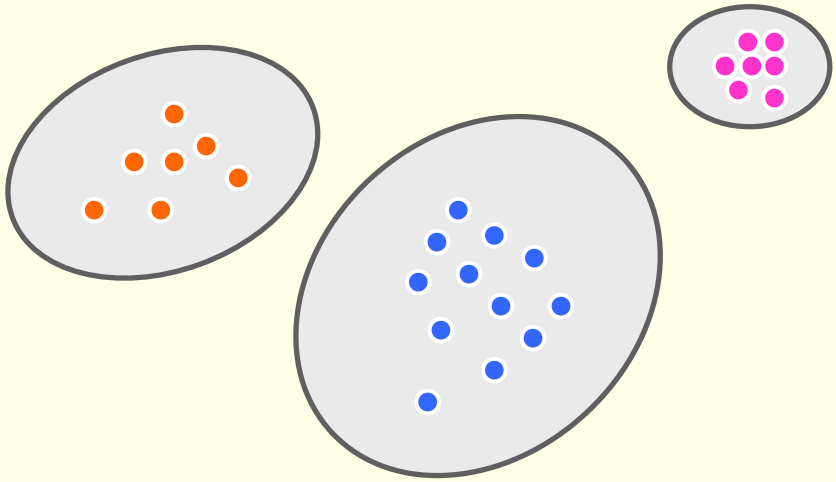
Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Αγρονόμων και Τοπογράφων Μηχανικών
Διατμηματικό Πρόγραμμα Μεταπτυχιακών Σπουδών
«ΓΕΩΠΛΗΡΟΦΟΡΙΚΗ»

Αλγόριθμοι Εξόρυξης Χωρικών Δεδομένων



Εφαρμογή σε Αλγόριθμους Συσταδοποίησης

Διπλωματική Εργασία
Κεχαγιά – Παρδάλη Ευθαλία

Υπεύθυνος Καθηγητής: Σελλής Τιμολέων

Αθήνα 2006

Πρόλογος

Η παρούσα διπλωματική εργασία εκπονήθηκε στα πλαίσια του Διατμηματικού Προγράμματος Μεταπτυχιακών Σπουδών «ΓΕΩΠΛΗΡΟΦΟΡΙΚΗ» της Σχολής Αγρονόμων και Τοπογράφων Μηχανικών του Εθνικού Μετσοβίου Πολυτεχνείου. Στοχεύει στην μελέτη και υλοποίηση ενδεικτικών αλγορίθμων για την αναγνώριση προτύπων από σημειακά δεδομένα, με στόχο κυρίως την ανεύρεση συσχετίσεων μεταξύ τους και την κατάταξή τους σε συγκεκριμένες κατηγορίες ή περιοχές.

Η εξόρυξη γνώσης από μεγάλο όγκο χωρικών δεδομένων (*spatial data mining*) αποσκοπεί στην ανακάλυψη κρυμμένων συσχετίσεων και χαρακτηριστικών που ενυπάρχουν στα στοιχεία. Ανάλογα με τη φύση των δεδομένων και τις ανάγκες της εφαρμογής μπορεί να απαιτείται διαφορετική επεξεργασία. Ειδικότερα, η μελέτη εστιάζει στους αλγόριθμους συσταδοποίησης χωρικών δεδομένων. Είναι σαφές πως το αντικείμενο δεν μπορεί να μελετηθεί διεξοδικά στο πλαίσιο μιας διπλωματικής εργασίας. Η εξόρυξη γνώσης από χωρικά – και όχι μόνο – δεδομένα με τη βοήθεια αλγορίθμων αποτέλεσε και αποτελεί αντικείμενο έρευνας από μια πληθώρα ερευνητικών ομάδων.

Στο σημείο αυτό θα ήθελα να ευχαριστήσω θερμά τον καθηγητή κ. Τίμο Σελλή για την ανάθεση και επίβλεψη της παρούσας εργασίας καθώς και για την βοήθειά του καθ' όλη τη διάρκεια εκπόνησης αυτής.

Ιδιαίτερες ευχαριστίες οφείλω στον κ. Κώστα Πατρούμπα, υποψήφιο διδάκτορα ΕΜΠ για την πολύτιμη συμβολή, υποστήριξη και συμπαράσταση που μου παρείχε, καθώς και για τον χρόνο που μου αφιέρωσε.

Σε προσωπικό επίπεδο θα ήθελα να ευχαριστήσω τους γονείς μου και τον κ. Δημήτρη Μαυρογιώργο για την υπομονή, την συμπαράσταση, την κατανόηση και την αγάπη τους.

Αθήνα, Μάϊος 2006

Περίληψη

Σε αυτήν την διπλωματική εργασία μελετάται η εξόρυξη γνώσης, ιδιαίτερα της χωρικής, από δεδομένα με την βοήθεια αλγορίθμων. Στόχος είναι η αναγνώριση προτύπων από σημειακά δεδομένα, με σκοπό κυρίως την ανεύρεση συσχετίσεων μεταξύ τους και την κατάταξή τους σε συγκεκριμένες κατηγορίες ή περιοχές.

Τα σύγχρονα συστήματα βάσεων δεδομένων (λ.χ. *Oracle, SQL*Server*) έχουν ήδη εντάξει κάποιες δυνατότητες χειρισμού μη χωρικών στοιχείων (λ.χ. πωλήσεις καταστημάτων). Ωστόσο, η επεξεργασία χωρικών οντοτήτων δύσκολα μπορεί να διεξαχθεί με τέτοια εργαλεία, καθώς η εγγενής φύση τέτοιων δεδομένων παραμένει ανεκμετάλλευτη. Κατά την τελευταία δεκαετία έχουν προταθεί αρκετοί αλγόριθμοι για εξόρυξη χωρικών ή πολυδιάστατων δεδομένων, ορισμένοι από τους οποίους θα μελετηθούν και θα υλοποιηθούν στα πλαίσια αυτής της διπλωματικής εργασίας.

Στο **Κεφάλαιο 1** παρουσιάζεται το ζήτημα της **Εξόρυξης Δεδομένων** (***Data Mining***) ή **Εξόρυξης Γνώσης από Δεδομένα** (*Knowledge Data Mining*). Επιχειρείται μία γενική θεώρηση σημαντικών εννοιών, οι οποίες σχετίζονται με

αυτή την σχετικά καινούρια και ραγδαία εξελισσόμενη επιστημονική περιοχή. Ενδεικτικά αναφέρονται εδώ ορισμένες από αυτές:

- **Κατηγοριοποίηση** (*Classification*)
- **Κανόνες Συσχετίσεων** (*Association Rules*)
- **Παρουσίαση Συνόψεων** (*Summarization*)
- **Συσταδοποίηση** (*Clustering*)
- **Πρόβλεψη** (*Prediction*)
- **Ανάλυση Χρονοσειρών** (*Time Series Analysis*)
- **Παλινδρόμηση** (*Regression*)
- **Κατηγοριοποίηση** (*Classification*).

Ακόμη, γίνεται επισκόπηση των αλγορίθμων οι οποίοι αφορούν στην εξόρυξη δεδομένων, στα πλαίσια των πιο πάνω εργασιών εξόρυξης (χωρικής) γνώσης.

Το **Κεφάλαιο** 2 αφορά στην **Εξόρυξη Χωρικών Δεδομένων**. Παρουσιάζεται το ζήτημα της Εξόρυξης Χωρικής Γνώσης και γίνεται αναφορά στις Χωρικές Ερωτήσεις. Αναπτύσσεται η **Οργάνωση Χωρικών Δεδομένων αναλύοντας τις δομές αυτών, παρουσιάζονται οι Βασικές Αρχές Εξόρυξης Γνώσης από Χωρικά Δεδομένα, καθώς και οι Αλγόριθμοι Εξόρυξης Χωρικών Δεδομένων, στα πλαίσια** της κατηγοριοποίησης, της συσταδοποίησης και των κανόνων χωρικών συσχετίσεων.

Το **Κεφάλαιο** 3 πραγματεύεται την **Συσταδοποίηση Χωρικών Δεδομένων**. Γίνεται μια ανάλυση του θεωρητικού πλαισίου της Συσταδοποίησης, εστιάζοντας στις Αρχές και τις Εφαρμογές της. Ακόμη, αναπτύσσονται οι Μέθοδοι Συσταδοποίησης και γίνεται Συγκριτική Θεώρηση των Αλγορίθμων Συσταδοποίησης που αναπτύχθηκαν στην ενότητα 1.4 του Κεφαλαίου 1. Τέλος, αναπτύσσονται οι αλγόριθμοι Χωρικής Συσταδοποίησης Βασιζόμενης στην Πυκνότητα Εφαρμογών με «Θόρυβο» *(DBSCAN - Density Based Spatial Clustering Of Applications With Noise)*, Προσέγγισης Καννάβου Στατιστικής Πληροφορίας *(STING - Statistical Information Grid Approach)*, και K -

κέντρων (*K – means).*

Στο **Κεφάλαιο 4** γίνεται **Πειραματική Αξιολόγηση** των τριών αλγορίθμων που αναπτύχθηκαν στις ενότητες 3.4.1 έως 3.4.3, αφού προηγηθεί υλοποίησή τους,. Η υλοποίηση γίνεται σε περιβάλλον *Visual Basic 6* και η οπτικοποίηση των αποτελεσμάτων σε περιβάλλον *ARCGIS – ArcMap 8.3.* Προκειμένου για την δοκιμαστική εκτέλεση των προγραμμάτων και την εκτίμηση των αποτελεσμάτων τους, χρησιμοποιούνται ως είσοδοι μικρά σύνολα σημειακών δεδομένων. Τα δεδομένα αυτά αφορούν σε μετρήσεις θερμοκρασίας και σε θέσεις ατυχημάτων στο οδικό δίκτυο.

