i)$, όπου j_i είναι το διάνυσμα της άγνωστης παραμέτρου, είναι γνωστά. Τότε η μεικτή πυκνότητα πιθανότητας για όλο το σύνολο δεδομένων εκφράζεται ως
$p(x|j) = \sum_{i=1}^{k} p(x|C_i, j_i)P(C_i)$,
όπου $j = (j_1, ..., j_k)$ και
$\sum_{i=1}^{k} P(C_i = 1.$

Μόλις βρεθεί υπολογιστεί το διάνυσμα της άγνωστης παραμέτρου j, η εκ των υστέρων πιθανότητα για την ανάθεση ενός σημείου σε μια συστάδα μπορεί εύκολα να υπολογιστεί με το θεώρημα του Bayes. Έτσι, οι mixtures μπορούν να κατασκευαστούν για οποιαδήποτε τύπο components, αλλά πιο συχνά χρησιμοποιούνται πυκνότητες πολυμεταβλητής Gauss, λόγω της πιο ολοκληρωμένης θεωρίας [16],[33]. Η εκτίμηση μέγιστης πιθανοφάνειας είναι μια σημαντική στατιστική προσέγγιση για την εκτίμηση των παραμέτρων [34] και λαμβάνει υπόψη την καλύτερη εκτίμηση ως αυτή που μεγιστοποιεί την πιθανότητα να δημιουργεί όλες τις παρατηρήσεις που δίνεται από τη συνάρτηση πυκνότητας

$p(x_1, ..., x_N)|j) =_{j=1}^{N} p(x_j|j)$, ή στη λογαριθμική μορφή

$l(j) = \sum_{j=1}^{N} lnp(x_j|J.$

Η καλύτερη εκτίμηση μπορεί να επιτευχθεί λύνοντας τις εξισώσεις $(\partial l(j))/(\partial j_i) = 0$.

Δυστυχώς, επειδή οι λύσεις των εξισώσεων αυτών δεν βρίσκονται αναλυτικά στις περισσότερες πριπτώσεις [36],[37],εφαρμόζονται επαναληπτικές μέθοδοι για την προσέγγιση των εκτίμήσεων μέγιστης πιθανοφάνειας. Η πιο γνωστή μέθοδος άπό αυτές είναι ο αλγόριθμος expectation

maximization(EM)[38]. Το κυριότερο μειονέκτημα του EM η ευαισθησία στην επιλογή των αρχικών παραμέτρων, το φαινόμενο του ιδιάζοντος πίνακα συνδιασπορών, η πιθανότητα σύγκλισης σε ένα τοπικό βέλτιστο, και ο αργός ρυθμός σύγκλισης [35],

- Η διέγερση των νευρώνων επηρεάζει και τη διέγερση άλλων νευρώνων που βρίσκονται κοντά του.

- Οι νευρώνες που βρίσκονται σε μεγάλες μεταξύ τους αποστάσεις φαίνεται να αλληλο αναχαιτίζονται.

- Οι νευρώνες φαίνεται να έχουν συγκεκριμένες διακριτές μεταξύ τους λειτουργίες.

Ο όρος αυτο-οργανώμενος δείχνει την ικανότητα αυτών των νευρωνικών δικτύων να οργανώνουν τους κόμβους σε συστάδες βάσει της μεταξύ τους ομοιότητας. Οι κόμβοι που βρίσκονται πιο κοντά μεταξύ τους έχουν μεγαλύτερη ομοιότητα απ'οτι οι κόμβοι που βρίσκονται μακριά. Αυτό αποτελεί και ένδειξη του πώς εκτελείται η πραγματική συσταδοποίηση. Με την πάροδο του χρόνου, οι κόμβοι του επιπέδου εξόδου ταιριάζουν με τους κόμβους του επιπέδου εισόδου και αναδύονται τα πρότυπα κόμβων του επιπέδου εξόδου. Το πιο γνωστό παράδειγμα SOFM είναι ο αυτο-οργανωμένος χάρτης Kohonen(Kohonen self-organizing map). Υπάρχει ένα επίπεδο εισόδου και ένα ειδικό επίπεδο, που παράγει τιμές εξόδου που συναγωνίζονται μεταξύ τους. Ως αποτέλεσμα δημιουργούνται πολλαπλές έξοδοι και επιλέγεται η καλύτερη. Αυτό το επιπλέον επίπεδο τεχνικά δεν είναι ούτε κρυφό επίπεδο ούτε επίπεδο εξόδου, και έτσι αναφερόμαστε σε αυτό ως το ανταγωνιστικό επίπεδο. Οι κόμβοι αυτού του επιπέδου θεωρούνται ως 2-διάστατα πλέγματα κόμβων όπως φαίνεται στο σχήμα. Κάθε κόμβος εισόδου συνδέεται με κάθε κόμβο του πλέγματος. Η διάδοση εμφανίζεται με την αποστολή της τιμής εισόδου κάθε κόμβου εισόδου σε κάθε κόμβο του ανταγωνιστικού επιπέδου. Όπως και στα κανονικά νευρωνικά δίκτυα, κάθε ακμή σχετίζεται με ένα βάρος και κάθε βάρος του ανταγωιστικού επιπέδου έχει μία συνάρτηση ενεργοποίησης. Έτσι, κάθε κόμβος του ανταγωνιστικού επιπέδου παράγει μια τιμή εξόδου, ο κόμβος με την καλύτερη έξοδο κερδίζει τον ανταγωνισμό και ορίζεται να είναι η έξοδος για τη συγκεκριμένη είσοδο. Ένα ελκυστικό χαρακτηριστικό των δικτύων Kohonen είναι ότι τα δεδομένα μπορούν να τροφοδοτήσουν παράλληλα τους πολλαπλούς ανταγωνιστικούς κόμβους. Η εκπαίδευση λαμβάνει χώρα ρυθμίζοντας τα βάρη έτσι ώστε η καλύτερη έξοδος να είναι ακόμη καλύτερη την επόμενη φορά που θα χρησιμοποιηθεί αυτή η είσοδος. Η έννοια "καλύτερη" ορίζεται υπολογίζοντας ένα μέτρο απόστασης.

Μια συνηθισμένη προσέγγιση είναιη αρχικοποίηση των βαρών των ακμών εισόδου του ανταγωνιστικού επιπέδου με κανονικοποιημένες τιμές. Η ομοιότητα μεταξύ των κόμβων εξόδου και των διανυσμάτων εισόδου ορίζεται τότε ως το εσωτερικό γινόμενο των δύο διανυσμάτων. Δοθείσης μιας πλειάδας εισόδου $X = x_1, ..., x_h$ και των βαρών των ακμών που αποτελούν εισόδους σε κάποιο ανταγωνιστικό κόμβο i ως $w_{1i}, ..., w_{hi}$, η ομοιότητα μεταξύ των X και i μπορεί να υπολογιστεί ως εξής:

$sim(X, i) = \sum_{j=1}^{h} x_j w_{ji}$

Ο ανταγωνιστικός κόμβος που έχει τη μεγαλύτερη ομοιότητα με τον κόμβο εισόδου κερδί-

ζει τον ανταγωνισμό. Βάσει αυτού, αυξάνονται τα βάρη που καταλήγουν στον κόμβο i, όπως επίσης και αυτά των κόμβων που τον περιβάλλουν άμεσα στη μήτρα. Αυτή είναι η φάση μάθησης. Δοθέντος ενός κόμβου i, χρησιμοποιούμε τον συμβολισμό N_i για να αναπαραστήσουμε την ένωση του i με τους κοντινούς του κόμβους στη μήτρα. Έτσι, η διαδικασία μάθησης χρησιμοποιεί τον ακόλουθο τύπο:
$Dw_{kj} = c(x_k - w_{kj}$ εάν $j \in N_i$ και $Dw_{kj} = 0$ διαφορετικά.

Στον τύπο αυτό, το c αναφέρεται στο ρυθμό μάθησης και στην πραγματικότητα μπορεί να μεταβάλλεται βάσει του κόμβου παρά να αποτελεί σταθερά. Η βασική ιδέα της μάθησης SOM είναι ότι μετά την είσοδο κάθε πλειάδας του συνόου εκπαίδευσης, τα βάρη του νικητή κόμβου και των γειτόνων του αλλάζουν ώστε να είναι πιο κοντά στην πλειάδα. Με την πάροδο του χρόνου αναδύεται ένα πρότυπο στους κόμβους εξόδου το οποίο είναι κοντα σάυτό του συνόλου εκπαίδευσης. Στην αρχή της διαδικασίας εκπαίδευσης, μπορεί να οριστεί αρκετά μεγάλη γειτονιά ενός κόμβου. Ωστόσο, η γειτονιά αυτή μπορεί να μειωθεί κατά τη διάρκεια της επεξεργασίας.

4.4 Συσταδοποίηση σε μεγάλες βάσεις δεδομένων

Οι αλγόριθμοι συσταδοποίησης που παρουσιάστηκαν προηγουμένως περιλαμβάνουν μερικούς από τους κλασικούς αλγόριθμους συσταδοποίησης. Ωστόσο, οι αλγόριθμοι αυτοί ενδέχεται να είναι ακατάλληλοι στην περίπτωση των δυναμικών βάσεων δεδομένων. Καταρχήν, επειδή οι περισσότεροι έχουν πολυπλοκότητα $O(n^2)$, θεωρούν ότι υπάρχει επαρκής μνήμη για την αποθήκευση των προς συσταδοποίηση δεδομένων και των δομών δεδομένων που χρειάζονται για την υποστηριξή τους. Δεδομένου όμως ότι οι μεγάλες βάσεις δεδομένων που χρειάζονται για την υποστήριξή τους. Δεδομένου όμως ότι οι μεγάλες βάσεις δεδομένων περιέχουν χιλιάδσ στοιχεία (ή και περισσότερα) οι παραδοχές αυτές δεν είναι ρεαλιστικές. Επιπλέον, η εκτέλεση I/O λειτουργιών είναι πολύ ακριβή λόγω της επαναληπτικής φύσης των αλγορίθμων. Εξαιτίας αυτών των περιορισμών της κύριας μνήμης, οι αλγόριθμοι δεν έχουν κλιμάκωση σε μεγάλες βάσεις δεδομένων. Οι τεχνικές συσταδοποίησης θα πρέπει να μπορούν να προσαρμόζονται καθώς η βάση δεδομένων αλλάζει. Οι κλασικοί ιεραρχικοί αλγόριθμοι συσταδοποίησης δεν είναι κατάλληλοι για σύνολα δεδομένων μεγάλης κλίμακας λόγω της τετραγωνικής πολυπλοκότητας χρόνου και χώρου. Ο αλγόριθμος K-means έχει πολυπλοκότητα χώρου $O(N + K)$. Επειδή το N είναι συνήθως πολύ μεγαλύτερο από το K και το d, η πολυπλοκότητα γίνεται σχεδόν γραμμική στον αριθμό των προτύπων. Ο αλγόριθμος K-means είναι αποτελεσματικός στη συσταδοποίηση μεγάλων βάσεων δεδομένων, και έχουν γίνει προσπάθειες για να ξεπεραστούν τα μειονεκτήματά του. [58],[59] . Για τη συσταδοποίηση μεγάλων βάσεων δεδομένων έχουν αναπτυχθεί πολλοί προτότυποι αλγόριθμοι [60][61][63][62][64][65]. Πολλοί από αυτούς κλιμακώνουν την υπολογιστική πολυπλοκότητα γραμμικά στο μέγεθος εισόδου και επιδεικνύουν τη δυνατότητα να χειρίζονται πάρα πολύ μεγάλες βάσεις δεδομένων. Προσέγγιση τυχαίας δειγματοληψίας, για παράδειγμα οι αλγόριθμοι CLARA[17] και CURE[18]. Το κλειδί είναι ότι το κατάλληλο μέγεθος μπορεί αποτελεσματικά να περιέχει τις σημαντικές γεωμετρικές ιδιότητες των συστάδων. Επιπλέον τα φράγματα Chernoff παρέχουν εκτίμηση για το κάτω φράγμα του ελαχίστου μεγέθους δείγματος, δοθείσης της χαμηλής πιθανότητας των σημείων σε κάθε συστάδα που χάνονται στο δειγματικό σύνολο [18]. Ο CLARA αναπαριστά κάθε συστάδα με ένα medoid και ο CURE επιλέγει συνολο από καλά διασκορπισμένα και κεντρικά συμπιεσμένα σημεία. Προσέγγιση τυχαίας αναζήτησης π.χ. η συσταδοποίηση μεγάλων εφαρμογών βασίζεται σε τυχαία αναζήτηση (CLARANS)[64].Ο CLARANS βλέπει τη συσταδοποίηση ως μια διαδικασία αναζήτησης σε ένα γράφο, στον οποίο κάθε κόμβος αντιστοιχεί σε ένα σύνολο K medoids . Αρχίζει με ένα τυχαίο κόμβο,(τρέχον κόμβος) και εξετάζει ένα σύνολο γειτόνων, ορισμένος ως ο κόμβος που αποτελείται από μόνο

ένα διαφορετικό πρότυπο, να αναζητήσει μια καλύτερη λύση, π.χ. ο οποιοσδήποτε γείτονας, με χαμηλότερο κόστος, γίνεται ο τρέχον κόμβος. Αν ο μέγιστος αριθμός γειτόνων, που ορίζεται από το χρήστη έχει επιτευχθεί, ο τρέχον κόμβος, γίνεται κόμβος νικητής. Αυτή η διαδικασία επαναλαμβάνεται μερικές φορές όπως ορίζεται από τους χρήστες. Παρόλο που ο CLARANS πετυχαίνει καλύτερη απόδοση από αλγόριθμους σαν τον CLARA, η συνολική πολυπλοκότητα χρόνου είναι ακόμα τετραγωνική, το οποίο κάνει τον CLARANS όχι και τόσο αποτελεσματικό σε πολύ μεγάλα σύνολα δεδομένων. Προσέγγιση με condensation, όπως ο αλγόριθμος BIRCH [21]. Η βασική ιδέα του αλγορίθμου BIRCH είναι η κατασκευή ενός δέντρου που διατηρεί όλη την απαραίτητη πληροφορία για τον υπολογισμό των αποστάσεων. Ένα κύριο χαρακτηριστικό του αλγορίθμου BIRCH είναι η χρήση του χαρακτηριστικού συσταδοποίησης (clustering feature), μιας τριάδας που περιέχει πληροφορία για τη συστάδα. Το χαρακτηριστικό συσταδοποίσης μιας συστάδας αποτελεί μια περίληψη της πληροφορίας της συστάδας. Από τον ορισμό αυτό είναι ξεκάθαρο πως ο αλγόριθμος BIRCH εφαρμόζεται μόνο σε αριθμητικά δεδομένα. Οι αλγόριθμοι BUBBLE και BUBBLE-FM[73] αποτελούν γενικεύσεις του BIRCH Προσέγγιση που βασίζεται στην πυκνότητα, όπως ο DBSCAN(Density based spatial clustering of applications with noise)[68] και ο DENCLUE(density based clustering)[62]. Ο αλγόριθμος DBSCAN απαιτεί η πυκνότητα σε μια γειτονιά για ένα αντικείμενο να είναι αρκετά μεγάλη αν αυτό ανήκει σε μια συστάδα. Ο DBSCAN δημιουργεί μια νέα συστάδα από ένα πρότυπο απορροφόντας όλα τα πρότυπα στη γειτονιά του. Η γειτονιά πρέπει να ικανοποιεί τιμή κατωφλίου για την πυκνότητα που την ορίζει ο χρήστης. Ο DBSCAN χρησιμοποιεί μια δομή R*-δέντρου για πιο αποδοτικά ερωτήματα. Ο αλγόριθμος DENCLUE αναζητά συστάδες με τοπικά μέγιστα της συνολικής συνάρτησης πυκνότητας, που αντικατοπτρίζει την εκτενή επιρροή των προτύπων στη γειτονιά τους. Προσέγγιση που βασίζεται σε Grid, όπως ο WaveCluster[65] και συσταδοποίηση fractal[72]. Ο αλγόριθμος WaveCluster τοποθετεί τα πρότυπα σε ένα σύνολο μονάδων που έχουν χωριστεί στον αρχικό χώρο των χαρακτηριστικών, και εφαρμόζει μετασχηματισμούς κυματιδίων σε αυτές τις μονάδες, για να απεικονίσει τα πρότυπα στο πεδίο συχνοτήτων. Η βασική ιδέα είναι ότι οι συστάδες μπορούν εύκολα να διαχωριστούν στο μετασχηματισμένο χώρο. Η fractal συσταδοποίηση συνδυάζει τις έννοιες της αυξητικής συσταδοποίησης και της διάστασης φρασταλ. Τα πρότυπα προστίθενται στις συστάδες αυξητικά, που ορίζονταιμέσω μιας αρχικής διαδικασίας, και αναπαρίστανται ως κελιά σε ένα πλέγμα, με την προϋπόθεση ότι η διάσταση φρασταλ της συστάδας πρέπει να μείνει σχετικά σταθερή.