Στο **Κεφάλαιο 5** εξάγονται **Συμπεράσματα** που αφορούν στο σύνολο της διπλωματικής εργασίας. Τα συμπεράσματα αναφέρονται αφενός στην συνολική βιβλιογραφική επισκόπηση του ζητήματος της εξόρυξης (χωρικής) γνώσης και, αφετέρου, στα ποιοτικά αποτελέσματα της πειραματικής αξιολόγησης των αποτελεσμάτων των υλοποιηθέντων αλγόριθμων συσταδοποίησης. Ειδικότερα, οι εξετασθέντες αλγόριθμοι δίνουν ως έναν βαθμό αξιόπιστα αποτελέσματα τα οποία θα μπορούσαν να αξιοποιηθούν, ενώ η δοκιμαστική τους εκτέλεση επιβεβαιώνει σε αρκετά σημεία το εξετασθέν θεωρητικό υπόβαθρο.

Κεφάλαιο 1
Εξόρυξη Δεδομένων

1.1 Εισαγωγή – Ορισμοί

Η Εξόρυξη Γνώσης ορίζεται ως η εύρεση πληροφοριών που είναι κρυμμένες σε μία βάση δεδομένων, η εξερευνητική ανάλυση δεδομένων, η ανακάλυψη καθοδηγούμενη από δεδομένα και η εξερευνητική μάθηση. Η σημερινή εξέλιξη στις λειτουργίες και στα προϊόντα της εξόρυξης γνώσης από δεδομένα είναι αποτέλεσμα πολλών χρόνων επιρροής από πολλούς επιστημονικούς κλάδους όπως είναι οι βάσεις δεδομένων, η ανάκτηση πληροφοριών, η στατιστική, οι αλγόριθμοι και η μηχανική μάθηση.

Ειδικότερα, πρόκειται για την διαδικασία «ανακάλυψης» ενδιαφερόντων και εν δυνάμει χρήσιμων προτύπων (*patterns*), υπαρκτών σε μεγάλες βάσεις δεδομένων. Ο όρος «εξόρυξη» χρησιμοποιείται προκειμένου να τονισθεί ότι τα πρότυπα συνιστούν ψήγματα πολύτιμης πληροφορίας προς ανακάλυψη, κρυμμένης μέσα σε μεγάλες βάσεις δεδομένων!

Ένα πρότυπο μπορεί να είναι μία στατιστική περίληψη (*summary statistic*), όπως ο μέσος όρος (*mean*), ο αριθμητικός μέσος (*median*), ή η τυπική απόκλιση (*standard deviation*) ενός συνόλου δεδομένων. Μέσω της εξόρυξης γνώσης αναζητούνται ταχύτατα και αυτόματα τοπικά και υψηλής χρησιμότητας πρότυπα, κάνοντας χρήση αλγορίθμων. [SC03]

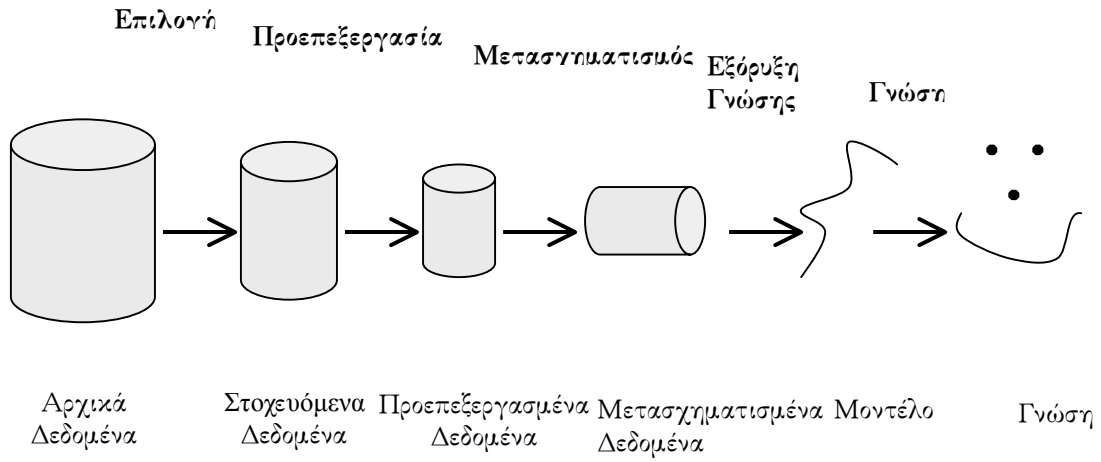
1.2 Εξόρυξη και Ανακάλυψη Γνώσης

Οι όροι «Εξόρυξη Γνώσης από Δεδομένα» και «Ανακάλυψη Γνώσης σε Βάσεις Δεδομένων» (*Knowledge Discovery in Databases*) συχνά χρησιμοποιούνται εναλλακτικά για την ίδια έννοια. Τελευταία, ο όρος «Ανακάλυψη Γνώσης σε Βάσεις Δεδομένων» έχει χρησιμοποιηθεί για να εκφράσει μια διαδικασία που αποτελείται από πολλά βήματα, ένα από τα οποία είναι η εξόρυξη γνώσης από δεδομένα. Έτσι:

- Η ανακάλυψη γνώσης σε βάσεις δεδομένων είναι η διαδικασία εύρεσης χρήσιμων πληροφοριών και προτύπων στα δεδομένα.
- Η εξόρυξη γνώσης από δεδομένα είναι η χρήση αλγορίθμων για την εξαγωγή πληροφοριών και προτύπων που παράγονται με την διαδικασία ανακάλυψης γνώσης σε βάσεις δεδομένων.

Ο ορισμός της ανακάλυψης γνώσης σε βάσεις δεδομένων περιλαμβάνει την λέξη – κλειδί «χρήσιμο». Εφόσον οι πληροφορίες που προκύπτουν από αυτήν την διαδικασία δεν είναι χρήσιμες, τότε δεν είναι στην πραγματικότητα πληροφορίες. Φυσικά το αν κάτι είναι χρήσιμο ή όχι είναι σχετική και υποκειμενική έννοια.

Η ανακάλυψης γνώσης σε βάσεις δεδομένων είναι μια διαδικασία που περιλαμβάνει πολλά διαφορετικά βήματα. Η είσοδος σε αυτή τη διαδικασία είναι τα δεδομένα και οι χρήσιμες πληροφορίες που επιθυμούν οι χρήστες είναι η έξοδος. Όμως, ο αντικειμενικός σκοπός δεν είναι από την αρχή ξεκάθαρος. Η διαδικασία από μόνη της είναι διαδραστική και συνήθως απαιτείται πολύς χρόνος για την ολοκλήρωσή της. Προκειμένου να διασφαλισθεί η χρησιμότητα και η ακρίβεια των αποτελεσμάτων αυτής της διαδικασίας, συνήθως χρειάζεται η συνεργασία ειδικών του πεδίου εφαρμογής με ειδικούς της διαδικασίας ανακάλυψης γνώσης σε βάσεις δεδομένων . Η συνολική διαδικασία της ανακάλυψης γνώσης σε βάσεις δεδομένων φαίνεται παραστατικά πιο κάτω (Σχήμα 1.1):



Σχήμα 1.1 : Διαδικασία Ανακάλυψης Γνώσης σε Βάσεις Δεδομένων
Πηγή: [Dun04]

Τα βήματα της διαδικασίας εξόρυξης γνώσης είναι:

1. Επιλογή: Σε αυτό το βήμα συλλέγονται δεδομένα από διάφορες βάσεις δεδομένων, αρχεία και μη ηλεκτρονικές πηγές.
2. Προεπεξεργασία: Εδώ μπορεί να εκτελεσθούν ενέργειες όπως διόρθωση ή αφαίρεση λανθασμένων δεδομένων, ή συλλογή και εκτίμηση ελλιπών δεδομένων.
3. Μετασχηματισμός: Τα δεδομένα που προέρχονται από διαφορετικές πηγές χρειάζεται να μετατραπούν σε ένα κοινό σχήμα για την περαιτέρω επεξεργασία τους. Κάποια από αυτά ίσως απαιτηθεί να κωδικοποιηθούν, ή να μετασχηματισθούν σε πιο χρήσιμα σχήματα. Μπορεί να ελαττωθούν τα δεδομένα, προκειμένου για την μείωση του αριθμού των τιμών αυτών που θα ληφθούν υπόψη.
4. Εξόρυξη γνώσης από δεδομένα: Με βάση το είδος της εξόρυξης που πρόκειται να εκτελεσθεί, σε αυτό το βήμα εφαρμόζονται αλγόριθμοι στα τροποποιημένα δεδομένα, ώστε να προκύψουν τα προσδοκώμενα αποτελέσματα.
5. Ερμηνεία / αξιολόγηση: Είναι πολύ σημαντικό το πώς θα παρουσιασθούν 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 διεπαφές χρήστη (*Graphic User Interface*).