Οι προηγούμενοι αλγόριθμοι δεν έχουν τη δυνατότητα να χειρίζονται δεδομένα μεγάλων δια-

στάσεων. Η απόδοση χειροτερεύει με την αύξηση της διάστασης. Εκτός από τις προηγούμενες προσεγγίσεις, διάφορες άλλες τεχνικές παίζουν επίσης σημαντικό ρόλο στη συσταδοποίηση μεγάλων βάσεων δεδομένων. Παράλληλοι αλγόριθμοι μπορούν πιο αποτελεσματικά να χρησιμοποιήσουν τους υπολογιστικούς πόρους, και έτσι βελτιώνουν τη συνολική απόδοση τόσο σε χρόνο αλλά και στο χώρο.[69][66][67]. Οι αυξητικές τεχνικές συσταδοποίησης δεν απαιτούν όλο εξαρχής το σύνολο των προτύπων αλλά χειρίζονται τα πρότυπα ένα-ένα τη φορά. Αν τα πρότυπα φαίνεται να είναι πολύ κοντα σε μια συστάδα σύμφωνα με κάποια κριτήρια, τότε αυτό μπαίνει στη συστάδα. Διαφορετικά, δημιουργείται μια νέα συστάδα για να αναπαραστίσει το πρότυπο. Ένα τυπικό παράδειγμα είναι η οικογένεια των αλγορίθμων ART[70][71].

5 Ο ΑΛΓΟΡΙΘΜΟΣ ΣΥΣΤΑΔΟΠΟΙΗΣΗΣ X-MEANS

5.1 Εισαγωγή

Ο αλγόριθμος K-means έχει τρία πολύ βασικά μειονεκτήματα. Είναι αργός και δεν έχει καλή κλιμάκωση σε σχέση με το χρόνο που χρειάζεται για να ολοκληρωθεί. Το δεύτερό, είναι ότι ο χρήστης πρέπει κάθε να δίνει την τιμή του K, δηλαδή τον αριθμό των συστάδων. Τέλος η εμπειρία έχει δείξει ότι όταν εκτελείται με σταθερό K, βρίσκει το χειρότερο τοπικό βέλτιστο, από το να αλλάζει δυναμικά την τιμή του K [9].

Η ταχύτητα βελτιώνεται σημαντικά ενσωματώνοντας το dataset σε ένα multiresolution kd-tree και αποθηκεύοντας κάποια στατιστικά δεδομένα στους κόμβους του. Μια προσεκτική ανάλυση της τοποθεσίας των centroid επιτρέπει γεωμετρικές 'αποδείξεις ' για τα σύνορα Voronoi, όπου δεν υπάρχει πουθενα προσέγγιση στους υπολογισμούς. Ένας επιπρόσθετος γεωμετρικός υπολογισμός, blacklisting, διατηρεί μια λίστα με τα centroid που πρέπει να εξεταστούν σε μια γειτονιά [14]. Η μέθοδος blacklisting δεν είναι μόνο γρήγορη, αλλά έχει και καλή κλιμάκωση στον αριθμό των centroid. Αυτός ο γρήγορος αλγόριθμος αποτελεί τη βάση για τον X-means, ένα νέο αλγόριθμο που εκτιμά γρήγορα την τιμή του K, και δρα μετά από κάθε τρέξιμο του K-means, όπου αποφασίζει τοπικά πιο υποσύνολο των centroid θα διαιρεθεί έτσι ώστε να ταιριάξουν καλύτερα στα δεδομένα. Η απόφαση διαίρεσης πραγματοποιείται με τον υπολογισμό του κριτηρίου BIC(Bayesian Information Criterion).

5.2 Ο κλασικός k-means

Θυμίζουμε τη λειτουργία του K-means[10].

Πριν την πρώτη επανάληψη τα κέντρα-centroids αρχικοποιούνται σε τυχαίες τιμές. Ο αλγόριθμος τερματίζει όταν οι θέσεις των centroid παραμένουν σταθερές κατά την διάρκεια μιας επανάληψης. Σε κάθε επανάληψη γίνονται οι παρακάτω δύο ενέργειες

1.Για κάθε σημείο x, βρίσκουμε το κέντρο που είναι πιο κοντά στο x.

2. Επαναπροσδιορίζεται η θέση του κέντρου, όπου για κάθε κέντρο, υπολογίζεται το κέντρο μάζας των σημείων που ανήκουν σε αυτό.

Θεωρούμε ότι[14]:

m_j οι συντεταγμένες του j κέντρου(i)είναι ο δείκτης του κέντρου που είναι πιο κοντά στο i σημείο. Για παράδειγμα, $m_{(i)}$ είναι το κέντρο που συνδέεται με το i-οστό σημείο σε μια επανάληψη. D είναι το σύνολο σημείων που δίνεται ως είσοδος, και $D_i \in D$ είναι το σύνολο το σημείων που έχουν το m_i ως το πιο κοντινό τους κέντρο.
$R = |D|$ και $R_i = |D_i|$.
Ο αριθμός των διαστάσεων είναι M, και ο πίνακας της Γκαουσιανής διακύμανσης είναι

$S = dig(s^2)$
Ο χρήστης αντί να δίνει ακριβώς την τιμή του K, δίνει ένα εύρος τιμών μέσα στο οποίο πιστεύει ότι λογικά βρίσκεται, και το αποτέλεσμα της συσταδοποίησης δεν είναι μόνο οι συστάδες, αλλά επίσης και ο αριθμός K, για τον οποίο έχουμε καλύτερα αποτελέσματα στο κριτήριο BIC.
Κατά βάση ο αλγόριθμος ξεκινά με K ίσο με το κάτω φράγμα του εύρους και συνεχιζει να προσθέτει κέντρα, όπου αυτά χρειάζονται μέχρι να επιτευχθεί το άνω φράγμα του εύρους. Κατά τη διάρκεια της διαδικασίας, το σύνολο των κέντρων που πετυχαίνει την καλύτερη βαθμολογία καταγράφεται, και αυτό είναι το τελικό σύνολο των συστάδων.
Ο αλγόριθμος αποτελείται από δύο λειτουργίες που επαναλαμβάνονται μέχρι την ολοκληρωσή του.
X-means:
Βήμα 1: Βελτίωση Παραμέτρων
Βήμα 2: Βελτίωση Δομής
Βήμα 3: Αν $K > K_{max}$ σταμάτα και ανέφερε το μοντέλο με την καλύτερη βαθμολογία που βρέθηκε
Διαφορετικά πήγαινε στο Βήμα 1.

Το πρώτο βήμα είναι απλό. Αποτελείται από την εκτέλεση-σύγκλιση του κοινού K-means Στο δεύτερο βήμα εντοπίζεται αν και που πρέπει να εμφανιστούν νέα κέντρα. Αυτό επιτυγχάνεται επιτρέποντας σε κάποια κέντρα να διαιρεθούν σε δύο. Στο σχήμα 5 απεικονίζεται μία ευσταθής λύση με 3 κέντρα. Επίσης απεικονίζονται και τα σύνορα των περιοχών που ανήκουν σε κάθε κέντρο. Το δεύτερο βήμα ξεκινά διαιρώντας κάθε κέντρο σε δύο κέντρα απογόνους (σχήμα 6). Αυτοί μετακινούνται σε μια απόσταση ανάλογη στο μέγεθος της περιοχής σε αντίθετες κατευθύνσεις κατά μήκος ενός τυχαίου διανύσματος. Στη συνέχεια σε κάθε γονική περιοχή εκτελείται τοπικά ο K-means(K=2) για κάθε ζευγάρι απογόνων κέντρων. Τα κέντρα απόγονοι

ανταγωνίζονται μεταξύ τους και μόνο στη γονική περιοχή. Το σχήμα 7 δείχνει το πρώτο βήμα όλων των τοπικών 2-means εκτελέσεων. Το σχήμα 8 δείχνει που καταλήγουν τελικά όλα τα παιδιά μετά των τερματισμό όλων των τοπικών 2-means. Σε αυτό το σημείο πραγματοποιείται ένας έλεγχος επιλογής μοντέλου σε όλα τα ζευγάρια. Σε κάθε περίπτωση ο έλεγχος ρωτά, αν υπάρχουν στοιχεία ότι δύο κέντρα απεικονίζουν την πραγματική δομή, ή το γονικό κέντρο απεικονίζει την κατανομή εξίσου καλά·Στην συνέχεια θα παρουσιαστούν οι λεπτομέρειες αυτού του ελέγχου. Σύμφωνα με το αποτέλεσμα του ελέγχου, είτε τα γονικά κέντρα είτε τα κέντρα απόγονοι διαγράφονται. Η προσδοκία είναι, τα κέντρα που έχουν ήδη ένα σύνολο σημείων που σχηματίζουν μία συστάδα στην πραγματική υποκείμενη κατανομή, δεν θα τροποποιηθούν από αυτή τη διαδικασία,θα επιζήσουν δηλαδή των απογόνων τους. Από την άλλη, περιοχές του χώρου που δεν αναπαρίστανται καλά από τα κέντρα, θα δέχονται περισσότερη προσοχή αυξάνοντας των αριθμό των κέντρων σε αυτές. Το σχήμα 9 δείχνει τι συμβαίνει όταν αυτός ο έλεγχος εφαρμοστεί και στα 3 ζευγάρια απογόνων του σχήματος 8. Επομένως ο χώρος αναζήτησης καλύπτει όλους τους πιθανούς 2^K σχηματισμούς διαίρεσης και προσδιορίζει ποιον να διερευνσήσει βελτιώνοντας το BIC τοπικά σε κάθε πριοχή. Στη συνέχεια ο αλγόριθμος ταλαντεύεται μεταξύ του πρώτου και του δεύτερου βήματος μέχρι το K να πάρει τη μέγιστη τιμή του.

5.3 Βαθμολογία BIC

Δίνονται τα δεδομένα D και μια οικογένεια εναλλακτικών μοντέλων M_j, όπου στην περίπτωση μας διαφορετικά μοντέλα αντιστοιχούν σε διαφορετικές τιμές του K. Για την επιλογή του καλύτερου μοντέλου χρησιμοποιείται η εκ των υστέρων πιθανότητες $Pr[M_j|D]$ για τη βαθμολόγηση των μοντέλων. Στην περίπτωση μας τα μοντέλα είναι όλα του που θεωρούνται από τον K-means (δηλαδή σφαιρικές Γκαουσιανές). Για να προσεγγιστούν οι εκ των υστέρων, μέχρι την κανονικοποίηση, χρησιμοποιείται ο ακόλουθος τύπος από τους Kass και Wasserman(1995). $BIC(M_j) = \hat{l}_j(D) - p_j/2 \cdot logR$,

όπου $\hat{l}_j(D)$ είναι η log-πιθανοφάνεια των δεδομένων σύμφωνα με το ϑ-οστό μοντέλο στο σημείο της μέγιστης πιθανοφάνειας, p_j είναι ο αριθμός των παραμέτρων στο M_j. Αυτό είναι γνωστό και κριτήριο Schwartz. Η εκτίμηση μέγιστης πιθανοφάνειας(MLE) για τη διασπορά, υπό την υπόθεση της σφαιρικής Γκαουσιανής, είναι

$\hat{s} = 1/R - K \sum_i (x_i - _{(i)})^2$

Οι σημειακές πιθανότητες είναι:

$\hat{P}(x_i) = R_{(i)}/R \cdot 1\sqrt{2\pi}\hat{s}^M exp(-1/2\hat{s}^2 \left\|x_i - m_{(i)}\right\|^2)$

Η log-πιθανοφάνεια των δεδομένων είναι:

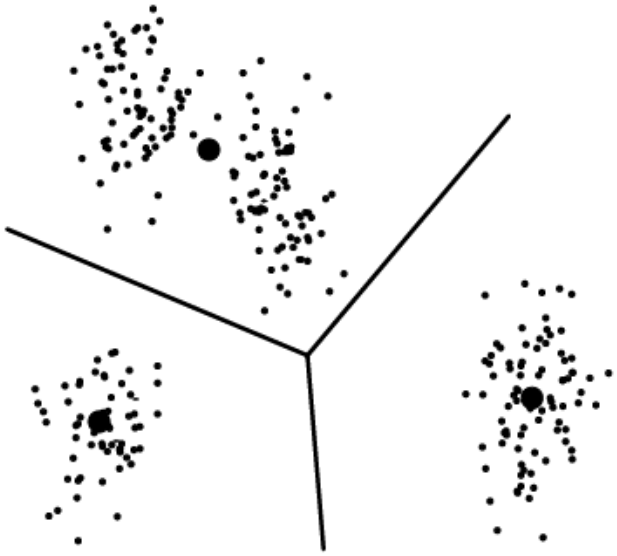
$l(D) = logP_iP(x_i) = \sum_i (\log 1/\sqrt{2\pi}s^M - 1/2s^2 \left\|x_i - m_{(i)}\right\|^2) + logR_{(i)}/R$

Με την τιμή του ν, $1 \leq n \leq K$ σταθερή εστιάζουμε μόνο στο σύνολο D_n των σημείων που ανήκουν στο κέντρο ν και χρησιμοποιώντας την εκτίμηση μέγιστης πιθανοφάνειας έχουμε:

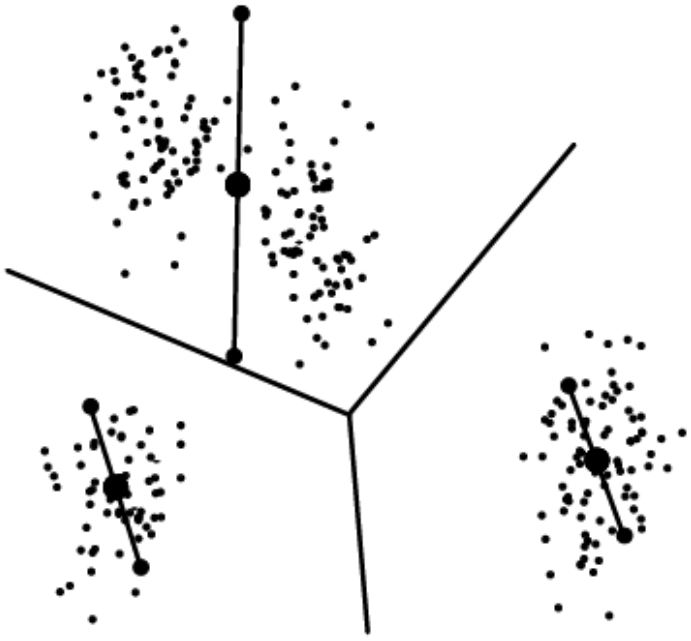
$\hat{l}(D_n) = -\frac{R_n}{2} \log(2\pi) - \frac{R_n \cdot M}{2} \log(\hat{s}^2) - \frac{R_n - K}{2} + R_n \log R_n - R_n \log R$

Ο αριθμός των τριών παραμέτρων P_j είναι απλά το άθροισμα των K-1 class probabilities, $M \cdot K$ οι συντεταγμένες των κέντρων, και μία εκτίμηση της διασποράς. για να επεκτείνουμε αυτό τον τύπο για όλα τα κέντρα εκτός από ένα, χρησιμοποιούμε το γεγονός ότι η log-πιθανοφάνεια των

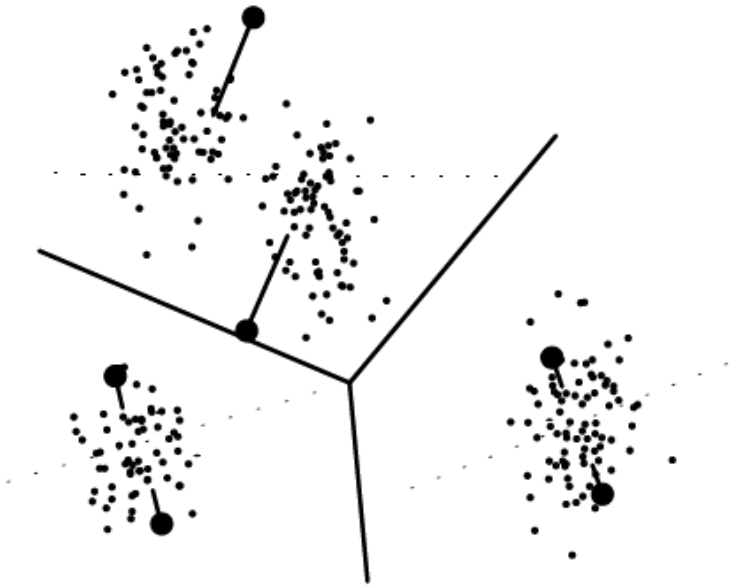
σημείων που ανήκουν σε όλα τα κέντρα υπό ερώτηση είναι το άθροισμα των log-πιθανοφανειών των ατομικών κέντρων, και αντικαθιστούμε το R με το συνολικό αριθμό σημείων που ανήκουν στα κέντρα υπό θεώρηση. Χρησιμοποιούμε τον τύπο BIC όπου ο K-means τελικά επιλέγει το καλύτερο μοντέλο και επίσης τοπικά σε όλους τους ελέγχους διαίρεσης των κέντρων.
Ο αλγόριθμος που περιγράφηκε ως τώρα μπορεί να υλοποιηθεί ως έχει για μικρά σύνολα δεδομένων. Δημιουργήθηκε όμως με τον περιορισμό ότι θα μπορεί να χρησιμοποιεί στατιστικά, έτσι ώστε να μπορεί να κλιμακώνεται σε σύνολα δεδομένων με πολλές εγγραφές. Η επιτάχυνση του αλγορίθμου επιτυγχάνεται αρχικά σε κάθε επανάληψη του K-means. Η διαδικασία είναι να προσδιορίσουμε, για κάθε σημείο πρότυπο, σε πιο κέντρο ανήκει. Τότε μπορούμε να υπολογίσουμε το κέντρο μάζας όλων των σημείων που ανήκουν σε ένα συγκεκριμένο κέντρο, και αυτό δημιουργεί τη νέα τοποθεσία για αυτό το κέντρο. Αμέσως παρατηρούμε ότι ένα υποσύνολο των σημείων, όλα ανήκουν σε ένα κέντρο, όπου είναι τόσο πληροφοριακό όσο το να το κάνουμε αυτό για ένα σημείο, δοθέντος ότι έχουμε επαρκή στατιστικά για το σύνολο(στην περίπτωσή μας τα επαρκή στατιστικά είναι ο αριθμός των σημείων και το διανυσματικό τους άθροισμα). Επειδή ένα kd-δεντρο επιβάλλει μια ιεραρχική δομή στο σύνολο των προτύπων και μπορούμε εύκολα να υπολογίσουμε αρκετά στατιστικά για τους κόμβους του και την κατασκευή, δημιουργεί μια φυσική επιλογή για τη διαμέριση των σημείων. Κάθε kd-κόμβος αναπαριστά ένα υποσύνολο του συνόλου των προτύπων. Έχει επίσης ένα bounding box που είναι ένας ελάχιστος υπερκύβος που περιλαμβάνει όλα τα σημεία στο υποσύνολο. Ακόμα περιέχει δείκτες σε δύο κόμβους απογόνους, που αναπαριστά μια διχοτόμηση των σημείων που περιέχει ο κόμβος γονέας.