Αυτές οι διαφορετικές επιρροές από το παρελθόν, οι οποίες οδήγησαν στην ανάπτυξη της περιοχής της εξόρυξης γνώσης από δεδομένα, συντέλεσαν στη δημιουργία διαφορετικών απόψεων για το τι είναι στην πραγματικότητα οι λειτουργίες της εξόρυξης γνώσης:

- Η **επαγωγή** χρησιμοποιείται για να οδηγηθούμε από μία πολύ εξειδικευμένη γνώση σε πιο γενικές πληροφορίες. Αυτό το είδος της τεχνικής συχνά υπάρχει στις εφαρμογές της τεχνητής νοημοσύνης.
- Επειδή ο πρωταρχικός αντικειμενικός στόχος της εξόρυξης γνώσης από δεδομένα είναι να περιγράψει μερικά χαρακτηριστικά ενός συνόλου δεδομένων από ένα γενικό μοντέλο, αυτή η προσέγγιση μπορεί να θεωρηθεί σαν ένα είδος **συμπίεσης**. Εδώ, τα λεπτομερή δεδομένα της βάσης δεδομένων «αφαιρούνται» και συμπιέζονται σε μία μικρότερη περιγραφή των χαρακτηριστικών των δεδομένων που βρίσκονται στο μοντέλο.
- Όπως διατυπώθηκε προηγουμένως, η διαδικασία της εξόρυξης γνώσης από δεδομένα μπορεί να θεωρηθεί από μόνη της σαν ένας τύπος διαδικασίας υποβολής **ερωτήσεων** στη σχετική βάση δεδομένων. Πράγματι, η έρευνα στην εξόρυξη γνώσης από δεδομένα τείνει προς την κατεύθυνση εκείνη όπου αναζητείται ο τρόπος ορισμού μιας ερώτησης εξόρυξης γνώσης και το κατά πόσο μπορεί να αναπτυχθεί μία γλώσσα ερωτήσεων που να περιλαμβάνει τόσους πολλούς διαφορετικούς τύπους επερωτήσεων εξόρυξης γνώσης.
- Η περιγραφή μιας μεγάλης βάσης δεδομένων μπορεί να θεωρηθεί σαν να χρησιμοποιούμε **προσέγγιση** προκειμένου να αποκαλυφθούν κρυμμένες πληροφορίες σχετικές με τα δεδομένα.
- Όταν εργαζόμαστε με μεγάλες βάσεις δεδομένων, η επίδραση του μεγέθους και η ικανότητα ανάπτυξης ενός αφηρημένου μοντέλου μπορούν να θεωρηθούν σαν ένας τύπος προβλήματος **αναζήτησης**. [Dun04]

1.3 Εξόρυξη Γνώσης και Αλγόριθμοι

Η εξόρυξη γνώσης από δεδομένα περιλαμβάνει πολλούς διαφορετικούς αλγορίθμους για να εκπληρωθούν διαφορετικές εργασίες. Όλοι αυτοί οι αλγόριθμοι επιχειρούν να ταιριάξουν ένα μοντέλο στα δεδομένα, εξετάζοντας τα τελευταία και καθορίζοντας το μοντέλο αυτό που είναι το πλησιέστερο στα χαρακτηριστικά τους.

1.3.1 Μοντέλα και Εργασίες στην Εξόρυξη Γνώσης από Δεδομένα

Οι αλγόριθμοι εξόρυξης γνώσης μπορεί να θεωρηθεί ότι αποτελούνται από τρία μέρη:

- Μοντέλο: Ο σκοπός του αλγορίθμου 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 μια τεχνική για να κάνουν αναζήτηση στα δεδομένα

Οι διάφοροι τύποι μοντέλων και ορισμένες από τις πιο συνήθεις εργασίες εξόρυξης γνώσης από δεδομένα που χρησιμοποιούν αυτό το είδος μοντέλου αναλύονται ευθύς αμέσως:

- **Προβλεπτικό Μοντέλο (*Predictive Model*)**: Κάνει μία πρόβλεψη για τις τιμές των δεδομένων, χρησιμοποιώντας γνωστά αποτελέσματα που έχει βρει από άλλα δεδομένα. Η μοντελοποίηση της πρόβλεψης μπορεί να γίνει με βάση τη χρήση ιστορικών δεδομένων. Η πρόβλεψη μπορεί να χρησιμοποιηθεί επίσης για να υποδηλώσει ένα συγκεκριμένο τύπο λειτουργίας εξόρυξης γνώσης από δεδομένα. Προκειμένου να καταστεί σαφής η έννοια του προβλεπτικού μοντέλου, παρατίθεται το εξής παράδειγμα: Η πρόβλεψη μιας πλημμύρας είναι δύσκολο πρόβλημα. Μία προσέγγιση περιλαμβάνει την χρήση οργάνων παρακολούθησης και ελέγχου που έχουν τοποθετηθεί σε

διάφορα σημεία του ποταμού. Αυτά τα όργανα συλλέγουν δεδομένα σχετικά με την πρόβλεψη της πλημμύρας: ύψος της στάθμης του νερού, ποσότητα βροχής, χρόνος, υγρασία, κοκ. Στην συνέχεια μπορεί να προβλεφθεί το ύψος της στάθμης του νερού σε ένα σημείο του ποταμού στο οποίο είναι πιθανό να δημιουργηθεί πλημμύρα, βάσει των δεδομένων που συλλέχθηκαν από αισθητήρες που βρίσκονται στον ποταμό πάνω από το σημείο αυτό. Η πρόβλεψη πρέπει να γίνει σε σχέση με το χρόνο που συλλέχθηκαν τα δεδομένα.

Οι πιο συνηθισμένες εργασίες εξόρυξης γνώσης από δεδομένα που χρησιμοποιούν αυτό το είδος μοντέλου, είναι η κατηγοριοποίηση, η παλινδρόμηση, η ανάλυση χρονολογικών σειρών και η πρόβλεψη:

- **Κατηγοριοποίηση (C*lassifiacation*): Απεικονίζει τα δεδομένα σε προκαθορισμένες ομάδες ή κατηγορίες – κλάσεις (*classes*).** Αναφέρεται συχνά σαν εποπτευόμενη μάθηση, επειδή οι κατηγορίες – κλάσεις καθορίζονται πριν ακόμη εξεταστούν τα δεδομένα. Η αναγνώριση προτύπου (*pattern recognition*) αποτελεί ένα είδος κατηγοριοποίησης, όπου ένα πρότυπο εισόδου κατηγοριοποιείται σε μία από διάφορες κατηγορίες, με βάση την εγγύτητά του ως προς αυτές τις προκαθορισμένες κατηγορίες.
- **Παλινδρόμηση (*Regression*): Χρησιμοποιείται για να απεικονιστεί ένα στοιχειώδες δεδομένο σε μία πραγματική μεταβλητή πρόβλεψης.** Περιλαμβάνει την εκμάθηση της συνάρτησης που κάνει αυτή την απεικόνιση. Η παλινδρόμηση προϋποθέτει ότι τα σχετικά δεδομένα ταιριάζουν με μερικά γνωστά είδη συνάρτησης και μετά καθορίζει την καλύτερη συνάρτηση αυτού του είδους που μοντελοποιεί τα δεδομένα που έχουν δοθεί. Ένα είδος ανάλυσης σφάλματος χρησιμοποιείται για να καθορίσει ποια συνάρτηση είναι η «καλύτερη».
- **Ανάλυση 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. Τρεις βασικές λειτουργίες πραγματοποιούνται στην ανάλυση χρονοσειρών: στη μία περίπτωση χρησιμοποιούνται μονάδες μέτρησης απόστασης για να καθορίσουν την ομοιότητα ανάμεσα σε διαφορετικές χρονοσειρές, στη δεύτερη εξετάζεται η δομή της χρονοσειράς για να καθορίσει (και ίσως να κατηγοριοποιήσει) την συμπεριφορά της και στην τρίτη χρησιμοποιούνται διαγράμματα χρονοσειρών για την πρόβλεψη μελλοντικών τιμών.

- **Πρόβλεψη (*Prediction*): Μπορεί να θεωρηθεί σαν ένα είδος κατηγοριοποίησης**. Αυτή η εργασία εξόρυξης γνώσης είναι διαφορετική από το μοντέλο πρόβλεψης, παρόλο που η διαδικασία πρόβλεψης αποτελεί έναν τύπο μοντέλου πρόβλεψης. Η διαφορά είναι ότι ως πρόβλεψη θεωρείται περισσότερο το να δίνεται τιμή σε μία μελλοντική κατάσταση, παρά σε μία τρέχουσα. Εδώ, γίνεται αναφορά σε μία είδος εφαρμογής, παρά σε μία προσέγγιση μοντελοποίησης. Οι εφαρμογές πρόβλεψης περιλαμβάνουν πρόγνωση πλημμύρων, αναγνώριση ομιλίας, μηχανική μάθηση και αναγνώριση προτύπου. Αν και μπορούν να προβλεφθούν οι μελλοντικές τιμές με τεχνικές ανάλυσης χρονοσειρών ή παλινδρόμησης, μπορούν να χρησιμοποιηθούν επίσης και άλλες προσεγγίσεις.

- **Περιγραφικό Μοντέλο (*Descriptive Model*)**: Αναγνωρίζει πρότυπα ή συσχετίσεις στα δεδομένα. Αντίθετα από το προβλεπτικό, λειτουργεί σαν ένα μέσο που διερευνά τις ιδιότητες των δεδομένων που εξετάζονται και όχι για να προβλέπει νέες ιδιότητες. Ένα παράδειγμα περιγραφικού μοντέλου είναι το εξής: Μία αλυσίδα πολυκαταστημάτων δημιουργεί ειδικούς καταλόγους, που στοχεύουν σε διάφορες δημογραφικές ομάδες, με βάση γνωρίσματα όπως το εισόδημα, ο τόπος διαμονής και τα φυσικά χαρακτηριστικά των δυνητικών πελατών (ηλικία, ύψος, βάρος κλπ). Προκειμένου να καθορίσει σε ποιους από τους πελάτες των διαφόρων καταλόγων θα σταλεί ταχυδρομικά διαφημιστικό υλικό και προκειμένου να δημιουργηθούν καινούργιοι και πιο συγκεκριμένοι κατάλογοι, η εταιρεία κάνει ομαδοποίηση των πιθανών πελατών βασιζόμενη στις προκαθορισμένες τιμές γνωρισμάτων. Τα αποτελέσματα της

συσταδοποίησης χρησιμοποιούνται στη συνέχεια από τη διεύθυνση προκειμένου να δημιουργηθούν ειδικοί κατάλογοι που θα διανεμηθούν στο πιο κατάλληλο τμήμα του πληθυσμού, βάσει της ομάδας που αντιστοιχεί σε αυτόν τον κατάλογο.

Οι πιο συνηθισμένες εργασίες εξόρυξης γνώσης από δεδομένα που χρησιμοποιούν το περιγραφικό μοντέλο, είναι η συσταδοποίηση, η παρουσίαση συνόψεων, οι κανόνες συσχετίσεων και η παρουσίαση ακολουθιών:

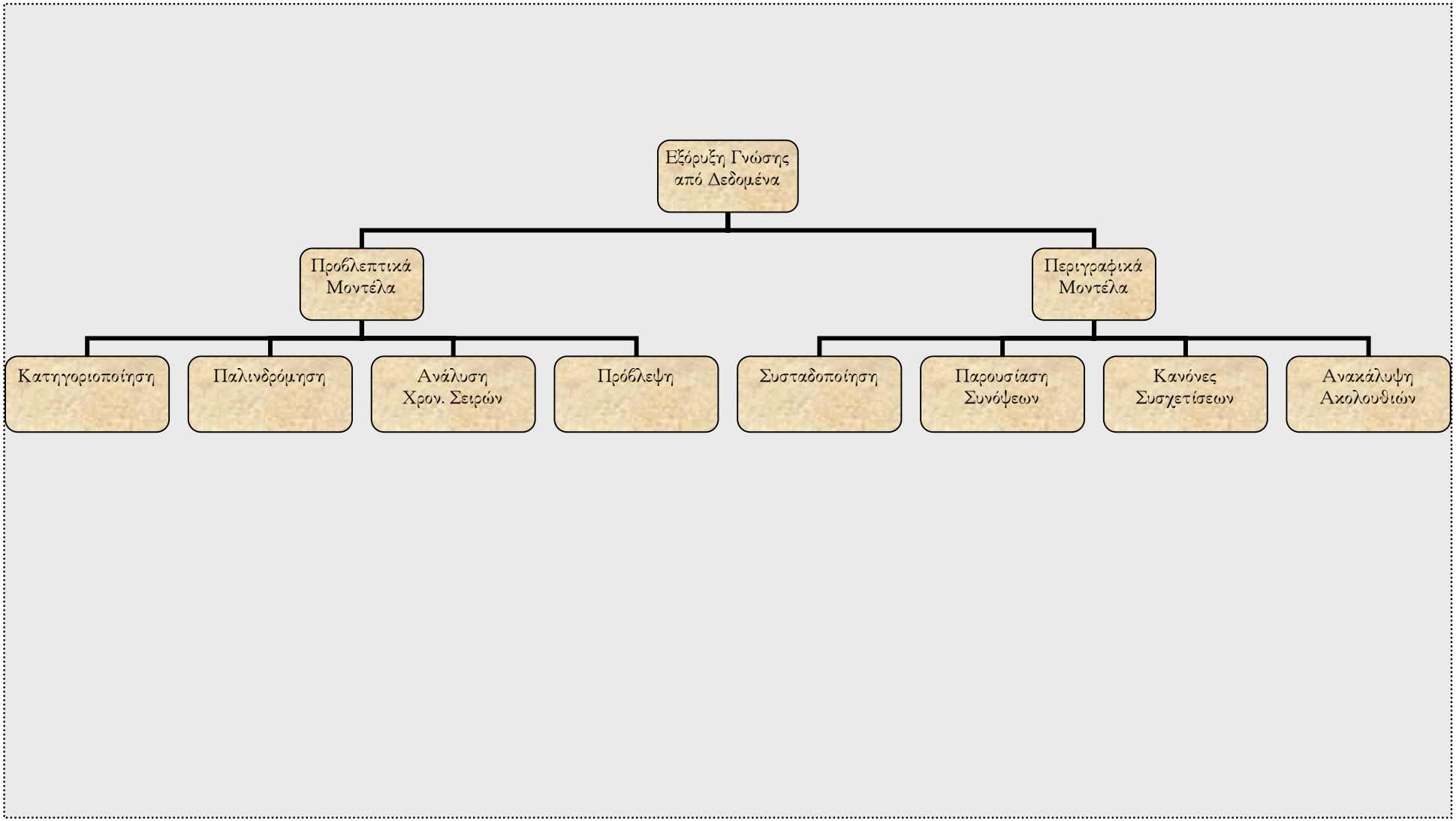
- **Συσταδοποίηση (*Clustering*)**: Είναι παρόμοια με την κατηγοριοποίηση, εκτός από το ότι οι συστάδες – ομάδες δεδομένων δεν είναι προκαθορισμένες, αλλά ορίζονται κυρίως από τα ίδια δεδομένα. Η συσταδοποίηση αναφέρεται εναλλακτικά και σαν μη εποπτευόμενη μάθηση, ή τμηματοποίηση. Μπορεί να θεωρηθεί σαν μια διαμέριση ή τμηματοποίηση των δεδομένων σε ομάδες, που μπορεί να είναι ή να μην είναι διακριτές μεταξύ τους. Συνήθως επιτυγχάνεται με τον καθορισμό της ομοιότητας, ως προς προκαθορισμένα γνωρίσματα, ανάμεσα στα δεδομένα. Τα πιο σχετικά δεδομένα κατατάσσονται στις ίδιες ομάδες. Εάν οι ομάδες δεν είναι προκαθορισμένες, χρειάζεται ένας ειδικός του πεδίου για να ερμηνεύσει τη σημασία των συστάδων που δημιουργούνται. Μια ειδική κατηγορία συσταδοποίησης ονομάζεται κατάτμηση (*segmentation*). Με την κατάτμηση, μια βάση δεδομένων χωρίζεται σε διακριτές ομάδες παρόμοιων εγγραφών που ονομάζονται τμήματα (*segments*). Η κατάτμηση συχνά θεωρείται πανομοιότυπη με την συσταδοποίηση. Κατά άλλους, η κατάτμηση θεωρείται σαν ειδικός τύπος συσταδοποίησης που εφαρμόζεται στην ίδια βάση δεδομένων.
- **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 κάποιου αριθμητικού γνωρίσματος). Εν ολίγοις, η παρουσίαση συνόψεων χαρακτηρίζει τα περιεχόμενα της βάσης δεδομένων.

- **Κανόνες Συσχετίσεων (*Association Rules*)**: 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. Ένας κανόνας συσχέτισης (*association rule*) είναι ένα μοντέλο που αναγνωρίζει ειδικούς τύπους συσχέτισης μεταξύ των δεδομένων. Η πιο διαδεδομένη προσέγγιση για την εύρεση κανόνων συσχετίσεων χρησιμοποιεί τα συχνά στοιχειοσύνολα (*frequent itemsets*), τα οποία ορίζονται ως τα στοιχειοσύνολα εκείνα των οποίων ο αριθμός των εμφανίσεων είναι πάνω από ένα κατώφλι *s*. Η προσέγγιση των συχνών στοιχειοσυνόλων i) εντοπίζει τα συχνά στοιχειοσύνολα βάσει του ορισμού τους και ii) δημιουργεί κανόνες από τα συχνά στοιχειοσύνολα. Οι συσχετίσεις συχνά χρησιμοποιούνται στις λιανικές πωλήσεις για να αναγνωρισθούν προϊόντα που συχνά αγοράζονται μαζί. Συσχετίσεις χρησιμοποιούνται επίσης σε πολλές άλλες εφαρμογές, όπως είναι η πρόβλεψη της αποτυχίας λειτουργίας των τηλεπικοινωνιακών διακοπτών. Η χρήση των κανόνων συσχετίσεων για τις όποιες αποφάσεις, πρέπει να γίνεται πολύ προσεκτικά επειδή υπάρχει ο κίνδυνος αυτές οι συσχετίσεις να είναι τυχαίες. Οι συσχετίσεις αυτές μπορεί να μην αντιπροσωπεύουν καμία έμφυτη σχέση ανάμεσα στα δεδομένα (κάτι που ισχύει για παράδειγμα στις συναρτησιακές εξαρτήσεις)
- **Ανακάλυψη Ακολουθιών**: Η ακολουθιακή ανάλυση (*sequential analysis*) ή αλλιώς ανακάλυψη ακολουθιών (*sequence discovery*) χρησιμοποιείται για να καθορισθούν σειριακά πρότυπα στα δεδομένα. Αυτά τα πρότυπα βασίζονται σε μία χρονική ακολουθία ενεργειών και είναι παρόμοια με τις συσχετίσεις στο ότι τα δεδομένα που εξάγονται

> συσχετίζονται, με τη διαφορά ότι η συσχέτισή τους αυτή βασίζεται στο χρόνο. [Dun04]

Οι διάφοροι τύποι μοντέλων και οι συνήθεις εργασίες εξόρυξης γνώσης από δεδομένα που χρησιμοποιούν αυτό το είδος μοντέλου φαίνονται παρακάτω (Σχήμα 1.2).