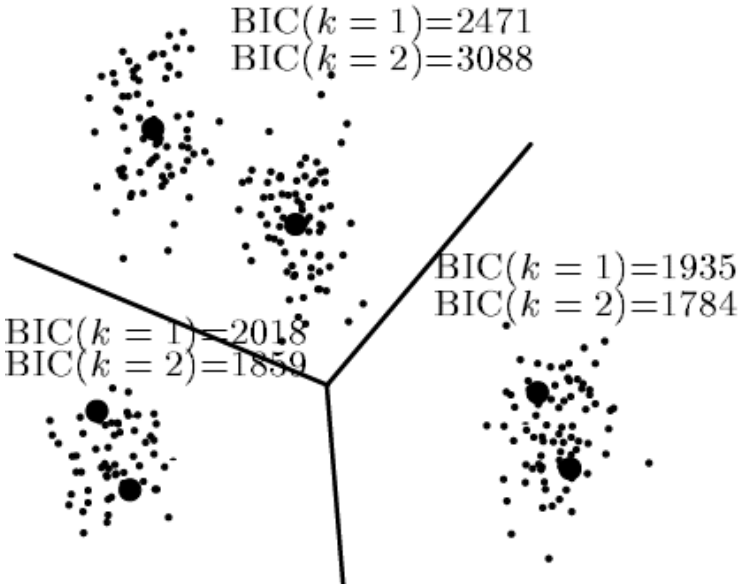
Σχήμα 5: Το αποτέλεσμα του K-means με 3 centroids



Σχήμα 6: Κάθε αρχικό centroid χωρίζεται σε δύο παιδιά.



Σχήμα 7: Το πρώτο βήμα του παράλληλου τοπικού 2-means. Η γραμμή σε κάθε κέντρο δείχνει προς τα που κινείται



Σχήμα 8: Το αποτέλεσμα μετά την ολοκλήρωση όλων των παράλληλων 2-means

Σχήμα 9: Τα κέντρα που απομένουν μετά από όλους τους τοπικούς ελέγχους

6 WEKA

Το WEKA(Waikato Environment for Knowledge Analysis)[84] είναι ένα πακέτο λογισμικού Μηχανικής Μάθησης, και αναπτύχθηκε από ερευνητές του πανεπιστήμιου του Waikato στη Νέα Ζηλανδία. Το WEKA είναι λογισμικό ανοιχτού κώδικα, και είναι υλοποιημένο σε γλώσσα προγραμματισμού Java. Το WEKA περιλαμβάνει μια συλογή από εργαλεία οπτικοποίησης και αλγόριθμους για ανάλυση δεδομένων και προγνωστική μοντελοποίηση, καθώς επίσης και γραφική διεπαφή χρήστη έτσι ώστε να είναι πιο εύκολη η πρόσβαση στις λειτουργίες του. Η πιο πρόσφατη έκδοση που βασίζεται πλήρως σε Java(WEKA 3), η οποία άρχισε να αναπτύσεται το 1997, χρησιμοποιείται σε πολλές διαφορετικές περιοχές εφαρμογών, ιδιαίτερα για εκπαιδευτικούς σκοπούς και έρευνα.

Τα ιδιαίτερα χαρακτηριστικά του WEKA είναι :

1. Είναι ελεύθερα διαθέσιμο υπό την GNU General Public Licence
2. Λειτουργεί σε σχεδόν όλες τις σύγχρονες υπολογιστικές πλατφόρμες επειδή είναι πλήρως υλοποιημένη στη γλώσσα προγραμματισμού Java.
3. Περιλαμβάνει μια κατανοητή συλλογή από τεχνικές προεπεξεργασίας δεδομένων και μοντελοποίησης.
4. Είναι εύκολο στη χρήση ακόμα και από αρχάριους λόγω του της γραφικής διεπαφής χρήστης που περιέχει.

Το WEKA υποστηρίζει αρκετές τυπικές εργασίες Data Mining, και ειδικότερα, προεπεξεργασία δεδομένων, συσταδοποίηση, ταξινόμηση, παλινδρόμηση, οπτικοποίηση, και επιλογή χαρακτηριστικών. Όλες οι τεχνικές του WEKA προϋποθέτουν και βασίζονται στην υπόθεση ότι τα δεδομένα είναι διαθέσιμα σε ένα αρχείο, όπου κάθε σημείο-πρότυπο περιγράφεται από ένα σταθερό αριθμό γνωρισμάτων(αριθμητικά ή κατηγορικά γνωρίσματα). Το WEKA παρέχει επίσης πρόσβαση σε SQL βάσεις δεδομένων και μπορεί να επεξεργαστεί αποτελέσματα που προκύπτουν από ένα ερώτημα μιας βάσης δεδομένων. Η βασική διεπαφή του WEKA είναι ο Explorer, αλλά στις ίδιες λειτουργίες μπορεί να έχει πρόσβαση κάποιος από το Knowledge-Flow αλλά και από το command-line. Υπάρχει επίσης και ο Experimenter, που επιτρέπει συστηματική σύγκριση της προγνωστικής απόδοσης των αλγορίθμων μηχανικής μάθησης του WEKA σε μια συλλογή από σύνολα δεδομένων. Η διεπαφή του Explorer έχει διάφορες καρτέλες, που δίνουν πρόσβαση στις βασικές συνιστώσες. Η καρτέλα Preprocess(προεπεξεργασία) έχει διευκολυνσεις για την εισαγωγή δεδομένων από μια βάση δεδομένων, ένα αρχεία CSV, κ.α., και για την προεπεξεργα-

σία αυτών των δεδομένων χρησιμοποιείται ο λεγόμενος αλγόριθμος φιλτραρίσματος. Αυτά τα φίλτρα μπορούν να χρησιμοποιηθούν για το μετασχηματισμό των δεδομένων (π.χ. η μετατροπη αριθμητικών γνωρισμάτων σε διακριτά), και δίνεται η δυνατότητα να διαγράφονται εγγραφές και γνωρίσματα σύμφωνα με κάποια κριτήρια. Η καρτέλα Classify (Κατηγοριοποίηση) δίνει τη δυνατότητα στο χρήστη να εφαρμόσει αλγόριθμους κατηγοριοποίησης και παλινδρόμησης, να εκτιμήσει την ακρίβεια του μοντέλου πρόβλεψης, να απεικονίσει καμπύλες ROC, κ.α. ή ακόμα και το ίδιο το μοντέλο (π.χ. ένα δέντρο απόφασης). Η καρτέλα Associate (συσχέτιση) παρέχει πρόσβαση σε μάθηση κανόνων συσχετίσεων, όπου προσπαθεί να προσδιορίσει όλες τις σημαντικές εσωτερικές σχέσεις μεταξύ των γνωρισμάτων στα δεδομένα. Η καρτέλα Cluster (συσταδοποίηση) δίνει πρόσβαση σε τεχνικές συσταδοποίησης στο WEKA, π.χ. στον αλγόριθμο K-means. Υπάρχει επίσης και υλοποίηση του αλγορίθμου Expectation maximization για την εκπαίδευση μήξης κανονικών κατανομών. Η επόμενη καρτέλα, Select attributes (επιλογή γνωρισμάτων), παρέχει αλγόριθμους για τον προσδιορισμό των πιο προγνωστικών γνωρισμάτων στο σύνολο των δεδομένων. Η τελευταία καρτέλα, Visualize (οπτικοποίηση), απεικονίζει έναν πίνακα με διαγράμματα διασποράς, όπου ατομικά διαγράμματα διασποράς μπορούν να επιλεγούν και να μεγενθυθούν, και να αναλυθούν περαιτέρω χρησιμοποιώντας διάφορους τελεστές επιλογής. Το RapidMiner, άλλοτε γνωστό και ως YALE (Yet Another Learning Enviroment), είναι ένα περιβάλλον για μηχανική μάθηση και Data Mining. Επιτρέπει τη δημιουργία πειραμάτων με χρήση ενός μεγάλου αριθμού τυχαία εμφωλεύσιμων τελεστών, που περιγράφονται σε αρχεία XML, και μπορούν να δημιουργηθούν με τη γραφική διεπαφή του RapidMiner. Εφαρμογές του RapidMiner καλύπτουν ερευνητικές εργασίες αλλά και εργασίες του πραγματικού κόσμου. Η αρχική έκδοση αναπτύχθηκε από το τμήμα τεχνητής Νοημοσύνης του πανεπιστημίου του Dortmund το 2001. Διανέμεται σύμφωνα με την GNU άδεια, από το SourceForge από το 2004. Το RapidMiner παρέχει πάνω από 500 τελεστές για όλες τις βασικές διαδικασίες μηχανικής μάθησης, συμπεριλαμβάνοντας input και output, και προεπεξεργασια δεδομένων, καθώς και οπτικοποίηση. Έχει γραφτεί στη γλώσσα προγραμματισμού Java και έτσι μπορεί να λειτουργήσει σε όλα τα γνωστά λειτουργικά συστήματα. Επίσης ενσωματώνει όλες τις τεχνικές του WEKA.

7 ΦΑΣΜΑΤΟΣΚΟΠΙΑ ΜΑΖΑΣ

Η φασματομετρία μάζας (mass spectrometry-MS) είναι μια αναλυτική τεχνική για τον προσδιορισμό της βασικής σύνθεσης ενός δείγματος ή μορίου. Χρησιμοποιείται επίσης για να αποσαφηνιστεί η χημική δομή των μορίων, όπως των πεπτιδίων και άλλων χημικών ενώσεων. Η βασική αρχή της φασματομετρίας είναι ο ιονισμός των χημικών ενώσεων για τη δημιουργία φορτισμένων μορίων ή μοριακών θραυσμάτων και η μέτρηση του λόγου της μάζας προς το φορτίο. Σε μια τυπική διαδικασία φασματομετρίας μάζας [74]:

1. Ένα δείγμα εισάγεται στη συσκευή φασματομέτρησης, και

2. Τα συστατικά του δείγματος ιονίζονται με μια πληθώρα μεθόδων, με αποτέλεσμα τη δημιουργία φορτισμένων σωματιδίων(ιόντα)

3. Τα ιόντα κατευθύνονται σε ένα ηλεκτρικό ή/και μαγνητικό πεδίο

4. Υπολογίζεται ο λόγος της μάζας προς το φορτίο των σωματιδίων βασίζεται σε λεπτομέρειες της κίνησης των ιόντων κατά την διεύλευση από το ηλεκτρομαγητικό πεδίο, και

5. Ανίχνευση των ιόντων, τα οποία κατά το προηγούμενο βήμα ταξιμονήθηκαν με βάση την τιμή m/z.

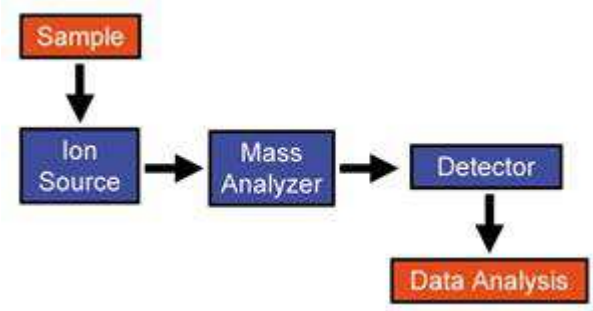
Οι συσκευές φασματομέτρησης μάζας αποτελούνται από τρία τμήματα:

1. από μια πηγή ιόντων, που μπορεί να μετατρέψει δείγμα αέριων μορίων σε ιόντα,

2. από έναν αναλυτή μάζας, που ταξινομεί τα ιόντα με βάση τη μάζας τους εφαρμόζοντας ηλεκτρομαγνητικά πεδία και

3. έναν ανιχνευτή που μετρά την τιμή ενός δείκτη ποσότητας και έτσι παρέχει δεδομένα για τον υπολογισμό της αφθονίας κάθε ιόντος που παρουσιάζεται.

Η τεχνική αυτή έχει ποιοτικές και ποσοτικές χρήσεις. Αυτές περιλαμβάνουν τον προσδιορισμό αγνώστων ενώσεων, τον προσδιορισμό της ισοτοπικής σύνθεσης των στοιχείων σε ένα μόριο, και τον προσδιορισμό της δομής μιας ένωσης παρατηρώντας τη θραυσματοποίηση. Άλλες χρήσεις περιλαμβάνουν τη μέτρηση της ποσότητας μιας ένωσης σε ένα δείγμα ή τη μελέτη των θεμελίων στη χημεία ιόντων αέριας φάσης (η χημεία των ιόντων και των ουδέτερων στο κενό). Η φασματομετρία χρησιμοποιείται ευρέως σε αναλυτικά εργαστήρια που μελετούν φυσικές, χημικές, ή βιολογικές ιδιότητες μια μεγάλης ποικιλίας ενώσεων.



Σχήμα 10: Ιονισμός -Διαχωρισμός Μαζών- Ανίχνευση Ιόντων

7.1 Μέθοδοι Ιονισμού

Η πηγή ιονισμού είναι το τμήμα του φασματόμετρου που ιονίζει το υπό εξέταση υλικό. Τα ιόντα μεταφέρονται μέσω ενός μαγνητικού ή ηλεκτρικού πεδίου στο διαχωριστή μαζών. Οι τεχνικές ιονισμού προσδιορίζουν και τον τύπο των δειγμάτων που μπορούν να αναλυθούν από τη φασματομετρία μάζας. Οι τεχνικές ηλεκτρονιακού και χημικού ιονισμού χρησιμοποιούνται για αέρια και υδρατμούς.Στο χημικό ιονισμό χρησιμοποιούνται αντιδράσεις μεταξύ ιόντων και μορίων για την παραγωγή ιόντων, χρησιμοποιώντας ένα αέριο όπως το μεθάνιο. Δύο τεχνικές που χρησιμοποιούνται συχνά με υγρά και στερεά βιολογικά δείγματα είναι αυτή του Ιονισμού με ηλεκτροψεκασμό (ESI-Electrospray Ionization) και MALDI (matrix-assisted laser desorption/ionization) που αναπτύχθηκε χωριστά από τους K. Tanaka και M. Karas και F. Hillenkamp [74]. Άλλες τεχνικές ιονισμού είναι οι: glow discharge, field desortion (FD),fast atom bombardment (FAB), thermospray, desorption/ionization on silicon(DIOS), Direct Analysis in Real Time (DART), atmospheric pressure chemical ionization (APCI), secondary ion mass spectrometry (SIMS), spark ionization και thermal ionization (TIMS) .