Σχήμα 1.2: Μοντέλα και Εργασίες στην Εξόρυξη Γνώσης από Δεδομένα
Πηγή: [Dun04]

1.4 Σημαντικοί Αλγόριθμοι στις Εργασίες Εξόρυξης Γνώσης από Δεδομένα

Στην ενότητα αυτή γίνεται μια σύντομη παρουσίαση βασικών αλγορίθμων που αφορούν στις εργασίες εξόρυξης γνώσης από δεδομένα, οι οποίες αναπτύχθηκαν προηγουμένως. Ειδικότερα, η μελέτη εστιάζεται σε τρεις από αυτές, την **κατηγοριοποίηση**, την **συσταδοποίηση** και τους **κανόνες συσχετίσεων**, που θα μας απασχολήσουν στην εξόρυξη γνώσης από χωρικά δεδομένα, γεγονός που καθιστά την αναφορά αυτή απαραίτητη.

Τονίζεται πως περαιτέρω ανάλυση σε αυτό το σημείο ξεφεύγει από τους σκοπούς της παρούσας μελέτης, καθώς και ότι δεν εξαντλούνται όλοι οι αλγόριθμοι οι οποίοι αναφέρονται στην βιβλιογραφία.

Λόγω του ότι η παρούσα εργασία εστιάζει στους αλγόριθμους εξόρυξης χωρικών δεδομένων, η εκτενής ανάλυση περιορίζεται σε ορισμένους εξ' αυτών, καθώς και σε κάποιους αλγόριθμους εξόρυξης γνώσης που βρίσκουν εφαρμογή τόσο σε χωρικά, όσο και σε μη χωρικά δεδομένα και οι οποίοι εξυπηρετούν τις ανάγκες αυτής της μελέτης. Σε επόμενη ενότητα θα γίνει πληρέστερη αναφορά σε αλγόριθμους που αφορούν στην εξόρυξη γνώσης από (χωρικά) δεδομένα μέσω της συσταδοποίησης. Σε αυτό το σημείο κάτι τέτοιο δεν κρίνεται σκόπιμο.

Αναφέρεται ότι ο αλγόριθμος $K-means$ (Κ πλησιέστεροι γείτονες) απαντάται τόσο στην κατηγοριοποίηση, όσο και στην συσταδοποίηση. Η διαφορά συσταδοποίησης και κατηγοριοποίησης έχει δωθεί στην ενότητα 1.3.1.

Οι αλγόριθμοι που εξετάζονται φαίνονται στον Πίνακα 1.1:

	ΕΡΓΑΣΙΑ ΕΞΟΡΥΞΗΣ ΓΝΩΣΗΣ	ΑΛΓΟΡΙΘΜΟΣ		
ΠΡΟΒΛΕΠΤΙΚΟ ΜΟΝΤΕΛΟ	ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ	**Αλγόριθμοι βασισμένοι στην στατιστική**		
		Αλγόριθμοι βασισμένοι στην απόσταση	*K* Πλησιέστεροι Γείτονες	
		Αλγόριθμοι βασισμένοι σε δένδρα απόφασης (*Decision Trees*)	*ID3*	
			C4.5 και *C5.0*	
			Δένδρα κατηγοριοποίησης και παλινδρόμησης (*CART – Classification and Regression Trees*)	
			Κλιμακούμενες τεχνικές για δένδρα απόφασης	Αλγόριθμος κλιμακούμενης παραλληλοποιήσιμης επαγωγής δένδρων αποφάσεων (*SPRINT – Scalable PaRallelizable Induction of Decision Trees*)
				Προσέγγιση *RainForest*
		Αλγόριθμοι βασισμένοι σε νευρωνικά δίκτυα	Διάδοση (*propagation*)	
			Εποπτευόμενη μάθηση του νευρωνικού δικτύου	
		Αλγόριθμοι βασισμένοι σε κανόνες	Δημιουργία κανόνων από ένα δένδρο απόφασης	
			Δημιουργία κανόνων από ένα νευρωνικό δίκτυο	
			Δημιουργία κανόνων χωρίς δένδρα απόφασης ή νευρωνικά δίκτυα	
		Συνδυαστικές τεχνικές	Σύνθεση από προσεγγίσεις	
			Πολλαπλές ανεξάρτητες τεχνικές - συνδυασμός πολλαπλών κατηγοριοποιητών (*combination of multiple classifiers*)	
ΠΕΡΙΓΡΑ	ΣΥΣΤΑΔΟΠΟΙΗΣΗ	**Ιεραρχικοί (*Hierarchical)***	Συσσωρευτικοί - *Agglomeratiuve*	Αλγόριθμος απλού συνδέσμου (*Single link algorithm*)
				Αλγόριθμος πλήρους συνδέσμου (*Complete link algorithm*)
				Αλγόριθμος μέσου συνδέσμου (*Average link algorithm*)
			Διαμεριστικοί - Divisive	Δένδρο ελάχιστης ζεύξης *(Minimum Spanning Tree - MST)*
				Αλγόριθμος συσταδοποίησης τετραγωνικού σφάλματος (*Squared error clustering algorithm*)
				Συσταδοποίηση K – κέντρα *(K – Means)*
				Αλγόριθμος πλησιέστερου γείτονα (*Nearest neighbor algorithm*)

ΑΦΙΚΟΜΟΝΤΕΛΟ				• Αλγόριθμος διαμερισμού γύρω από μέσους (*Partitioning Around Medoids - PAM*) • Συσταδοποίηση μεγάλων εφαρμογών *(Clustering LARge Applications – CLARA)*: βελτίωση πολυπλοκότητας χρόνου του *PAM* • Συσταδοποίηση μεγάλων εφαρμογών βασιζόμενη σε τυχαία αναζήτηση (*Clustering LARge Applications based upon randomized Search - CLARANS*): βελτίωση του αλγορίθμου *CLARA*
				Αλγόριθμος Ενέργειας Δεσμού (*Bond Energy Algorithm – BEA*)
				Συσταδοποίηση με γενετικούς αλγόριθμους
				Συσταδοποίηση με Νευρωνικά Δίκτυα
		Κατηγορικοί - *Categorical*	Εύρωστη συσταδοποίηση με χρήση συνδέσμων *(Robust Clustering using links – ROCK)*	
		Μεγάλες βάσεις δεδομένων – *Large DB*	*Balanced Iterative Reducing and clustering using Hierarchies- BIRCH*	
			Density Based Spatial Clustering of Applications with Noise- DBSCAN	
			Συσταδοποίηση με χρήση αντιπροσώπων (*Clustering Using Representatives - CURE)*	
	ΚΑΝΟΝΕΣ ΣΥΣΧΕΤΙΣΕΩΝ	**Βασικοί Αλγόριθμοι**	Αλγόριθμος *Apriori*	
			Αλγόριθμος της Δειγματοληψίας (*Sampling*)	
			Αλγόριθμος της Διαμέρισης (*Partitioning*)	
		Παράλληλοι και κατανεμημένοι αλγόριθμοι	Αλγόριθμος κατανομής μετρητών (*Count Distribution Algorithm - CDA*)	
			Αλγόριθμος κατανομής δεδομένων (*Data Distribution Algorithm - DDA*)	

Πίνακας 1.1: Σημαντικοί Αλγόριθμοι στις Εργασίες Εξόρυξης Γνώσης από Δεδομένα

Πηγή: [Dun04]

Ειδικότερα, σε ό,τι αφορά στην **κατηγοριοποίηση**, καμία τεχνική αυτής δεν υπερτερεί πάντα σε σχέση τις άλλες αναφορικά με την ακρίβεια της κατηγοριοποίησης. Ωστόσο, υπάρχουν πλεονεκτήματα και μειονεκτήματα σχετικά με τη χρήση της κάθε μιας. Οι προσεγγίσεις της παλινδρόμησης επιβάλλουν στα δεδομένα να ταιριάξουν σε ένα προκαθορισμένο μοντέλο. Εάν επιλεγεί ένα γραμμικό μοντέλο, τότε τα δεδομένα ταιριάζουν σε αυτό το μοντέλο, έστω και εάν στην πραγματικότητα μπορεί να μην είναι γραμμικά. Αυτό απαιτεί τη χρήση γραμμικών δεδομένων. Η τεχνική K πλησιέστερου γείτονα απαιτεί μόνο ότι τα δεδομένα θα είναι τέτοια ώστε να μπορούν να υπολογιστούν Μ αποστάσεις. Αυτό μπορεί στη συνέχεια να εφαρμοστεί ακόμα και σε μη αριθμητικά δεδομένα. Ο χειρισμός των ακραίων σημείων γίνεται με το να κοιτάζει η μέθοδος μόνο τους Κ πλησιέστερους γείτονες. Οι τεχνικές που βασίζονται σε δένδρα αποφάσεων είναι εύκολες στην κατανόηση, αλλά μπορεί να οδηγήσουν σε υπερπροσαρμογή. Προκειμένου να αποφευχθεί κάτι τέτοιο, μπορούν να χρησιμοποιηθούν τεχνικές κλαδέματος. Ο $ID3$ μπορεί να. εφαρμοσθεί μόνο σε κατηγορικά δεδομένα. Βελτιστοποιήσεις σε. αυτό το σημείο έχουν εμφανιστεί με τον $C4.5$ και τον $C5$, οι οποίοι επιτρέπουν τη χρήση συνεχών δεδομένων, όπως επίσης και βελτιωμένες τεχνικές διάσπασης. Ο $CART$ δημιουργεί μερικά δυαδικά δένδρα και έτσι μπορεί να οδηγήσει στη δημιουργία δένδρων με μεγάλο βάθος. (www. cis.temple.edu ˜ ingargio/cis587/readings/id3-c45.html)

Η εξέταση των παραπάνω προσεγγίσεων με βάση την ανάλυση πολυπλοκότητας οδηγεί στο συμπέρασμα ότι είναι πολύ αποτελεσματικές. Αυτό οφείλεται στο γεγονός ότι με το που θα χτισθεί το μοντέλο, η εφαρμογή του στην κατηγοριοποίηση είναι σχετικά απλή. Οι στατιστικές τεχνικές, όπως η παλινδρόμηση απαιτούν σταθερό χρόνο για να κατηγοριοποιήσουν μία πλειάδα με το που τα μοντέλα έχουν χτιστεί. Οι προσεγγίσεις που βασίζονται στην απόσταση χρειάζονται επίσης σταθερό χρόνο αλλά απαιτούν κάθε πλειάδα να συγκριθεί είτε με έναν αντιπρόσωπο κάθε κατηγορίας είτε με όλα τα στοιχεία στο σύνολο εκπαίδευσης. Υποθέτοντας ότι υπάρχουν q από αυτά, η προσέγγιση K – πλησιέστερου γείτονα απαιτεί $O(q)$ χρόνο για κάθε πλειάδα. Οι τεχνικές

κατηγοριοποίησης με δένδρα απόφασης, απαιτούν έναν αριθμό από συγκρίσεις ο οποίος είναι (στη χειρότερη περίπτωση) ίσος με το πιο μακρύ μονοπάτι από τη ρίζα προς ένα φύλλο. Έτσι, απαιτούν $O(log(q))$ χρόνο ανά πλειάδα. Οι προσεγγίσεις τύπου νευρωνικού δικτύου και πάλι απαιτούν μια πλειάδα να διαδοθεί μέσω του γράφου. Επειδή το μέγεθος του γράφου είναι σταθερό, μπορεί να θεωρηθεί ότι η εκτέλεση γίνεται σε σταθερό χρόνο. Έτσι, όλοι οι αλγόριθμοι απαιτούν $O(n)$ χρόνο για να κατηγοριοποιήσουν n στοιχεία μίας βάσης δεδομένων.

Αναφορικά με τους αλγορίθμους **συσταδοποίησης**, οι τεχνικές απλού συνδέσμου, πλήρους συνδέσμου και μέσου συνδέσμου είναι ιεραρχικές τεχνικές με πολυπλοκότητα $O(n^2)$. Οι παραπάνω τεχνικές είναι συσσωρευτικές. αλλά είναι παράλληλα και διαμεριστικές και δημιουργούν τις συστάδες με φορά από πάνω προς τα κάτω (*top - down*). Επίσης, οι τεχνικές αυτές υποθέτουν ότι υπάρχουν όλα τα δεδομένα ταυτόχρονα και συνεπώς δεν είναι αυξητικές

Τόσο ο *k – means*, όσο και οι τεχνικές τετραγωνικού σφάλματος είναι επαναληπτικές τεχνικές και απαιτούν $O(tkn)$ χρόνο. Η τεχνική πλησιέστερου γείτονα δεν είναι επαναληπτική, αλλά στην περίπτωση αυτή ο αριθμός των συστάδων δεν είναι προκαθορισμένος. Συνεπώς, η πολυπλοκότητα χειρότερης περίπτωσης μπορεί να είναι $O(n^2)$. Ο αλγόριθμος *BIRCH* φαίνεται να είναι αρκετά αποδοτικός. Ο αλγόριθμος *CURE* αποτελεί βελτίωση των ανωτέρω, επιτυγχάνει καλύτερη κλιμάκωση μέσω δειγματοληψίας και διαμερισμού και αναπαριστά μια συστάδα με πολλαπλά σημεία αντί ενός. Η χρήση πολλαπλών σημείων επιτρέπει στη συγκεκριμένη προσέγγιση να εντοπίζει μη σφαιρικές συστάδες. Με τη δειγματοληψία, ο *CURE* επιτυγχάνει πολυπλοκότητα χρόνου $O(n)$. Ωστόσο, ο *CURE* δεν χειρίζεται αποδοτικά τα κατηγορικά δεδομένα. Αυτό βέβαια του επιτρέπει να είναι πιο ανθεκτικός στις αρνητικές συνέπειες των ακραίων σημείων. Οι αλγόριθμοι *k - means* και *PAM* στηρίζονται στην επαναπροσδιορισμό των στοιχείων στις συστάδες, ο οποίος δεν οδηγεί πάντα στον εντοπισμό μιας καθολικά βέλτιστης ανάθεσης. Τα αποτελέσματα του *k - means* είναι αρκετά ευαίσθητα στην ύπαρξη ακραίων σημείων. Ο αλγόριθμος *BIRCH* είναι ταυτόχρονα δυναμικός και κλιμακούμενος. Ωστόσο, εντοπίζει μόνο σφαιρικές

συστάδες. Ο αλγόριθμος *DBSCAN* στηρίζεται στην πυκνότητα. Η πολυπλοκότητα χρόνου του *DBSCAN* μπορεί να βελτιωθεί στο *O(nlog(n))* με κατάλληλα χωρικά ευρετήρια. Η απόδοση των γενετικών αλγορίθμων εξαρτάται εξ' ολοκλήρου από την τεχνική που επιλέγεται για την αναπαράσταση των επιμέρους στοιχείων, από το πώς γίνεται η διασταύρωση, και από το κριτήριο τερματισμού που χρησιμοποιείται.

Οι αλγόριθμοι **κανόνων συσχετίσεων** μπορούν να ταξινομηθούν ως προς τις ακόλουθες διαστάσεις:

- **Στόχος**: είτε να δημιουργούν όλους τους κανόνες που ικανοποιούν μια δεδομένη τιμή για την υποστήριξη και το επίπεδο εμπιστοσύνης, είτε να δημιουργούν κάποιο υποσύνολο των κανόνων με βάση τους δεδομένους περιορισμούς.
- **Τύπος**: αφορά στους κανόνες συσχετίσεων, οι οποίοι μπορεί να είναι κανονικοί ή πιο εξελιγμένοι.
- **Τύπος δεδομένων**: κατηγορικά χωρικά (ή μη χωρικά) δεδομένα. Κανόνες μπορούν επίσης να παραχθούν για άλλους τύπους δεδομένων, όπως το απλό κείμενο (εξόρυξη γνώσης από τον παγκόσμιο ιστό).
- **Τεχνική**: η πιο κοινή στρατηγική για τη δημιουργία κανόνων συσχέτισης είναι αυτή της εύρεσης των συχνών στοιχειοσυνόλων.
- **Στρατηγική στοιχειοσυνόλων**: τα στοιχειοσύνολα μπορούν να μετρηθούν με διάφορους τρόπους. Η πιο απλοϊκή προσέγγιση είναι η δημιουργία όλων των στοιχειοσυνόλων και το μέτρημα αυτών. Καθώς αυτό είναι συνήθως πολύ απαιτητικό σε χώρο, η πιο κοινή προσέγγιση είναι η από κάτω προς τα πάνω (*bottom - up*) προσέγγιση που χρησιμοποιείται από τον *Apriori* αλγόριθμο, η οποία εκμεταλλεύεται την ιδιότητα των συχνών στοιχειοσυνόλων. Εναλλακτικά, θα μπορούσε να χρησιμοποιηθεί μια από πάνω προς τα κάτω τεχνική.
- **Στρατηγική συναλλαγών**: προκειμένου να μετρηθούν τα στοιχειοσύνολα, πρέπει να γίνει ένα πέρασμα των συναλλαγών της βάσης δεδομένων. Θα

μπορούσαν να μετρηθούν όλες οι συναλλαγές, ή μόνο ένα δείγμα, ή οι συναλλαγές θα μπορούσαν να διαιρεθούν σε διαμερίσεις.