7.2 Μέθοδοι διαχωρισμού μαζών

[75] Ο διαχωρισμός μαζών είναι το κρισιμότερο στάδιο στην πορεία μιας ανάλυσης και έχουν αναπτυχθεί μέθοδοι διαχωρισμού που διαφέρουν ως προς τα χαρακτηριστικά τους, τις δυνατότητες και τους περιορισμούς τους. Γενικά, δεν φαίνεται να υπάρχει μία μέθοδος διαχωρισμού μαζών που να είναι εξ' ίσου αποτελεσματική σε όλες τις εφαρμογές της ανάλυσης με φασματογραφία μάζας. Οι συνηθέστερες μέθοδοι διαχωρισμού μαζών (φίλτρα μαζών) είναι α) μαγνητικού πεδίου β) τετραπόλου και γ) χρόνου πτήσης. Ολα τα φίλτρα μαζών βασίζονται στην εφαρμογή ηλεκτρικών και μαγνητικών πεδίων που επιδρούν στην κίνηση φορτισμένων σωματιδίων (ιόντων). Οι

μαθηματικές σχέσεις που διέπουν την αλληλεπίδραση των ηλεκτρικών και μαγνητικών πεδίων με τα ιόντα έχουν διατυπωθεί από τον Νεύτωνα (δεύτερος νόμος) και τον νόμο της δύναμης του Lorentz, αντίστοιχα[75]:

$F = mg(1)$

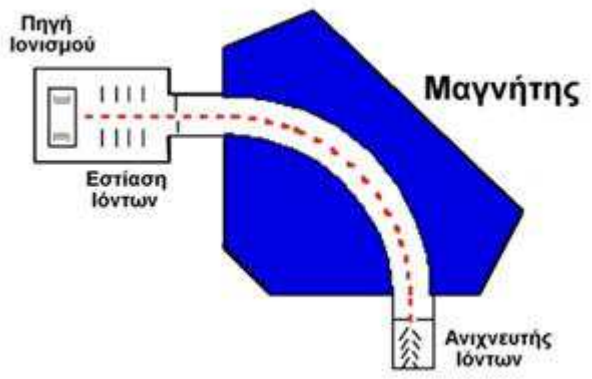
$F = z(E + (vB))$ (2)

όπου F είναι η το διάνυσμα της δύναμης που ασκείται στο ιόν, m η μάζα του, g το διάνυσμα της επιτάχυνσης, z το φορτίο του ιόντος, E το διάνυσμα της έντασης του ηλεκτρικού πεδίου, v το διάνυσμα της ταχύτητας του ιόντος και B το διάνυσμα της έντασης του μαγνητικού πεδίου. Από την σχέση (1) φαίνεται ότι η δύναμη F προκαλεί επιτάχυνση ανάλογη της μάζας του σωματιδίου, ενώ από την σχέση (2) φαίνεται ότι η εφαρμοζόμενη δύναμη F είναι ανάλογη του ιοντικού φορτίου z. Συνεπώς, και οι τρείς προαναφερόμενες μέθοδοι διαχωρίζουν τα ιόντα ανάλογα με τον λόγο m/z, όπου m είναι η μοριακή μάζα του ιόντος και z το φορτίο του. Στην πλειονότητα των περιπτώσεων τα παραγόμενα ιόντα είναι απλά φορτισμένα (z=1) και οι εξαιρέσεις ($z > 1$) πρέπει να αναζητηθούν σε μόρια με πολύ χαμηλά δυναμικά ιονισμού (π.χ. θειούχες ενώσεις ή χημικές ενώσεις βαρέων στοιχείων). Παρ' όλα αυτά, η δημιουργία ιόντων με φορτίο $z > 1$ δεν συμβαίνει σε μεγάλο βαθμό και οι εντάσεις τους σε ένα φάσμα μάζας είναι μικρές.

1. Μαγνητικό πεδίο

Η μέθοδος του μαγνητικού πεδίου (magnetic sector mass filter) είναι η αρχαιότερη και ταυτόχρονα η περισσότερο ακριβής και δαπανηρή, ενώ μία απλή διάταξη του παρουσιάζεται στο επόμενο σχήμα.

Τα ιόντα μετά την δημιουργία τους εστιάζονται κατάλληλα με "φακούς ιόντων" (ανομοιογενή ηλεκτρικά πεδία που συγκλίνουν τις τροχιές των ιόντων) με σκοπό την αύξηση της ευαισθησίας της μεθόδου και οδηγούνται σε ένα χώρο όπου μαγνητικό πεδίο έντασης B εφαρμόζεται κάθετα στην κατεύθυνση κίνησης των ιόντων. Τα ιόντα εισέρχονται στο μαγνητικό πεδίο έχοντας ευθύγραμμη ομαλή πορεία με ταχύτητα v, και αναγκάζονται να ακολουθήσουν κυκλική τροχιά, της οποίας η ακτίνα εξαρτάται από τον λόγο m/z. Η κινητική ενέργεια ενός ιόντος πρίν εισέλθει στο μαγνητικό πεδίο είναι



Σχήμα 11: Μέθοδος Μαγνητικού πεδίου

$T = zV_a = \frac{mv^2}{2}$

ενώ η ταχύτητα v δίνεται από:

$v = \sqrt{\frac{2zV_a}{m}}$

όπου V_a το εφαρμοζόμενο δυναμικό επιτάχυνσης των ιόντων. Από τον νόμο της δύναμης του Lorentz, το μαγνητικό πεδίο εφαρμόζει μία δύναμη zvB η οποία ισούται με την φυγόκεντρο δύναμη mv^2/r, καθώς το ιόν κινείται σε κυκλική τροχιά, με ακτίνα r:

$zvB = \frac{mv^2}{r}$

Αντικατάσταση με την τιμή v της ταχύτητας, προκύπτει τελικά:

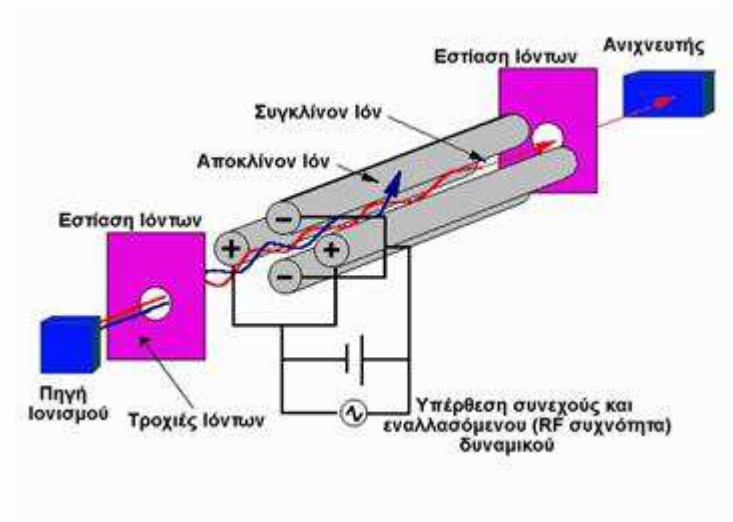
$\frac{m}{z} = \frac{B^2r^2}{2V_a}$

Συνεπώς, αν ο ανιχνευτής βρίσκεται σε συγκεκριμένη θέση που αντιστοιχεί σε μία τροχιά με ακτίνα r, μόνο τα ιόντα με συγκεκριμένο λόγο m/z, ο οποίος ικανοποιεί την ανωτέρω ισότητα φθάνουν στον ανιχνευτή. Η σάρωση σε όλες τις επιθυμητές μάζες (ή ακριβέστερα τον λόγο m/z) γίνεται είτε μεταβάλλοντας την ένταση του μαγνητικού πεδίου B ή μεταβάλλοντας το αρχικό δυναμικόVa επιτάχυνσης των ιόντων. Ο δεύτερος τρόπος έχει το πλεονέκτημα της έλλειψης

υστέρησης της μεταβολής του μαγνητικού πεδίου, όμως η ευαισθησία είναι σχετικά ανάλογη του λόγου m/z. Η αναπόφευκτη θερμική κίνηση των ιόντων προκαλεί μία μείωση της θεωρητικά άπειρης διακριτικής ικανότητας, η οποία μπορεί να βελτιωθεί σημαντικά με την χρήση σχισμών (slits) οι οποίες τοποθετούνται πριν την είσοδο στο μαγνητικό πεδίο και περιορίζουν την είσοδο σε ιόντα που κινούνται σε συγκεκριμένη κατεύθυνση. Η αύξηση της διακριτικής ικανότητας συντελεί στη δυνατότητα διαχωρισμού ιόντων με διαφορές μάζας της τάξεως των 0.01 αμυ. Γιά παράδειγμα, ο διαχωρισμός των CO, N2, $H_2C = CH_2$ και της ελεύθερης ρίζας $H_2C = N$ με μοριακές μάζες 27.9949, 28.0062, 28.0312 και 28.0187, αντίστοιχα, είναι εφικτός. Επιπλέον, η μαγνητική μέθοδος διαχωρισμού μαζών χαρακτηρίζεται από εξαιρετικά καλή ευαισθησία, αναπαραγωγίσιμα φάσματα μάζας, υψηλή δυναμική περιοχή (μεγάλη κλίμακα μαζών) και μπορεί να εφαρμοστεί σε ποσοτικές εφαρμογές. Το σημαντικότερο μειονέκτημα της μεθόδου είναι το υψηλότατο κόστος.

2. Τετραπολικό φίλτρο μαζών (quadrupole mass filter)

Αποτελεί μία από τις νεώτερες μεθόδους διαχωρισμού μαζών, οικονομικότερη από την χρήση του μαγνητικού φίλτρου μαζών με αποτέλεσμα την σημαντικά ευρύτερη χρήση της. Ο διαχωρισμός μαζών επιτυγχάνεται με την υπέρθεση ενός εναλλασομένου ηλεκτρικού πεδίου έχοντας συχνότητα στην περιοχή των ραδιοκυμάτων (RF, radiofrequency) σε ένα συνεχές (DC) ηλεκτρικό πεδίο. Το DC-RF πεδίο εφαρμόζεται σε τέσσερις παράλληλες ράβδους, όπως δείχνεται στο επόμενο σχήμα.



Σχήμα 12: Μέθοδος Τετραπολικού φίλτρου μαζών

Σε κάθε ζεύγος αντιθέτως ευρισκομένων ράβδων το ηλεκτρικό πεδίο περιγράφεται με την σχέση:

$W = 2[U + Vcos(wt)]$

όπου U το δυναμικό του συνεχούς πεδίου, V το μέγιστο του δυναμικού του εναλλασσομένου πεδίου, ω η κυκλική συχνότητα του πεδίου RF και t ο χρόνος. Τα ιόντα που κατευθύνονται στην κατεύθυνση του Z άξονα (ορίζεται ως παράλληλος με τις ράβδους) αναγκάζονται να κινηθούν σε μία τροχιά ταλάντωσης ανάμεσα στους άξονες X και Υ. Η μαθηματική ανάλυση της κίνησης ενός ιόντος στο συνδυασμό των πεδίων είναι αρκετά πολύπλοκη, και έχει επιτευχθεί χρησιμοποιώντας κλασσική μηχανική (εξισώσεις Matthieu). Η περιγραφή της κίνησης του ιόντος μπορεί να γίνει με την εισαγωγή των παραμέτρων:

$a = 8zUmr_0^2w^2$

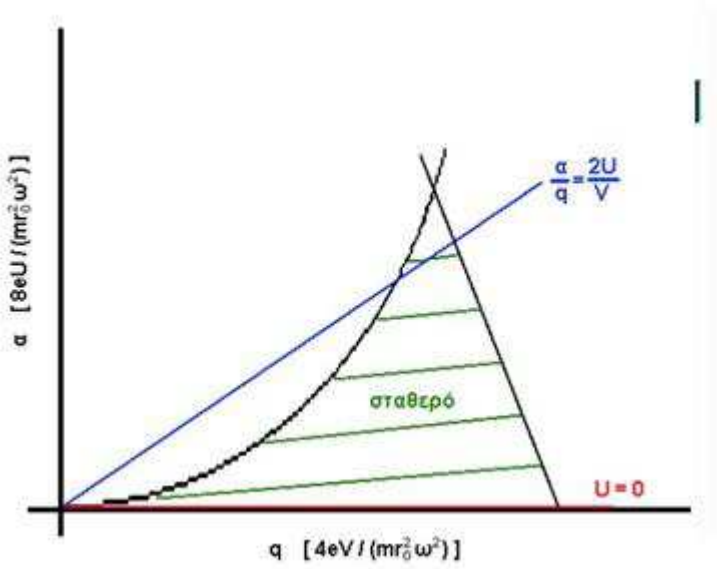
και

$q = \frac{4zV}{mr_0^2w^2}$

όπου z το φορτίο του ιόντος, m η μάζα του και ro η ακτίνα του πεδίου (μέσον της απόστασης ανάμεσα στα κέντρα αντιθέτως ευρισκομένων ράβδων).

Οι υπολογισμοί δείχνουν οτι μόνο οι παράμετροι α και q αρκούν για τον προσδιορισμό της τροχιάς του ιόντος. Όταν το πλάτος της ταλάντωσης του ιόντος παραμένει σταθερό στο χρόνο η τροχιά χαρακτηρίζεται ως σταθερή, ενώ αν το πλάτος της ταλάντωσης αυξάνει με την πάροδο του χρόνου, τότε η τροχιά χαρακτηρίζεται ως ασταθής. Όπως είναι προφανές, μία σταθερή τροχιά έχει σαν αποτέλεσμα την δίοδο ενός ιόντος διαμέσου του χώρου καθ' όλο το μήκος των ράβδων, ενώ αντίθετα, αν η τροχιά είναι ασταθής, το ιόν τελικά προσκρούει πάνω στις ράβδους, εξουδετερώνεται και δεν κατορθώνει να φτάσει στον ανιχνευτή (που βρίσκεται στο άλλο άκρο των ράβδων). Η ανάλυση της τροχιάς ενός ιόντος φαίνεται στο ακόλουθο διάγραμμα σταθερότητας $a = f(q)$.

Μόνο τα ιόντα με τιμές α και q στην (σχεδόν) τριγωνική γραμμοσκιασμένη περιοχή εκτελούν σταθερές ταλαντώσεις. Γιά ένα δεδομένο λόγο $2U/V$, ο λόγος a/q είναι ο ίδιος γιά όλες τις μάζες και τα σημεία λειτουργίας γιά όλες τις μάζες βρίσκονται σε μία ευθεία γραμμή στο διάγραμμα σταθερότητας, η οποία διέρχεται από την αρχή των αξόνων. Το διάγραμμα σταθερότητας και οι εκφράσεις των παραμέτρων α και χ δείχνουν ότι:



Σχήμα 13: Ανάλυση τροχιάς ιόντος

α) η μάζα m είναι ανάλογη του V (με την ω σταθερή).

β) η διακριτική ικανότητα Dm/m είναι σταθερή για σταθερό λόγο U/V και επίσης δεν μεταβάλλεται με την μάζα m.

γ) η διακριτική ικανότητα μπορεί να ρυθμιστεί με την μεταβολή του U.

δ) οι παράμετροι a και q εξαρτώνται από το λόγο m/z.

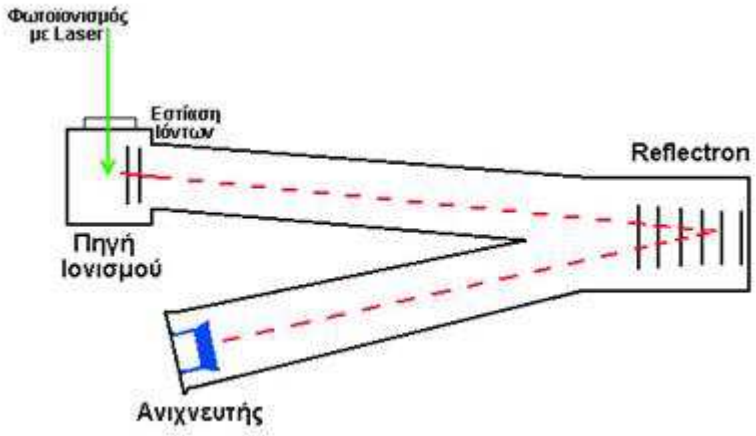
Αξίζει να παρατηρηθεί ότι αν το εναλλασσόμενο πεδίο μηδενιστεί τότε όλα τα σημεία βρίσκονται στον q-άξονα και για ένα πολύ ασθενές εναλλασσόμενο πεδίο όλα τα ιόντα εκτελούν σταθερές ταλαντώσεις. Επιπλέον, λόγω της θερμικής κίνησης των ιόντων, η ανεξαρτησία της διακριτικής ικανότητας από την μάζα ισχύει μόνο για επαρκή χρόνο παραμονής στο τετραπολικό πεδίο και αρκετά στενή κατανομή ταχυτήτων στον άξονα Z. Η διαπερατότητα των ιόντων δια μέσου του τετραπολικού φίλτρου μειώνεται με την αύξηση της διακριτικής ικανότητας (αύξηση του U) καθ' όσον τα σημεία λειτουργίας ανέρχονται στο διάγραμμα σταθερότητας. Η σάρωση της κλίμακας μαζών συνήθως γίνεται είτε με μεταβολή της κυκλικής συχνότητας ω, ή με παράλληλη μεταβολή των εντάσεων των πεδίων U και V, κρατώντας τον λόγο U/V σταθερό.

Τα πλεονεκτήματα του τετραπολικού φίλτρου μαζών είναι το μικρό κόστος και η αναπαραγωγισιμότητα των λαμβανομένων φασμάτων μάζας. Μειονεκτήματα του είναι η χαμηλή διακριτική ικανότητα (1 amu) και η εξάρτηση της διαπερατότητας ενός ιόντος από την μάζα του. Οι τετραπολικοί φασματογράφοι μάζας χρησιμοποιούνται ευρέως στην καθημερινή χημική ανάλυση και συνήθως ακολουθούν τους αέριους χρωματογράφους (Gas Chromatography, GC) οι οποίοι

επιτυγχάνουν τον διαχωρισμό των συστατικών ενός μίγματος για την ευκολότερη ταυτοποίηση κάθε συστατικού.

3. Χρόνος πτήσης (Time-of-flight)

Η τεχνική του χρόνου πτήσης χρησιμοποιεί την εξάρτηση της ταχύτητας από την μάζα ενός ιόντος που κινείται σε χώρο απαλλαγμένο ηλεκτρομαγνητικών πεδίων.



Σχήμα 14: Μέθοδος χρόνου πτήσης

Το σχηματιζόμενο ιόν επιταχύνεται σε ταχύτητα ν κάτω από την επίδραση ηλεκτρικού πεδίου δυναμικού V:
$T = zV = \frac{mV^2}{2}$
ενώ η ταχύτητα του ν δίνεται από την σχέση:
$v = \sqrt{\frac{2zV}{m}}$

Μετά την αρχική επιτάχυνση, το ιόν εισέρχεται σε χώρο απουσία ηλεκτρικού πεδίου στον οποίο κινείται ευθύγραμμα και ομαλά μέχρι τον ανιχνευτή, σε απόσταση L, απαιτώντας χρονικό διάστημα t:

$t = Lv = L\sqrt{\frac{m}{2zV}}$

Συνεπώς, ο απαιτούμενος χρόνος εξαρτάται από την μάζα του ιόντος και ως εκ τούτου επιτυγχάνεται η διάκριση των ιόντων ανάλογα της μάζας τους. Η τεχνική προαπαιτεί τον ακριβή

καθορισμό του χρόνου εκκίνησης του ιόντος από την πηγή ιονισμού. Είναι προφανές ότι η δημιουργία ιόντων και η τελική ανίχνευση τους δεν είναι δυνατόν να πραγματοποιείται με συνεχή τρόπο, αλλά μόνο με την μορφή παλμών. Ως εκ τούτου, η τεχνική συνιστάται στην χημική ανάλυση διαδικασιών που εκτελούνται παλμικά (φωτοδιάσπαση και απόξεση επιφανειών, με παλμικό laser).

7.3 Μέθοδοι ανίχνευσης ιόντων

Οι μέθοδοι ανίχνευσης ιόντων [74] [75] [76]στηρίζονται στο γεγονός της δευτερογενούς εκπομπής ηλεκτρονίων κατά την πρόσπτωση ηλεκτρονίων ή ιόντων σε επιφάνεια, κατάλληλα επιστρωμένη με ειδικά υλικά ή πολωμένη σε πολύ υψηλό δυναμικό (kV). Συνήθως χρησιμοποιείται το 'φαινόμενο χιονοστιβάδας ' (avalanche effect), κατά το οποίο η αρχική εκπομπή ηλεκτρονίων δίνει αφορμή σε επιπλέον εκπομπή από πολλαπλές προσπτώσεις των δευτερογενώς παραγομένων ηλεκτρονίων και την τελική ενίσχυση του ασθενέστατου ρεύματος των αρχικών ιόντων σε μετρήσιμο ρεύμα ηλεκτρονίων. Οι συνηθέστερες τεχνικές ανίχνευσης ιόντων είναι:

1. Κύπελλο (cup) Faraday. Αποτελείται από μία κοίλη επιφάνεια, η οποία συνδέεται με ένα ηλεκτρόμετρο (συσκευή μέτρησης ηλεκτρικού φορτίου). Φορτίζεται αρνητικά με την πρόσπτωση ηλεκτρονίων στην επιφάνεια της με συνέπεια την ανταπόκριση του ηλεκτρομέτρου. Είναι μία καθαρά αναλογική μέθοδος και δεν μπορεί να λειτουργήσει με παλμικές διατάξεις. Η ευαισθησία του Faraday cup είναι σχετικά μικρή, με σημαντικό πλεονέκτημα το χαμηλό κόστος.

2. Ηλεκτρονοπολλαπλασιαστής (Secondary Electron Multiplier, SEM). Αποτελείται από μία διαδοχική σειρά δυνόδων (μεταλλικές επιφάνειες -συνήθως κράμα Cu-Be- οι οποίες εκπέμπουν δευτερογενή ηλεκτρόνια) πολωμένων σε υψηλό δυναμικό. Μπορεί να λειτουργήσει σε αναλογικές ή ψηφιακές διατάξεις, έχει καλή ευαισθησία αλλά σχετικά υψηλό κόστος.

3. Ανιχνευτής Daly. Αποτελείται από μία μεταλλική επιφάνεια με δυνατότητα εκπομπής δευτερογενών ηλεκτρονίων τα οποία επιταχύνονται προς ένα σπινθηριστή (scintillator) και το εκπεμπόμενο φώς ανιχνεύεται από φωτοπολλαπλασιαστή.

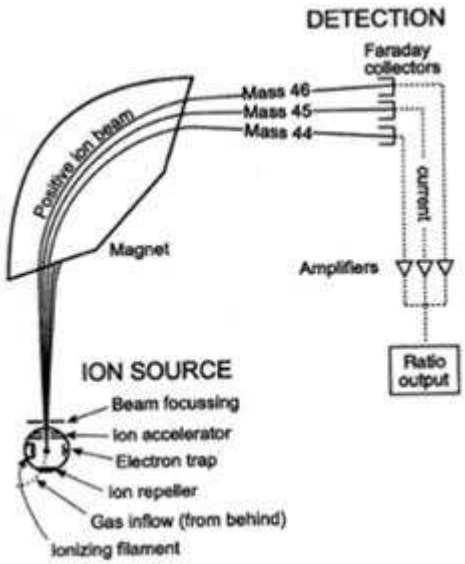
4. Channeltron. Αποτελείται από μία δύνοδο σε σχήμα χωνιού έχοντας ειδική επίστρωση υλικού με δυνατότητα εκπομπής δευτερογενών ηλεκτρονίων των οποίων ο αριθμός αυξάνεται σημαντικά με το ¨φαινόμενο χιονοστοιβάδασ¨, δίνοντας τελικά μία ένα ρεύμα εύκολα μετρήσιμο με ένα ηλεκτρόμετρο.

5. Πλάκα μικροδιόδων (Microchannel plate). Αποτελείται από μία συστοιχία γυάλινων τριχοειδών σωλήνων με εσωτερική διάμετρο 10-25 μμ, τα οποία είναι επιστρωμένα με κατάλληλο υλικό το οποίο έχει δυνατότητα εκπομπής δευτερογενών ηλεκτρονίων. Ταυτόχρονα τα τριχοει-

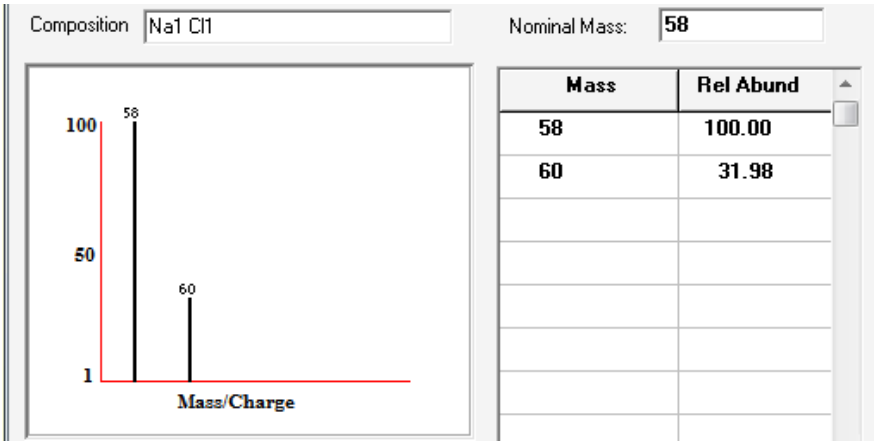
δή βρίσκονται πολωμένα σε υψηλή τάση και με την βοήθεια του ¨φαινομένου χιονοστοιβάδας' παράγεται ρεύμα ενισχυμένο κατά 103 ως 104 σε σχέση με το αρχικό ρεύμα ιόντων. Ο ανιχνευτής έχει μεγάλη ευαισθησία, ενώ η δυνατότητα κατασκευής μεγάλης επιφάνειας από πολλές οπές τριχοειδών επιτρέπει την χρήση του σε συστήματα μέτρησης πολλών μαζών ταυτόχρονα (multichannel detector, focal plane array detectors) σε συνδυασμό με φίλτρο μαζών μαγνητικού πεδίου.

7.4 Απλοποιημένο παράδειγμα

Το ακόλουθο παράδειγμα [74] περιγράφει τη λειτουργία ενός φασματόμετρου μαγνητικού πεδίου. Θεωρούμε ένα δείγμα του χλωριούχου νατρίου (άλας). Στην πηγή ιόντων, το δείγμα ατμοποιείται (μετατρέπεται σε αέριο) και ιονίζεται (μετασχηματισμένο σε ηλεκτρικά φορτισμένα σωματίδα) στα ιόντα νατρίου (Na+) και χλωρίου (Cl^-). Τα άτομα και τα ιόντα νατρίου είναι μονοισοτοπικά, με μια μάζα περίπου 23 amu . Τα άτομα και τα ιόντα χλωρίου έρχονται σε δύο ισότοπα με μάζες σε 35 amu (σε μια φυσική αφθονία περίπου 75 τοις εκατό) και 37 amu περίπου 37 (σε μια φυσική αφθονία περίπου 25 τοις εκατό). Το τμήμα ανάλυσης του φασματομέτρου περιέχει ηλεκτρικά και μαγνητικά πεδία, τα οποία ασκούν δυνάμεις στα ιόντα που ταξιδεύουν μέσω αυτών των πεδίων. Η ταχύτητα ενός φορτισμένου σωματισίου μπορεί να αυξηθεί ή να μειωθεί περνώντας μέσα από τοηλεκτρικού πεδίου, και η κατεύθυνσή του μπορεί να αλλάξει από το μαγνητικό πεδίο. Το μέγεθος της εκτροπής της τροχιάς του κινούμενου ιόντος εξαρτάται από το λόγο μάζα-φορτίου . Από το δεύτερο νόμο κίνησης του Newton , τα ελαφρύτερα ιόντα εκτρέπονται από τη μαγνητική δύναμη περισσότερο από τα βαρύτερα ιόντα. Τα ρεύματα (streams) των ταξινομημένων ιόντων περνούν από τη συσκευή ανάλυσης στον ανιχνευτή, ο οποίος καταγράφει τη σχετική αφθονία κάθε ιόντος. Αυτές οι πληροφορίες χρησιμοποιούνται για να καθορίσουν τη χημική σύνθεση στοιχείων του αρχικού δείγματος (δηλ. ότι και το νάτριο και το χλώριο είναι παρόντα στο δείγμα) και την ισοτοπική σύνθεση των συστατικών της (η αναλογία των 35Cl 37Cl).



Σχήμα 15: Σχέδιο ενός απλού φασματογράφου μάζας μαγνητικού πεδίου



Σχήμα 16: Φάσμα Μάζας του άλατος

7.5 Χρωματογραφικές τεχνικές σε συνδυασμό με φασματομετρία μάζας

Μια σημαντική βελτίωση στη μαζική επίλυση και στις ικανότητες προσδιορισμού μάζας της φασματομετρίας μάζας, είναι η συνδυαστική χρήση της με χρωματογραφικές τεχνικές διαχωρισμού.

7.5.1 Αέρια Χρωματογραφία

Ένας συνήθης συνδυασμός είναι η αέρια χρωματογραφία και η φασματομετρία μάζας (GC-MS). Σε αυτή την τεχνική, ένας αέριος χρωματογράφος χρησιμοποιείται για να διαχωρίσει διαφορετικές ενώσεις. Το σύνολο αυτό των διαχωρισμένων ενώσεων τροφοδοτείται στην περιοχή

ιονισμού, μια μεταλλική ίνα στην οποία εφαρμόζεται τάση. Η ίνα εκπέμπει ηλετρόνια που ιονίζουν τις ενώσεις. Τα ιόντα μπορούν στη συνέχεια να θραυματιστούν, παράγοντας προβλέψιμα πρότυπα. Τα άθικτα ιόντα καθώς και τα θραύσματα περνούν στη συσκευή ανάλυσης του φασματόμετρου μάζας, όπου τελικά ανιχνεύονται.

7.5.2 Υγρή χρωματογραφία

Όπως και η αέρια χρωματογραφία MS (GS-MS), έτσι και η υγρή χρωματογραφία φασματομετρίας μάζας (LC-MS) διαχωρίζει τις ενώσεις χρωματογραφικά πριν εισαχθούν στην περιοχή ιονισμού και τον φασματογράφο μάζας. Διαφέρει από τη GS-MS στο ότι η κινητή φάση είναι υγρή, συχνά ένα μίγμα νερού και διαλύτη, αντί για αέριο. Συχνά στην LC-MS χρησιμοποιείται ιονισμός με ηλεκτροψεκασμό. Υπάρχουν επίσης και νέες τεχνικές ιονισμού με χρήση laser.

7.6 Δεδομένα και Ανάλυση Φασμάτων Μάζας

Η φασματομετρία μάζας παράγει διάφορους τύπους δεδομένων. Η πιο συνηθισμένη αναπαράσταση είναι το φάσμα μάζας (mass spectrum) . Κάποιοι τύποι δεδομένων φασματομετρίας μάζας αναπαρίστανται καλύτερα ως χρωματογράμματα μάζας. Οι τύποι των χρωματογραφημάτων περιλαμβάνουν , selected ion monitoring (SIM), total ion current (TIC), και selected reaction monitoring (SRM). Άλλοι τύποι δεδομένων φασματομετρίας μάζας αναπαρίστανται καλά ως τρισδιάστατοι χάρτες. Σε αυτή τη μορφή, στο άξονα x βρίσκονται οι τιμές m/z , η intensity βρίσκεται στον άξονα y, και στον άξονα z καταγράφεται η παράμετρος του χρόνου.
Η ανάλυση δεδομένων φασματομετρίας μάζας είναι ένα περίπλοκο αντικείμενο μιας και εξαρτάται από το πείραμα που παράγει τα δεδομένα. Υπάρχουν γενικές υποδιαιρέσεις των δεδομένων που είναι στοιχειώδη για την κατανόηση οποιονδήποτε δεδομένων. Τα φασματόμετρα μάζας δουλεύουν είτε σε λειτουργία αρνητικών ιόντων είτε θετικών ιόντων. Είναι πολύ σημαντικό να ξέρει κάποιος αν τα ιόντα είναι θετικά ή αρνητικά φορτισμένα. Αυτό είναι συχνά σημαντικό για τον προσδιορισμό της ουδέτερης μάζας αλλά μαρτυρά και κάτι για τη φύση των μορίων. Διαφορετικοί τύποι περιοχών ιονισμού παράγουν διαφορετικά θραύσματα από τα ίδια αρχικά μόρια. Κατανοώντας την προέλευση του δείγματος, κάποιες προσδοκίες μπορούν να θεωρηθούν για τα συστατικά των μορίων αλλά και τα θραύσματά τους.
Επειδή η ακριβής δομή ή η ακολουθία των πεπτιδίων ενός μορίου αποκωδικοποιείται μέσω των θραυσμάτων μάζας, η ερμηνεία των φασμάτων μάζας απαιτεί το συνδυασμό διαφόρων τεχνικών. Συνήθως η πρώτη στρατηγική για τον προσδιορισμό μιας άγνωστης ένωσης είναι η σύγκριση του πειραματικού φάσματος μάζας με μια βιβλιοθήκη φασμάτων μάζας. Αν η αναζήτηση δεν αποδόσει, τότε γίνεται χειροκίνητη ερμηνεία ή ερμηνεία μέσω λογισμικού. Η προσομοίωση της διαδικασίας ιονισμού και θραυματοποίησης που συμβαίνει σε ένα φασματογράφο μάζας είναι το αρχικό εργαλείο που προσδιορίζει τη δομή ή την ακολουθία των πεπτιδίων ενός μορίου. Μια a priori δομική πληροφορία θραυματίζεται in silico και το προκύπτον πρότυπο συγκρίνεται με το παρατηρηθέν φάσμα. Μία τέτοια προσομοίωση υποστηρίζεται από μια βιβλιοθήκη θραυματοποίησης που περιέχει δημοσιευμένα πρότυπα γνωστών αντιδράσεων αποσύνθεσης. Λογισμικό που επωφελείται από αυτή την ιδέα έχει αναπτυχθεί για μικρά μόρια και πρωτείνες. Ένας άλλος τρόπος ερμηνείας των φασμάτων μάζας περιλαμβάνει φάσματα με ακριβώς μετρήσιμη μάζα. Μία τιμή m/z με ακέραια προσέγγιση μπορεί να αναπαριστά έναν τεράστιο αριθμό πιθανών δομών ιόντων. Ακριβέστεροι αριθμοί μαζών μειώνουν σημαντικά τον αριθμό των υποψήφιων μοριακών τύπων, αν και κάθε ένας μπορεί ακόμα να αναπαριστά ένα μεγάλο αριθμό δομικά διαφορετικών

ενώσεων. Ένας αλγόριθμος, που ονομάζεται γενήτρια τύπων, υπολογίζει όλους τους τύπους μορίων που μπορούν θεωρητικά να ταιριάζουν με ένα δοσμένο φάσμα μάζας με κάποιο περιθώριο λάθους.