- **Δομή δεδομένων στοιχειοσυνόλων**: Η πιο κοινή δομή δεδομένων που χρησιμοποιείται για την αποθήκευση των υποψήφιων στοιχειοσυνόλων, όπως επίσης και των μετρητών τους, είναι ένα δένδρο κατακερματισμού. Ένα δένδρο κατακερματισμού (*hash tree*) είναι ένα δέντρο αναζήτησης πολλαπλών δρόμων, όπου η διακλάδωση που θα ακολουθηθεί σε κάθε επίπεδο του δένδρου καθορίζεται με την εφαρμογή μίας συνάρτησης κατακερματισμού, σε αντίθεση με την σύγκριση των τιμών των κλειδιών με τα σημεία διάσπασης του κόμβου. Ένας κόμβος - φύλλο στο δένδρο κατακερματισμού περιέχει τους υποψηφίους που κατακερματίζονται σε αυτό, αποθηκευμένους με κάποια διάταξη. Κάθε εσωτερικός κόμβος στην πραγματικότητα περιέχει έναν πίνακα κατακερματισμού με συνδέσμους προς τους κόμβους κλειδιά. Τα δένδρα κατακερματισμού παρέχουν μία αποτελεσματική τεχνική για την αποθήκευση, προσπέλαση και μέτρημα των στοιχειοσυνόλων. Είναι αποτελεσματικά στην αναζήτηση, εισαγωγή και διαγραφή στοιχειοσυνόλων.
- **Βελτιστοποίηση**: Αυτές οι τεχνικές αποσκοπούν στο να βελτιώσουν την απόδοση ενός αλγορίθμου, δεδομένης της κατανομής των δεδομένων (ανομοιομορφία) ή της ποσότητας της κύριας μνήμης.
- **Αρχιτεκτονική**: Έχουν προστεθεί σειριακοί, παράλληλοι όσο και κατανεμημένοι αλγόριθμοι.
- **Στρατηγική παραλληλισμού**: Έχουν χρησιμοποιηθεί τόσο ο παραλληλισμός των δεδομένων, όσο και ο παραλληλισμός των εργασιών.

Πιο κάτω (Πίνακας 3.1) γίνεται μία σύγκριση των αλγορίθμων εύρεσης κανόνων συσχέτισης οι οποίοι αναφέρθηκαν προηγούμενα (m είναι ο αριθμός των στοιχείων).

Διαμέριση	Περάσματα	Δομή δεδομένων	Παραλληλισμός
Apriori	m+1	Δένδρο κατακερματισμού	όχι
Sampling	2	Δεν προδιαγράφεται	όχι
Partitioning	2	πίνακας κατακερματισμού	όχι
CDA	m+1	δένδρο κατακερματισμού	δεδομένων
DDA	m+1	δένδρο κατακερματισμού	εργασιών

Πίνακας 1.2: Σύγκριση των Αλγορίθμων Εύρεσης Κανόνων Συσχέτισης
Πηγή: [DXGH00]

1.5 Θέματα Σχετικά με την Εξόρυξη Γνώσης από Δεδομένα

Υπάρχουν πολλά σημαντικά θέματα υλοποίησης που σχετίζονται με την εξόρυξη γνώσης από δεδομένα. Ένα από αυτά είναι η 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, οι οποίοι είναι απαραίτητοι για να ταυτοποιήσουν τα δεδομένα εκπαίδευσης και να ορίσουν τα επιθυμητά αποτελέσματα, ενώ σχετίζονται και με το θέμα της ερμηνείας των αποτελεσμάτων. Σε αντίθετη περίπτωση θα ήταν χωρίς νόημα για το μέσο χρήστη.
- των ειδικών της συγκεκριμένης τεχνικής εξόρυξης γνώσης, οι οποίοι χρησιμοποιούνται προκειμένου να μορφοποιήσουν τις ερωτήσεις και να βοηθήσουν στην ερμηνεία των αποτελεσμάτων. Η οπτικοποίηση των αποτελεσμάτων των αλγορίθμων εξόρυξης γνώσης είναι χρήσιμη για να δούμε και να κατανοήσουμε ευκολότερα τα αποτελέσματα αυτά.

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. Όταν το μοντέλο δεν ταιριάζει σε μελλοντικές καταστάσεις, το φαινόμενο καλείται υπερπροσαρμογή (*overfitting*). Αυτό μπορεί να συμβαίνει εξαιτίας υποθέσεων που γίνονται για τα δεδομένα ή απλά μπορεί να συμβαίνει εξαιτίας του μικρού μεγέθους των δεδομένων εκπαίδευσης. Η υπερπροσαρμογή μπορεί επίσης να εμφανιστεί και σε περιπτώσεις όπου δεν αλλάζουν τα δεδομένα.

Αναφορικά με τις ακραίες τιμές (*outliers*), υπάρχουν συχνά πολλές καταχωρήσεις δεδομένων που δεν ταιριάζουν σωστά στο μοντέλο που έχει αναπτυχθεί. Αυτό συμβαίνει συχνά στις πολύ μεγάλες βάσεις δεδομένων, Εάν το μοντέλο που θα δημιουργηθεί περιλαμβάνει αυτές τις ακραίες τιμές, τότε ίσως να μη συμπεριφέρεται σωστά για τα μη ακραία δεδομένα.

Ένα άλλο θέμα είναι τα μεγάλα σύνολα δεδομένων, τα οποία δημιουργούν προβλήματα όταν εφαρμόζονται αλγόριθμοι εξόρυξης γνώσης που έχουν σχεδιαστεί για μικρά σύνολα δεδομένων. Πολλές εφαρμογές μοντελοποίησης έχουν εκθετική πολυπλοκότητα και γι' αυτό το λόγο είναι αναποτελεσματικές στα μεγαλύτερα σύνολα δεδομένων. Αποτελεσματικό εργαλείο για να αντιμετωπισθεί το πρόβλημα της κλιμάκωσης είναι η δειγματοληψία.

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 μίας εργασίας. Η χρήση άλλων γνωρισμάτων μπορεί απλά να αυξήσει τη συνολική πολυπλοκότητα και να μειώσει την απόδοση ενός αλγορίθμου. Μία λύση στο πρόβλημα των πολλαπλών διαστάσεων είναι να μειωθούν τα γνωρίσματα, κάτι που αναφέρεται ως «μείωση των πολλαπλών διαστάσεων» (*dimensionality reduction*).

Κατά τη διάρκεια της φάσης της προεπεξεργασίας στη διαδικασία εξόρυξης γνώσης από δεδομένα, τα δεδομένα που λείπουν μπορούν να συμπληρωθούν με κατ' εκτίμηση 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. Επιπλέον, μερικά γνωρίσματα στη βάση δεδομένων ίσως να μην έχουν ενδιαφέρον όσον αφορά στη συγκεκριμένη εργασία εξόρυξης γνώσης που πραγματοποιείται. Ακόμη, μερικές τιμές των γνωρισμάτων μπορεί να είναι άκυρες ή λανθασμένες (θόρυβος). Αυτές οι τιμές συνήθως διορθώνονται πριν τρέξουμε την εφαρμογή της εξόρυξης γνώσης από δεδομένα. Επιπρόσθετα, οι βάσεις δεδομένων δεν μπορούν να θεωρηθούν στατικές. Όμως, οι περισσότεροι αλγόριθμοι εξόρυξης γνώσης υποθέτουν ότι κάτι τέτοιο συμβαίνει. Αυτό απαιτεί ο αλγόριθμος να ξανατρέχει από την αρχή κάθε φορά που αλλάζει η βάση δεδομένων (δεδομένα που αλλάζουν).

Ένα από τα ζητούμενα, τέλος, είναι να προσδιοριστεί η ενδεικνυόμενη χρήση για μια πληροφορία που προήλθε από τη λειτουργία της εξόρυξης γνώσης. Η αποτελεσματική ερμηνεία των αποτελεσμάτων θεωρείται μερικές φορές πιο δύσκολο έργο από το τρέξιμο ενός αλγορίθμου. [Dun04]

Στα επόμενα παρουσιάζονται με συντομία μια σειρά από έννοιες σχετικές με την εξόρυξη γνώσης και καταδεικνύεται πώς αυτές σχετίζονται με την τελευταία.

1.5.1 Βάσεις Δεδομένων

Οι αλγόριθμοι που δεν αποδίδουν καλά όταν υπάρχει κλιμάκωση των δεδομένων, όπως πράγματι συμβαίνει στις πραγματικές ογκώδεις βάσεις δεδομένων, είναι περιορισμένης χρήσης. Με αυτό συσχετίζεται το γεγονός ότι οι τεχνικές πρέπει να λειτουργούν ανεξάρτητα από το μέγεθος της διαθέσιμης κύριας μνήμης.