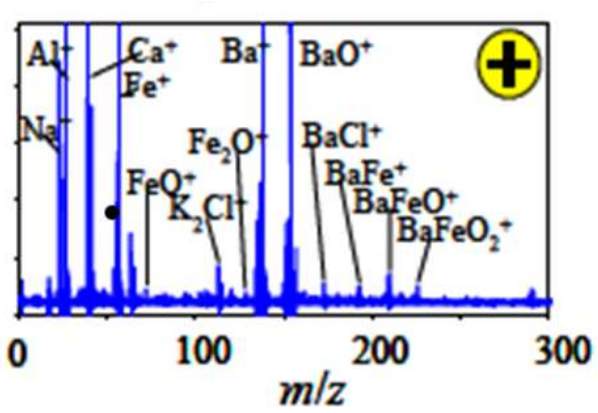
Μια άλλη τεχνική ανάλυσης φασμάτων μάζας που έχει εφαρμοστεί [77][78][79][81], είναι η συσταδοποίηση (clustering) των φασμάτων . Ο σκοπός είναι η ομαδοποίηση των φασμάτων, έτσι ώστε να προσδιοριστούν κοινά πρότυπα της χημικής σύστασης των σωματιδίων που εμφανίζονται στα δεδομένα. Η συσταδοποίηση έχει εφαρμοστεί σε φάσματα μάζας αιωρούμενων σωματιδίων για μελέτη της ατμόσφαιρας, αλλά και σε φασματομετρία (MS/MS) πεπτιδίων[80]. Στα πειράματα MS/MS συχνά δημιουργούνται πλεονάζοντα φάσματα των ίδιων πεπτιδίων. Η συσταδοποίηση των φασμάτων MS/MS εκμεταλεύεται αυτό τον πλεονασμό, προσδιορίζοντας τα πολλαπλά φάσματα του ίδιου πεπτιδίου και αντικαθιστώντας αυτά με έναν αντιπρόσωπο του φάσματος. Η ανάλυση μόνο των αντιπροσώπων φασμάτων έχει ως αποτέλεσμα την σημαντική επιτάχυνση αναζητήσεων σε μια βάση MS/MS.

8 ΣΥΣΤΑΔΟΠΟΙΗΣΗ ΦΑΣΜΑΤΩΝ ΜΑΖΑΣ ΜΕ ΤΟΝ X-Means

Για την ανάλυση φασμάτων μάζας με τεχνικές data mining έχουν αναπτυχθεί εφαρμογές λογισμικού όπως το Enchilada [82] και το YAADA[83] , που είναι toolbox του Matlab. Οι τεχνικές συσταδοποίσης των Enchilada και YAADA περιλαμβάνουν τους αλγόριθμους k-means, k-medians, ART-2a . Στη μελέτη [78] πραγματοποιείται μία σύγκριση των ART-2a και DBSCAN σε συσταδοποίηση φασμάτων μάζας, όπου παρουσιάζονται με παρόμοια αποτελέσματα.

Στη παρούσα εργασία εφαρμόζεται ο αλγόριθμος X-means , ο οποίος συμπεριλαμβάνεται στους αλγόριθμους συσταδοποίησης του Weka . Το βασικό χαρακτηριστικό του αλγορίθμου είναι ότι δε χρειάζεται να δώσουμε τον ακριβή αριθμό του πλήθους των συστάδων εκ των προτέρων.

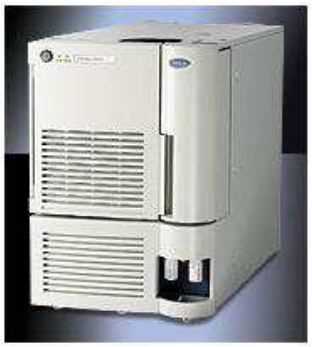
Κάθε φάσμα μάζας είναι ένα n-διάστατο διάνυσμα με κάθε όρισμα να αναπαριστά μια τιμή m/z , δηλαδή ένα ιόν, και η τιμή κάθε ορίσματος είναι η σχετική αφθονία του ιόντος στο φάσμα. Στο σχήμα 17 παρουσιάζεται ένα φάσμα μάζας θετικών ιόντων με εύρος τιμών m/z 1 έως 300.



Σχήμα 17: Φάσμα μάζας με τιμές 1 έως 300

Το δείγμα που θα χρησιμοποιηθεί αποτελείται από 101 διαφορετικά ιόντα, μάζας από 150 έως 250, δήλαδή το dataset αποτελείται από διανύσματα με 101 ορίσματα.

Το δείγμα 1200 περίπου φασμάτων προέρχεται από τον ανιχνευτή μαζών Waters 3100. Ο Waters 3100 είναι ένας μικρός ανιχνευτής μαζών, και λειτουργεί με ιονισμό ατμοσφαιρικής πίεσης απλού τετραπόλου. Είναι σχεδιασμένος ειδικά για αναλύσεις υγρής χρωματογραφίας LS/MS και αποτελεί μία αξιόπιστη λύση για τα εργαστήρια.



Σχήμα 18: Ο φασματογράφος μάζας Waters 3100

Το dataset μετατράπηκε από μορφή αρχείου .csv σε μορφή αρχείου .arff για επεξεργασία από το Weka. Στη συνέχεια τα φάσματα μάζας κανονικοποιήθηκαν (σχήμα 19)για να βεβαιώσουμε ότι η συσταδοποίηση να λάβει υπόψη το σχετικό μέγεθος των κορυφών και όχι το απόλυτο. Ως μέγιστο πλήθος συστάδων για την εκτέλεση του αλγορίθμου επιλέγεται το 10 και απόσταση μεταξύ των διανυσμάτων η ευκλείδια απόσταση. Αν το αποτέλεσμα είναι 10 συστάδες τότε μεγαλώνουμε το εύρος

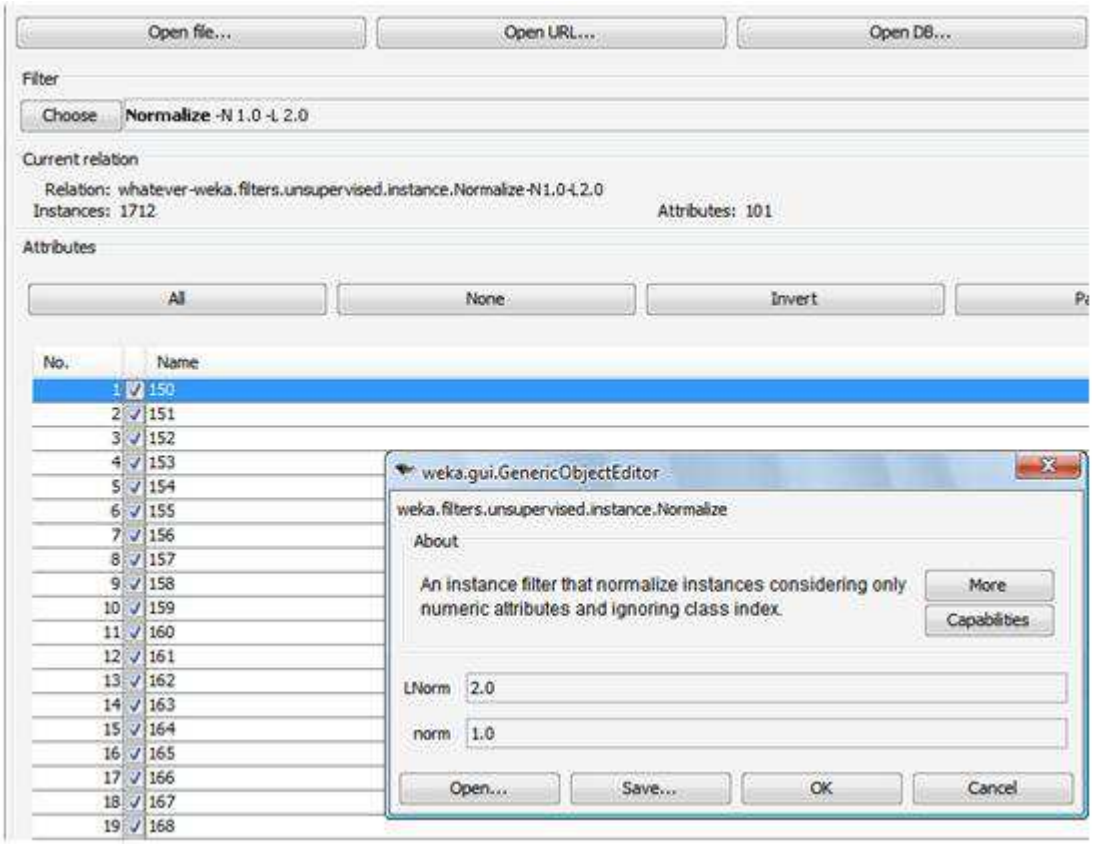
Το σύνολο των συστάδων που προκύπτουν είναι 4, όπου το 63% (1074) των φασμάτων ανήκουν στη δεύτερη συστάδα.

Στο σχήμα 22 εμφανίζεται το φάσμα μάζας της δεύτερης συστάδας με όλες τις κορυφές και στο σχήμα 23 το φάσμα με κορυφές αυτές που έχουν σχετική αφθονία μεγαλύτερη του 0,05. Οι κορυφές με τη μεγαλύτερη αφθονία έχουν τιμές m/z 158, 196, 214.

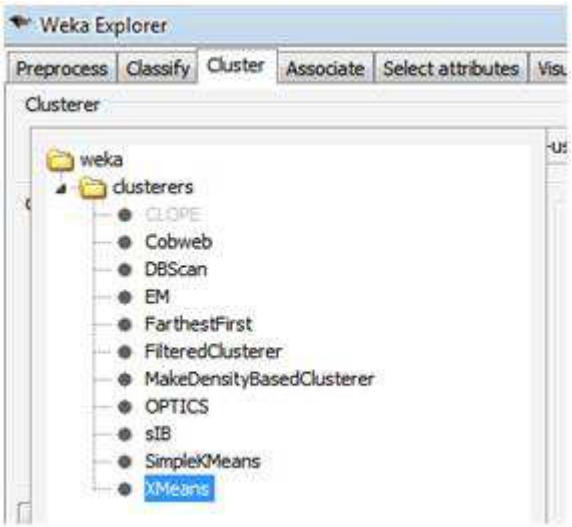
Για να μπορεί να ερμηνεύσει κάποιος πλήρως το αποτέλεσμα της συσταδοποίησης πρέπει να γνωρίζει καλά το δείγμα καθώς και το μηχανισμό παραγωγής των ιόντων. Το γεγονός ότι το 63% των φασμάτων ανήκουν στην ίδια συστάδα σημαίνει ότι το δείγμα δείγμα έχει παρόμοια χημική σύσταση. Ένας απλός και γρήγορος τρόπος για να έχουμε μια εκτίμηση για τη χημική σύσταση των συστάδων, είναι η χρήση ειδικού λογισμικού. Αυτό που χρησιμοποιήθηκε είναι η demo έκδοση του προγράμματος "Mass Spec Calc Pro"(MSCP). Το MSCP δέχεται ως παραμέτρους την τιμή της μάζας καθώς και τα χημικά στοιχεία της χημικής ένωσης. Τα στοιχεία αυτά μπορεί να είναι ένας συνδυασμός των C, H, O, N, Cl, Si, S . Με δεδομένα λοιπόν τη μάζα προς το φορτίο και ότι τα στοιχεία της χημικής ένωσης είναι ένας συνδυασμός των C, H, O, N , τότε εμφανίζονται όλοι οι δυνατοί συνδυασμοί των στοιχείων αυτών ως ιόντα χημικών ενώσεων με τη δεδομένη τιμή m/z .

Στο σχήμα 25 εμφανίζονται τα ιόντα μάζας 158, στο σχήμα 26 εμφανίζονται τα ιόντα μάζας 196

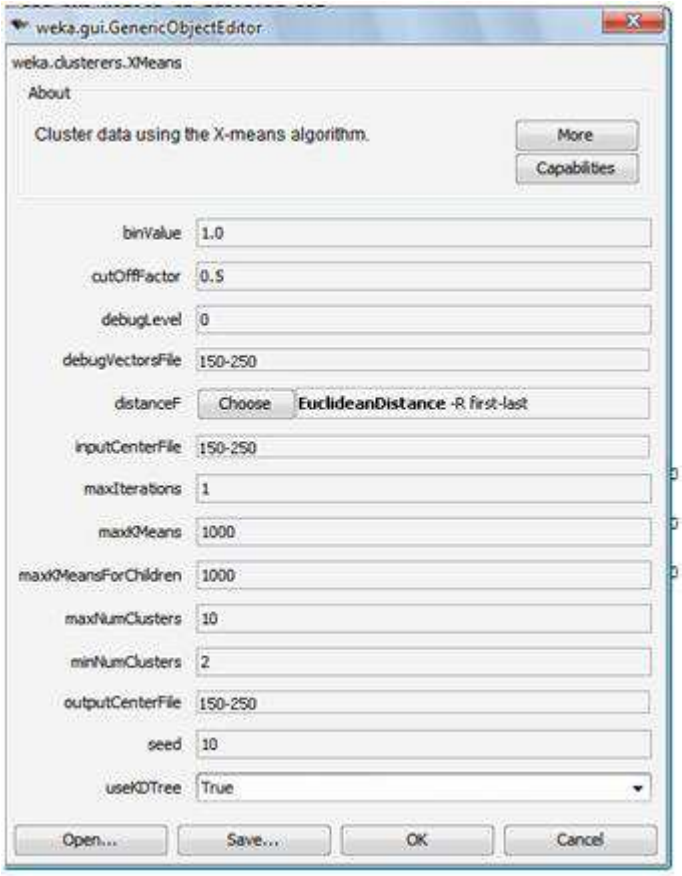
και στο σχήμα 27 τα ιόντα μάζας 214. Αυτές είναι οι κυριότερες τιμές m/z που εμφανίζονται. Οι φασματογράφοι μάζας καταγράφουν διακριτές τιμές m/z . Γι αυτό το λόγο στα πιθανά ιόντα εμφανίζεται και η διαφορά της πραγματικής τους μάζας από τη διακριτή τιμή. Η αφθονία μιας τιμής m/z δεν είναι απαραίτητο να οφείλεται μόνο σε ένα ιόν αλλά και σε περισσότερα.
Έχει ενδιαφέρον να συγκρίνουμε το αποτέλεσμα της συσταδοποίησης του Xmeans με τον απλό K-means. Ως αριθμό συστάδων για τον K-means θα χρησιμοποιηθεί το πλήθος (4) το συστάδων που προέκυψε από την εκτέλεση του αλγορίθμου X-means για το ίδιο dataset. Στο φάσμα μάζας της επικρατέστερης συστάδας εντοπίζονται κορυφές για τις ίδιες τιμές m/z (158,195,196,214).Παρατηρούμε λοιπόν ότι το αποτέλεσμα είναι ίδιο και για τους δύο αλγόριθμους. Το βασικό όμως πλεονέκτημα του X-means έναντι του K-means είναι η δυνατότητα εντοπισμού ταυτόχρονα των συστάδων αλλά και το πλήθος τους. Ο X-means αποτελεί μία καλή και αξιόπιστη λύση Clustering σε πολλούς τομείς.



Σχήμα 19: Κανονικοποίηση των φασμάτων

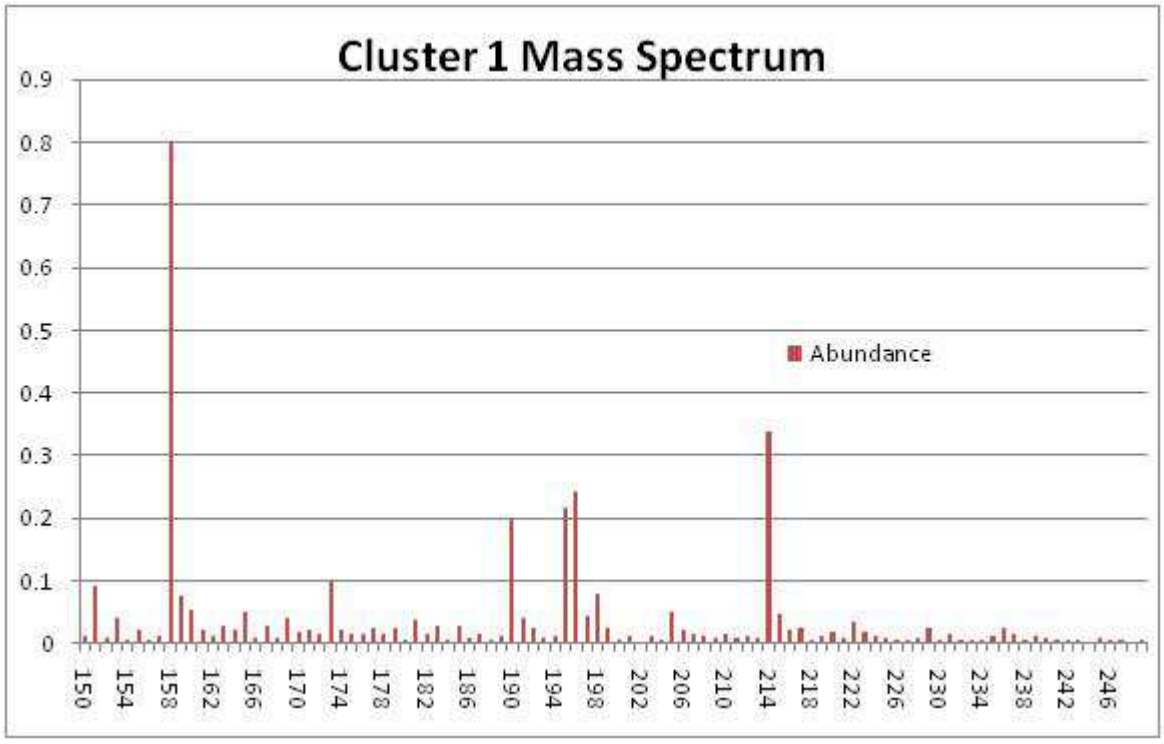


Σχήμα 20: Επιλογή του Xmeans

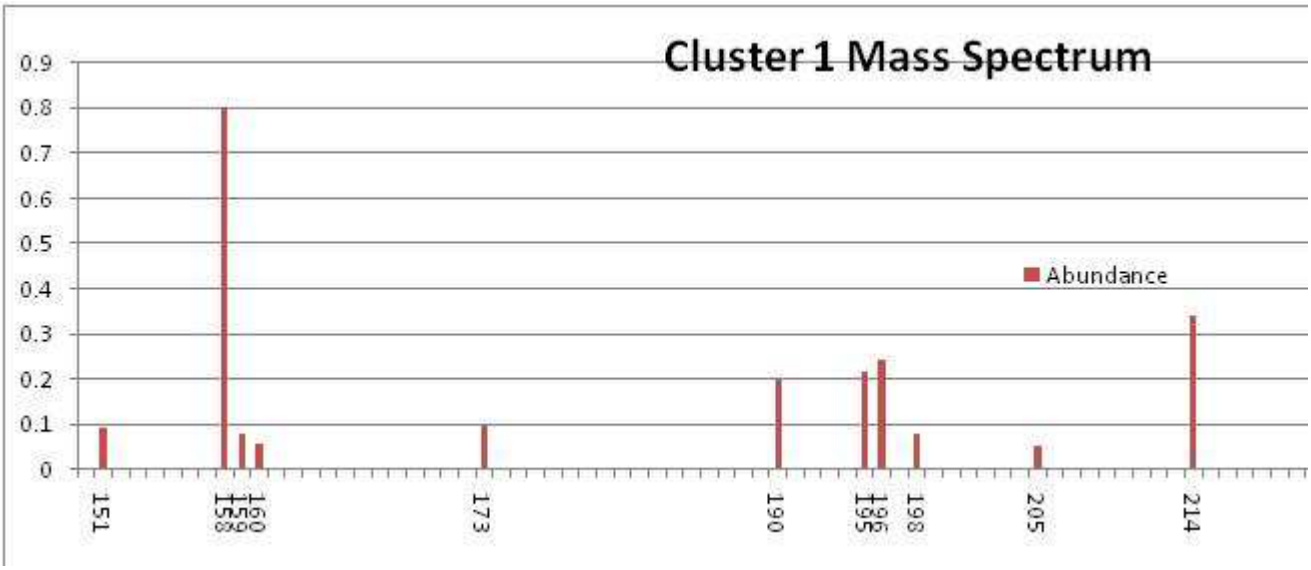


Σχήμα 21: Επιλογή παραμέτρων

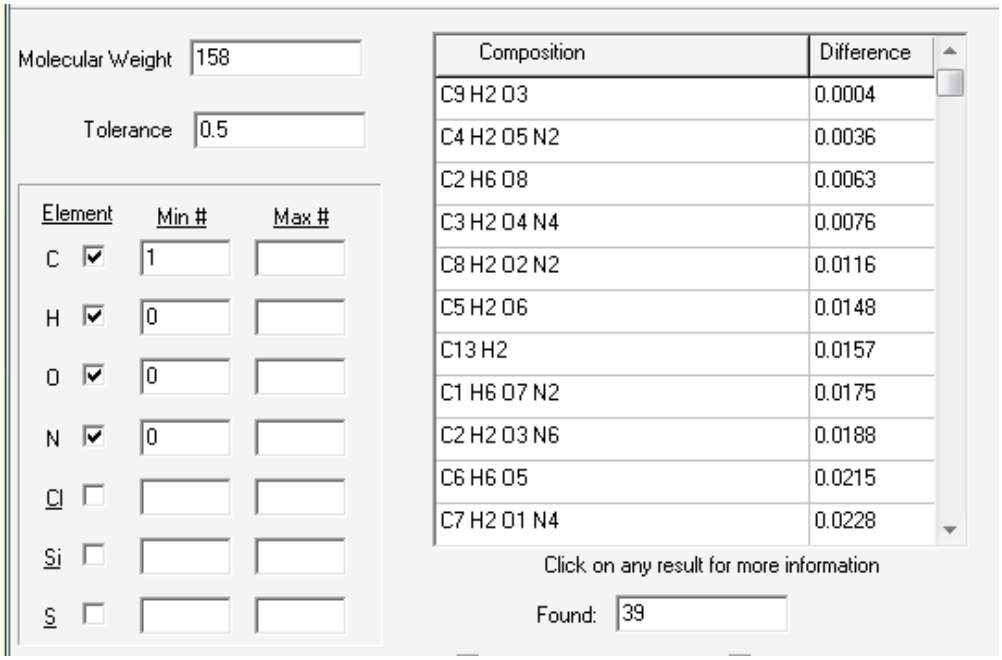
Σχήμα 22: Το αποτέλεσμα του αλγορίθμου είναι 4 συστάδες φασμάτων



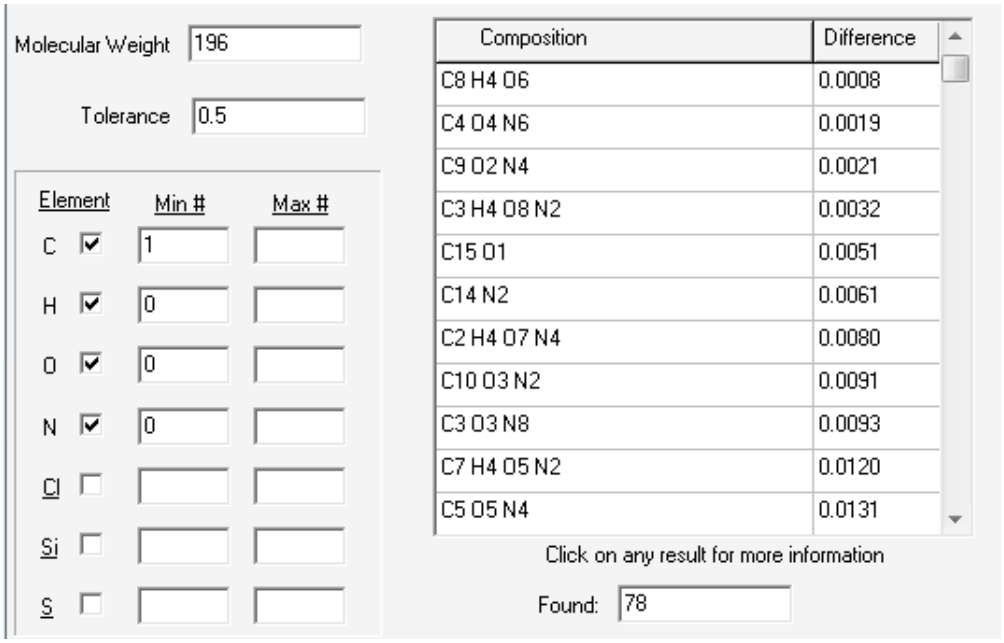
Σχήμα 23: Το φάσμα μάζας της δεύτερης συστάδας



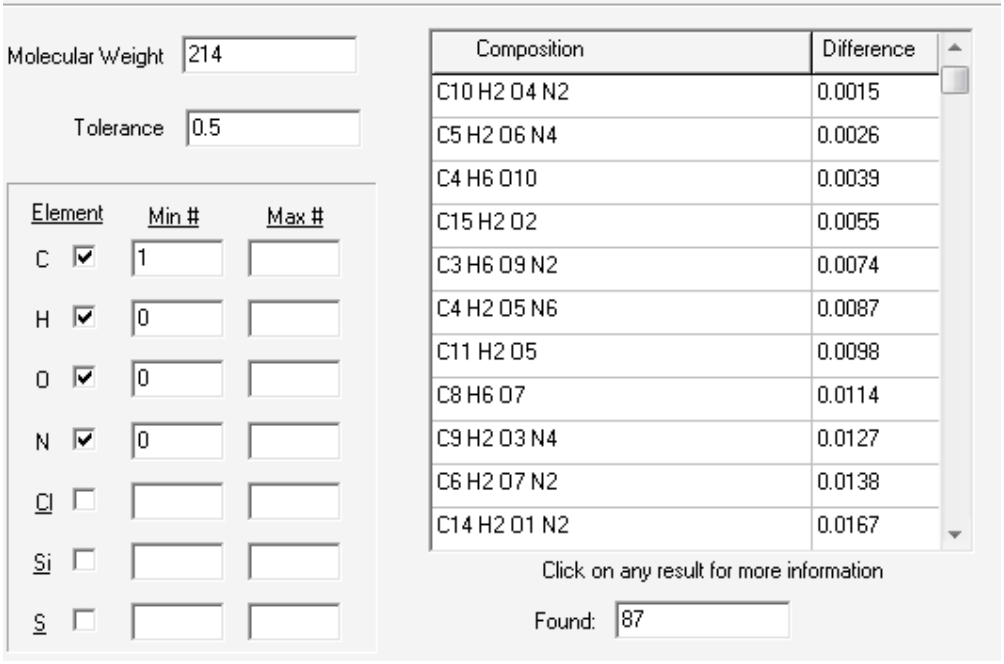
Σχήμα 24: Το φάσμα μάζας της δεύτερης συστάδας με τις πιο σημαντικές κορυφές



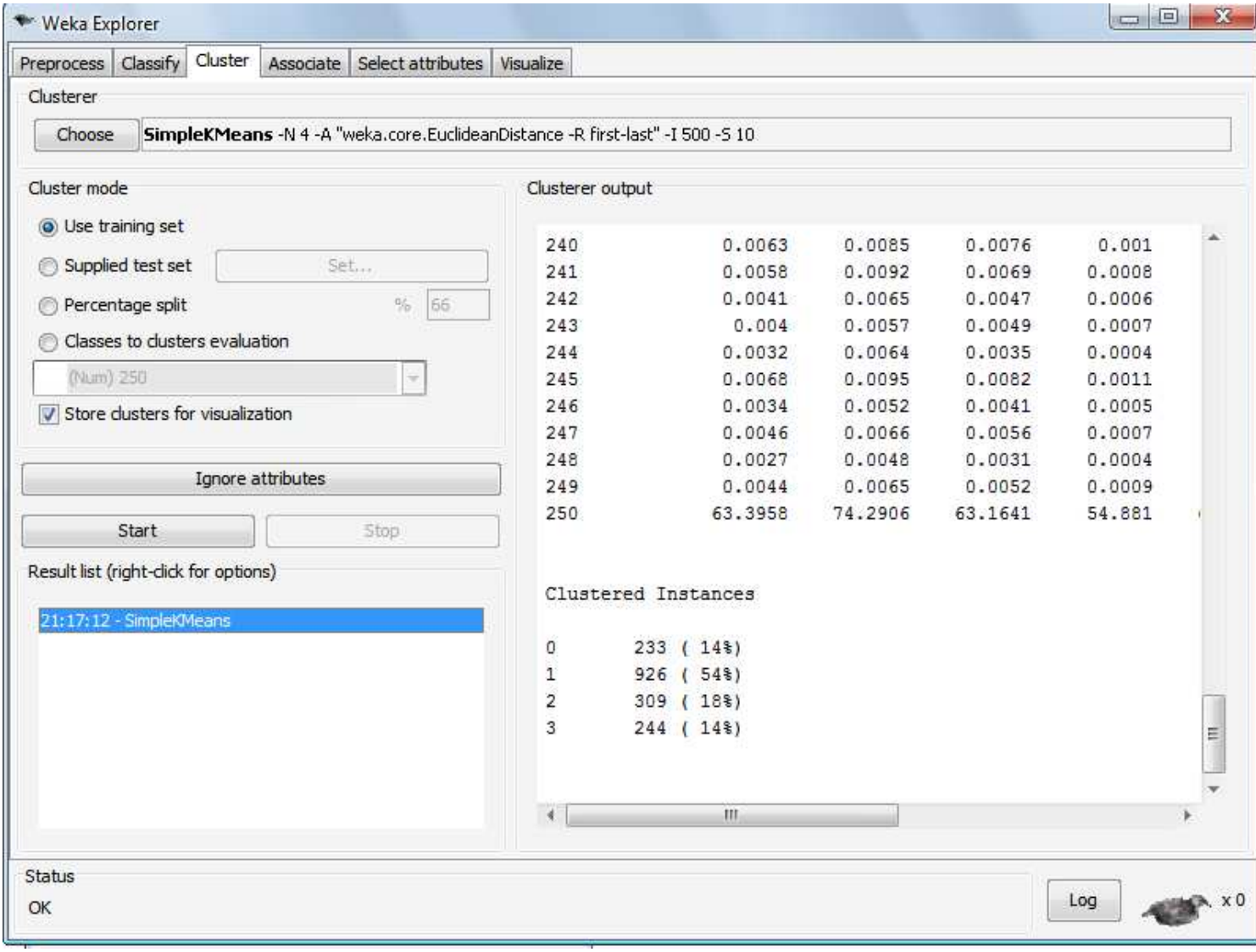
Σχήμα 25: Ιόντα μάζας 158



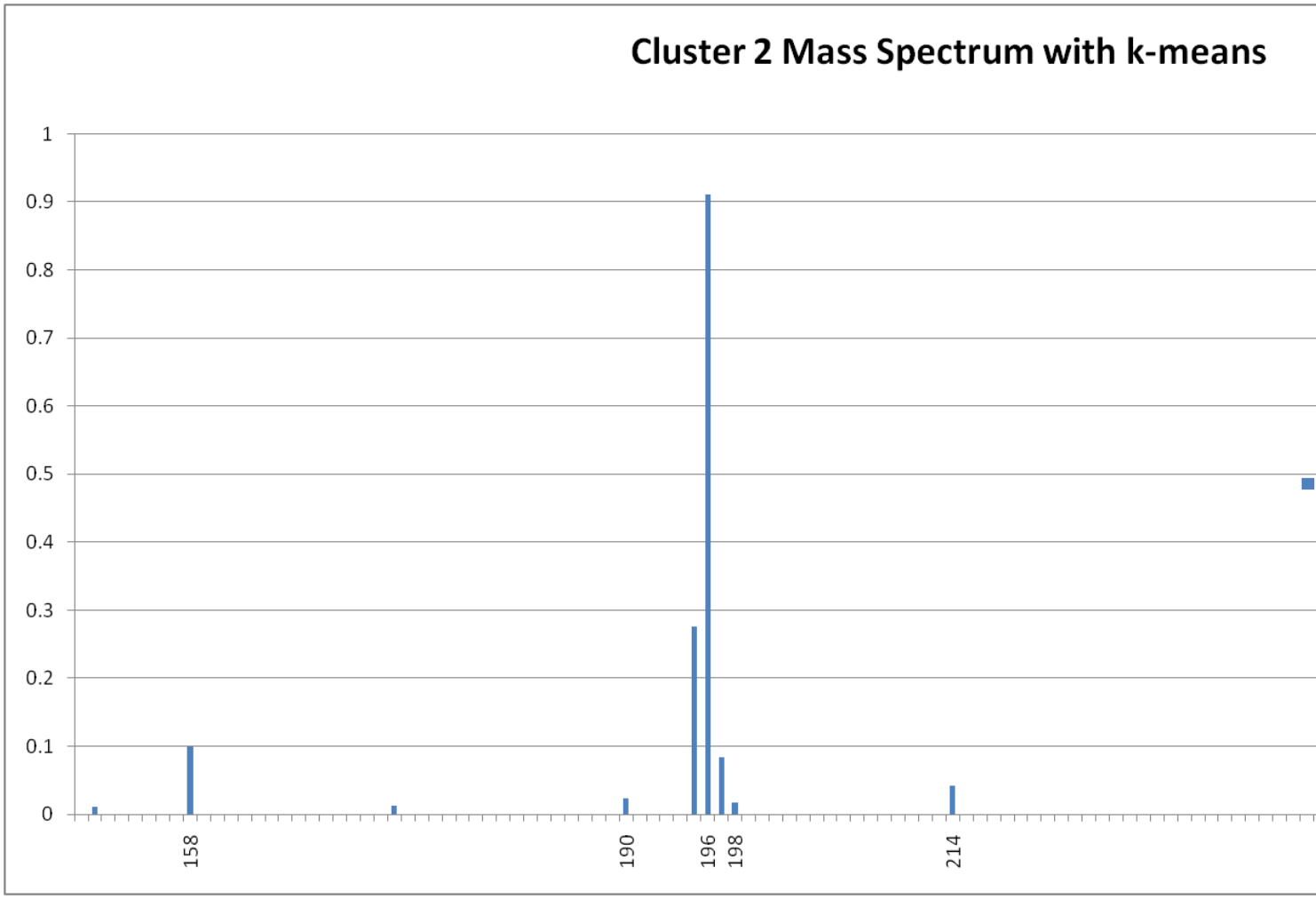
Σχήμα 26: Ιόντα μάζας 196



Σχήμα 27: Ιόντα μάζας 214



Σχήμα 28: Οι συστάδες με τον αλγόριθμο k-means



Σχήμα 29: Το φάσμα μάζας της επικρατέστερης συστάδας με τον k-means

Αναφορές

[1] Data Mining Introductory and Advanced Topics by Margaret H. Dunham,ISBN 9788177587852

[2] Usama Fayyad, Gregor Piatetsky-Shapiro, and Padhraic Smyth. The KDD process for extracking useful knowlwdge from volumes of data. Journal of the ACM, 39(11):27-34, November 1996

[3] Usama Fayyad, Gregor Piatetsky-Shapiro, and Padhraic Smyth.From data mining to knowlwdge discovery:An overview. In Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy, eds., Advances in Knowledge Discovery and Data Mining, pages 1-34, AAAI/MIT Press, 1996

[4] Jain, A.K. and Dubes, R.C. 1988. Algorithms for Clustering Data. Prentice-Hall advanced reference series. Prentice-Hall, Inc., Upper Saddle River, NJ

[5] S. Theodoridis and K. Koutroumbas. Pattern Recognition. Academic Press, 1999.

[6] H. Romesburg. Cluster analysis for researchers. Lifetime Learning Publications,1984.

[7] C.D. Michener R.R. Sokal. A statistical method for evaluating systematic relationships.U. Kansas Sci. Bull., 38:1409-1438, 1958.

[8] L. Kaufman and P. Rousseeuw. Finding groups in data: An introduction to cluster analysis. John Wiley Sons, 1990.

[9] Survey of Clustering Algorithms, Rui Xu; Wunsch, D., II;Dept. of Electr. Comput. Eng., Univ. of Missouri-Rolla, Rolla, MO, USA,Neural Networks, IEEE Transactions on May 2005,Volume: 16,Issue:3,On page(s): 645 - 678,ISSN: 1045-9227

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

[11] Estimating the number of clusters using a windowing technique,B. Boutsinas, D. K. Tasoulis and M. N. Vrahatis,Department of Business Administration and Artificial Intelligence Research Center (UPAIRC), University of Patras, Thursday, June 15, 2006,

Pattern Recognition and Image AnalysisCoverageVolume 16 / 2006 - Volume 20 / 2010, ISSN: 1054-6618

[12] Huang, Z.; Chen, L.; Cai, J.-Y.; Gross, D.; Musicant, D.; Ramakrishnan, R.; Schauer, J.; Wright, S.J.;Wisconsin Univ., Madison, WI, USA, Data Mining, 2004. ICDM '04. Fourth IEEE International Conference on 1-4 Nov. 2004,page(s): 122 - 129,ISBN: 0-7695-2142-8

[13] Under a Creative Commons License. Atmospheric Chemistry and Physics Cluster Analysis of the Organic Peaks in Bulk Mass Spectra Obtained During the 2002 New England Air Quality Study with an Aerodyne Aerosol Mass Spectrometer (2006) by C. Marcolli , M. R. Canagaratna , D. R. Worsnop , R. Bahreini , J. A. De Gouw , C. Warneke , P. D. Goldan , W. C. Kuster , E. J. Williams , B. M. Lerner , J. M. Roberts , J. F. Meagher , F. C. Fehsenfeld , M. Marchewka , S. B. Bertman , A. M. Middlebrook

[14] X-means: Extending K-means with Efficient Estimation of the Number of Clusters (2000) by Dau Pelleg , Andrew Moore In Proceedings of the 17th International Conf. on Machine Learning

[15] Data clustering: a review by: A. K. Jain, M. N. Murty, P. J. Flynn ACM Comput. Surv., Vol. 31, No. 3. (September 1999), pp. 264-323.

[16] B. Everitt, S. Landau, and M. Leese, Cluster Analysis. London: Arnold, 2001.

[17] L. Kaufman and P. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis: Wiley, 1990.

[18] S. Guha, R. Rastogi, and K. Shim, "CURE: An efficient clustering algorithm for large databases," in Proc. ACM SIGMOD Int. Conf. Management of Data, 1998, pp. 73-84.

[19] ROCK: A robust clustering algorithm for categorical attributes," Inf. Syst., vol. 25, no. 5, pp. 345-366, 2000.

[20] G. Karypis, E. Han, and V. Kumar, "Chameleon: Hierarchical clustering using dynamic modeling," IEEE Computer, vol. 32, no. 8, pp. 68-75, Aug. 1999.

[21] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: An efficient data clustering method for very large databases," in Proc. ACM SIGMOD Conf. Management of Data, 1996, pp. 103-114.

[22] G. Liu, Introduction to Combinatorial Mathematics. New York: Mc- Graw-Hill, 1968.

[23] K. Stoffel and A. Belkoniene, "Parallel K-means clustering for large data sets," in Proc. EuroPar"99 Parallel Processing, 1999, pp.1451-1454.

[24] G. Ball and D. Hall, "A clustering technique for summarizing multivariate data," Behav. Sci., vol. 12, pp. 153-155, 1967.

[25] Z. Huang, "Extensions to the K-means algorithm for clustering large data sets with categorical values," Data Mining Knowl. Discov., vol. 2,pp. 283-304, 1998.

[26] S. Gupata, K. Rao, and V. Bhatnagar, "K-means clustering algorithm for categorical attributes," in Proc. 1st Int. Conf. Data Warehousing and Knowledge Discovery (DaWaK"99), Florence, Italy, 1999, pp. 203-208.

[27] P. Hansen and N. Mladenoviz, "J-means: A new local search heuristic for minimum sum of squares clustering," Pattern Recognit., vol. 34, pp. 405-413, 2001.

[28] T. Kanungo, D. Mount, N. Netanyahu, C. Piatko, R. Silverman, and A. Wu, "An efficient K-means clustering algorithm: Analysis and implementation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 7, pp. 881-892, Jul. 2000.

[29] G. Patanj and M. Russo, "The enhanced-LBG algorithm," Neural Netw., vol. 14, no. 9, pp. 1219-1237, 2001.

[30] Fully automatic clustering system," IEEE Trans. Neural Netw., vol. 13, no. 6, pp. 1285-1298, Nov. 2002.

[31] M. Su and C. Chou, "A modified version of theK-means algorithm with a distance based on cluster symmetry," IEEE Trans. Pattern Anal. Mach. Intell., vol. 23, no. 6, pp. 674-680, Jun. 2001.

[32] K. Wagstaff, S. Rogers, and S. Schroedl, "Constrained K-means clustering with background knowledge," in Proc. 8th Int. Conf. Machine Learning, 2001, pp. 577-584.

[33] X. Zhuang, Y. Huang, K. Palaniappan, and Y. Zhao, "Gaussian mixture density modeling, decomposition, and applications," IEEE Trans. Image Process., vol. 5, no. 9, pp. 1293-1302, Sep. 1996.

[34] R. Duda, P. Hart, and D. Stork, Pattern Classification, 2nd ed. New York: Wiley, 2001.

[35] Model-Based clustering, discriminant analysis, and density estimation," J. Amer. Statist. Assoc., vol. 97, pp. 611-631, 2002.

[36] M. Figueiredo and A. Jain, "Unsupervised learning of finite mixture models," IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 3, pp. 381-396, Mar. 2002.

[37] J. Friedman, "Exploratory projection pursuit," J. Amer. Statist. Assoc., vol. 82, pp. 249-266, 1987.

[38] G. McLachlan and T. Krishnan, The EM Algorithm and Extensions. New York: Wiley, 1997.

[39] G. Celeux and G. Govaert, "A classification EM algorithm for clustering and two stochastic versions," Comput. Statist. Data Anal., vol. 14, pp. 315-332, 1992.

[40] F. Harary, Graph Theory. Reading, MA: Addison-Wesley, 1969

[41] A. Jain and R. Dubes, Algorithms for Clustering Data. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[42] V. Estivill-Castro and I. Lee, "AMOEBA: Hierarchical clustering based on spatial proximity using Delaunay diagram," in Proc. 9th Int. Symp. Spatial Data Handling (SDH"99), Beijing, China, 1999, pp. 7a.26-7a.41.

[43] R. Sharan and R. Shamir, "CLICK: A clustering algorithm with applications to gene expression analysis," in Proc. 8th Int. Conf. Intelligent Systems for Molecular Biology, 2000, pp. 307-316.

[44] A. Ben-Dor, R. Shamir, and Z. Yakhini, "Clustering gene expression patterns," J. Comput. Biol., vol. 6, pp. 281-297, 1999.

[45] G. Liu, Introduction to Combinatorial Mathematics. New York: Mc- Graw-Hill, 1968.

[46] U. Maulik and S. Bandyopadhyay, "Genetic algorithm-based clustering technique," Pattern Recognit., vol. 33, pp. 1455-1465, 2000.

[47] K. Krishna and M. Murty, "Genetic K-means algorithm," IEEE Trans.

[48] L. Tseng and S. Yang, "A genetic approach to the automatic clustering problem," Pattern Recognit., vol. 34, pp. 415-424, 2001.

[49] J. Holland, Adaption in Natural and Artificial Systems. Ann Arbor,MI: Univ. Michigan Press, 1975.

[50] L. Hall, I. Fzyurt, and J. Bezdek, "Clustering with a genetically optimized approach," IEEE Trans. Evol. Comput., vol. 3, no. 2, pp. 103-112, 1999.

[51] D. Fogel, "An introduction to simulated evolutionary optimization," IEEE Trans. Neural Netw., vol. 5, no. 1, pp. 3-14, Jan. 1994.

[52] M. Cowgill, R. Harvey, and L. Watson, "A genetic algorithm approach to cluster analysis," Comput. Math. Appl., vol. 37, pp. 99-108, 1999.

[53] Alex Berson and Stephen J. Smith. Data Warehousing, Data Mining, and OLAP. McGraw-Hill, 1997.

[54] P. Sneath, "The application of computers to taxonomy," J. Gen. Microbiol., vol. 17, pp. 201-226, 1957.

[55] T. Sorensen, "A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyzes of the vegetation on Danish commons," Biologiske Skrifter, vol. 5, pp. 1-34, 1948.

[56] E. Forgy, "Cluster analysis of multivariate data: Efficiency vs. interpretability of classifications," Biometrics, vol. 21, pp. 768-780, 1965.

[57] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in Proc. 5th Berkeley Symp., vol. 1, 1967, pp. 281-297.

[58] Z. Huang, "Extensions to the K-means algorithm for clustering large data sets with categorical values," Data Mining Knowl. Discov., vol. 2, pp. 283-304, 1998.

[59] C. Ordonez and E. Omiecinski, "Efficient disk-based K-means clustering for relational databases," IEEE Trans. Knowl. Data Eng., vol. 16, no. 8, pp. 909-921, Aug. 2004.

[60] P. Bradley, U. Fayyad, and C. Reina, "Scaling clustering algorithms to large databases," in Proc. 4th Int. Conf. Knowledge Discovery and Data Mining (KDD"98), 1998, pp. 9-15.

[61] Clustering very large databases using EM mixture models," in Proc. 15th Int. Conf. Pattern Recognition, vol. 2, 2000, pp. 76-80.

[62] A. Hinneburg and D. Keim, "An efficient approach to clustering in large multimedia databases with noise," in Proc. 4th Int. Conf. Knowledge Discovery and Data Mining (KDD"98), 1998, pp. 58-65.

[63] M. Ester, H. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in Proc. 2nd Int. Conf. Knowledge Discovery and Data Mining (KDD"96), 1996, pp. 226-231.

[64] R. Ng and J. Han, "CLARANS: A method for clustering objects for spatial data mining," IEEE Trans. Knowl. Data Eng., vol. 14, no. 5, pp. 1003-1016, Sep.-Oct. 2002.

[65] G. Sheikholeslami, S. Chatterjee, and A. Zhang, "WaveCluster: A multiresolution clustering approach for very large spatial databases," in Proc. 24th VLDB Conf., 1998, pp. 428-439.

[66] C. Olson, "Parallel algorithms for hierarchical clustering," Parallel Comput., vol. 21, pp. 1313-1325, 1995.

[67] K. Stoffel and A. Belkoniene, "Parallel K-means clustering for large data sets," in Proc. EuroPar"99 Parallel Processing, 1999, pp. 1451-1454.

[68] S. Eschrich, J. Ke, L. Hall, and D. Goldgof, "Fast accurate fuzzy clustering through data reduction," IEEE Trans. Fuzzy Syst., vol. 11, no. 2, pp. 262-270, Apr. 2003.

[69] E. Dahlhaus, "Parallel algorithms for hierarchical clustering and applications to split decomposition and parity graph recognition," J. Algorithms, vol. 36, no. 2, pp. 205-240, 2000.

[70] G. Carpenter and S. Grossberg, "A massively parallel architecture for a self-organizing neural pattern recognition machine," Comput. Vis. Graph. Image Process., vol. 37, pp. 54-115, 1987.

[71] The ART of adaptive pattern recognition by a self-organizing neural network," IEEE Computer, vol. 21, no. 3, pp. 77-88, Mar. 1988.

[72] D. Barbara and P. Chen, "Using the fractal dimension to cluster datasets," in Proc. 6th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2000, pp. 260-264.

[73] V. Ganti, R. Ramakrishnan, J. Gehrke, A. Powell, and J. French, "Clustering large datasets in arbitrary metric spaces," in Proc. 15th Int. Conf. Data Engineering, 1999, pp. 502-511.

[74] http://en.wikipedia.org/wiki/Mass$_s pectrometry$

[75] Mass Spectrometry, Chemie Laboratory Ntes, University of Crete, Panos Papagianakopoulos, Giannis Lazarou

[76] http://masspec.scripps.edu/

[77] Cluster analysis of the organic peaks in bulk mass spectra obtained during the 2002 New England Air Quality Study with an Aerodyne aerosol mass spectrometer C. Marcolli1, M. R. Canagaratna2, D. R. Worsnop, R. Bahreini,et al. Atmos. Chem Phys Discuss 6, 4601-4641,2006

[78] Cluster analysis of single particle mass spectra measured at Flushing, NY Liming Zhou, Philip K. Hopke ., Prasanna Venkatachari Center for Air Resources Engineering and Science, Department of Chemical Engineering, Clarkson University, P.O. Box 5708, Potsdam, NY 13699-5708, USA

[79] User-Friendly Clustering for Atmospheric Data Analysis ,Benjamin J. Anderson David R. Musicant Anna M. Ritz,Andrew Ault Deborah Gross Melanie Yuen Dept. of Chemistry Carleton College Northfield, MN 55057,Markus Gdlli TSI, Inc. 500 Cardigan Road Shoreview

[80] Clustering Millions of Tandem Mass Spectra Ari M. Frank,*,- Nuno Bandeira,- Zhouxin Shen,- Stephen Tanner,- Steven P. Briggs,- Richard D. Smith,and Pavel A. Pevzner,

Department of Computer Science and Engineering, University of California, San Diego, La Jolla, California 92093-0404, Department of Biology, University of California, San Diego, La Jolla, California 92093-0346, Bioinformatics Program, University of California, San Diego, La Jolla, California 92093-0419, and Biological Sciences Division, Pacific

[81] The EDAM Project: Mining Atmospheric Aerosol Datasets 1 Raghu Ramakrishnan, James J.Schauer, Lei Chen, Zheng Huang, Martin M. Shafer

[82] http://pages.cs.wisc.edu/ chenl/edam2/edamOverview.pdf

[83] http://www.yaada.org/description.html

[84] http://www.cs.waikato.ac.nz/ml/weka/