Τα πραγματικά δεδομένα έχουν θόρυβο και πολλές ελλιπείς τιμές γνωρισμάτων. Οι αλγόριθμοι θα πρέπει να μπορούν να δουλεύουν ακόμα και παρουσία αυτών των προβλημάτων, ενώ πολλοί αλγόριθμοι εξόρυξης γνώσης από δεδομένα δουλεύουν με στατικές βάσεις δεδομένων. Αυτό δεν μπορεί να θεωρηθεί ρεαλιστική υπόθεση, ενώ ορισμένοι αλγόριθμοι μπορεί μεν να δουλεύουν καλά αλλά να είναι δυσνόητοι και δύσχρηστοι, άρα μη αποδεκτοί από τους χρήστες. [Dun04]

1.5.2 Στατιστική

Η όλη διαδικασία της εξόρυξης γνώσης ομοιάζει σε σημαντικό βαθμό με την στατιστική. Επ' αυτού, μπορεί να πει κανείς ότι ένας τρόπος να ιδωθεί η εξόρυξη γνώσης είναι ως διαδικασία φιλτραρίσματος, πριν την εφαρμογή αυστηρών στατιστικών εργαλείων. Ο ρόλος αυτής της διαδικασίας είναι να «οργώσει» τα δεδομένα και να παράξει ενδιαφέρουσες υποθέσεις, οι οποίες εν συνεχεία μπορούν αν επιβεβαιωθούν κάνοντας χρήση της στατιστικής. Κάτι τέτοιο είναι παρόμοιο με την χρήση των R – δένδρων, προκειμένου να ανακτηθούν τα Ελάχιστα Περιβάλλονται Ορθογώνια (*Minimum Bounding Rectangle - MBR*), ώστε να απαντηθεί μια σειρά από ερωτήματα. Τα R – δένδρα και τα *MBR* παρέχουν ένα γρήγορο φίλτρο, ώστε να

αναζητηθούν στο χώρο υποψήφιοι οι οποίοι δύνανται να ικανοποιούν ένα ερώτημα. [SC03]

Απλές έννοιες στατιστικής, όπως ο καθορισμός μιας κατανομής δεδομένων ή ο υπολογισμός της μέσης τιμής και της απόκλισης, μπορούν να θεωρηθούν ως τεχνικές εξόρυξης γνώσης. Κάθε μία από αυτές συνιστά ένα περιγραφικό μοντέλο για τα υπό θεώρηση δεδομένα

Η έρευνα στη στατιστική έχει συνεισφέρει με πολλούς από τους αλγόριθμους που έχουν προταθεί για εξόρυξη γνώσης. Η διαφορά βρίσκεται στους σκοπούς, στο γεγονός ότι οι στατιστικοί χειρίζονται μικρότερα και πιο σχηματοποιημένα σύνολα δεδομένων, καθώς και στην έμφαση της εξόρυξης γνώσης στην χρήση τεχνικών μηχανικής μάθησης

Κατά πολλούς η κύρια διαφορά ανάμεσα στην εξόρυξη γνώσης και την στατιστική είναι ότι η εξόρυξη γνώσης προορίζεται για χρήση από ένα στέλεχος της επιχείρησης και όχι από ένα στατιστικό. Έτσι, η εξόρυξη γνώσης (ειδικά από τη σκοπιά των βάσεων δεδομένων) περιλαμβάνει όχι μόνο την μοντελοποίηση, αλλά επίσης την ανάπτυξη αποδοτικών αλγορίθμων (και δομών δεδομένων) για την εκτέλεση της μοντελοποίησης σε μεγάλα σύνολα δεδομένων. [Dun04]

1.5.3 Ασαφή Σύνολα και Ασαφής Λογική

Ένα σύνολο φυσιολογικά ορίζεται ως μια συλλογή αντικειμένων. Ένα ασαφές σύνολο είναι ένα σύνολο *F*, του οποίου η συνάρτηση συμμετοχής σε αυτό *f*, είναι μία συνάρτηση με πεδίο τιμών τους πραγματικούς αριθμούς του διαστήματος [0,1].

Τα ασαφή σύνολα χρησιμοποιούνται σε πολλούς τομείς της πληροφορικής και των βάσεων δεδομένων. Στο πρόβλημα της κατηγοριοποίησης, σε όλες τις εγγραφές της βάσης δεδομένων ανατίθεται μία από τις προκαθορισμένες κατηγορίες. Μια κοινή προσέγγιση για την λύση του προβλήματος της κατηγοριοποίησης είναι να ανατεθεί μία συνάρτηση συμμετοχής σε κάθε εγγραφή, για κάθε κατηγορία. Στην

εγγραφή τότε εκχωρείται η κατηγορία με την υψηλότερη τιμή συνάρτησης συμμετοχής. Παρομοίως, τα ασαφή σύνολα μπορούν να χρησιμοποιηθούν για να περιγράψουν άλλες λειτουργίες εξόρυξης γνώσης. Οι κανόνες συσχετίσεων παράγονται δίνοντας μια τιμή εμπιστοσύνης, η οποία υποδεικνύει τον βαθμό ισχύος του κανόνα σε ολόκληρη την βάση δεδομένων. Αυτό μπορεί να θεωρηθεί ως μία συνάρτηση συμμετοχής.

1.5.4 Ανάκτηση Πληροφοριών

Η Ανάκτηση Πληροφοριών και πιο πρόσφατα οι ψηφιακές βιβλιοθήκες και η αναζήτηση στο διαδίκτυο συνεπάγονται την ανάκτηση επιθυμητών πληροφοριών από δεδομένα κειμένου. Η ιστορική εξέλιξη της ανάκτησης πληροφοριών βασίστηκε στην αποτελεσματική χρήση των βιβλιοθηκών και εμπίπτει στις εργασίες κατηγοριοποίησης.

1.5.5 Συστήματα Υποστήριξης Αποφάσεων

Τα Συστήματα Υποστήριξης Αποφάσεων (*Decision Support Systems*) συνιστούν ευφυή υπολογιστικά συστήματα και εργαλεία που βοηθούν τους διευθυντές στην λήψη αποφάσεων και στην επίλυση προβλημάτων. Διαφέρουν από τα παραδοσιακά συστήματα διαχείρισης βάσεων δεδομένων, στο ότι οι περισσότερες ερωτήσεις αναφέρονται σε ειδικό σκοπό και προσαρμοσμένες πληροφορίες. Η εξόρυξη γνώσης μπορεί να θεωρηθεί ως μία συλλογή εργαλείων που βοηθούν την συνολική διεργασία που επιτελούν τα συστήματα υποστήριξης αποφάσεων, δηλαδή τα συστήματα αυτά μπορούν να χρησιμοποιήσουν εργαλεία εξόρυξης γνώσης.

1.5.6 Μοντελοποίηση σε Διαστάσεις

Η μοντελοποίηση σε διαστάσεις (*dimensional modeling*) είναι ένας διαφορετικός τρόπος να ιδωθούν δεδομένα και να διατυπωθούν ερωτήματα σε μία βάση δεδομένων. Μπορεί να χρησιμοποιηθεί σε ένα σύστημα υποστήριξης αποφάσεων σε συνδυασμό με εργασίες εξόρυξης γνώσης. Αν και δεν κρίνεται απαραίτητο, για σκοπούς αποδοτικότητας τα δεδομένα μπορούν να αποθηκευθούν χρησιμοποιώντας διαφορετικές δομές δεδομένων, ενώ οι εφαρμογές υποστήριξης

αποφάσεων συχνά απαιτούν οι πληροφορίες να ανακτώνται μέσω πολλών διαστάσεων.

1.5.7 Άμεση Αναλυτική Επεξεργασία

Τα συστήματα Άμεσης Αναλυτικής Επεξεργασίας (*Online Analytical Processing - OLAP*) στοχεύουν στο να παρέχουν πιο πολύπλοκα αποτελέσματα ερωτήσεων από τα παραδοσιακά συστήματα βάσεων δεδομένων. Αντίθετα με τις ερωτήσεις βάσεων δεδομένων, οι εφαρμογές άμεσης αναλυτικής επεξεργασίας συνήθως περιλαμβάνουν ανάλυση των πραγματικών δεδομένων. Μπορούν να θεωρηθούν ως επεκτάσεις μερικών βασικών συναθροιστικών λειτουργιών, διαθέσιμων στην δομημένη γλώσσα ερωτημάτων (*SQL*). Αυτή η περαιτέρω ανάλυση των δεδομένων και η σαφής φύση των ερωτήσεων άμεσης αναλυτικής επεξεργασίας είναι αυτό που διαφοροποιεί τις εφαρμογές αυτού του τύπου από τις παραδοσιακές. Τα εργαλεία της άμεσης αναλυτικής επεξεργασίας μπορούν επίσης να χρησιμοποιηθούν σε συστήματα υποστήριξης αποφάσεων.

1.5.8 Μηχανές αναζήτησης στο Διαδίκτυο

Ως αποτέλεσμα του τεράστιου όγκου δεδομένων που βρίσκεται στο διαδίκτυο και του γεγονότος ότι αυτός συνεχώς μεγαλώνει, προκαλεί ενδιαφέρον η συλλογή επιθυμητών πληροφοριών. Οι μηχανές αναζήτησης χρησιμοποιούνται για να προσπελαύνουν τα δεδομένα και μπορεί να τις φανταστεί κανείς σαν συστήματα ερωτήσεων όπως είναι τα συστήματα ανάκτησης πληροφοριών. Όπως στα τελευταία, οι ερωτήσεις των μηχανών αναζήτησης μπορούν να δηλωθούν με λέξεις – κλειδιά, με δυαδική λογική, κλπ. Η διαφορά εντοπίζεται κατά βάση στα δεδομένα πάνω στα οποία γίνεται η αναζήτηση (σελίδες με ετερογενή δεδομένα και πολλούς υπερσυνδέσμους), καθώς και στην αρχιτεκτονική που χρησιμοποιείται.

1.5.9 Μηχανική Μάθηση

Η Τεχνητή Νοημοσύνη περιλαμβάνει πολλές τεχνικές εξόρυξης γνώσης, όπως τα νευρωνικά δίκτυα και η κατηγοριοποίηση. Ωστόσο, η τεχνητή νοημοσύνη είναι πιο γενική και περιλαμβάνει περιοχές εκτός της παραδοσιακής εξόρυξης γνώσης.

Οι εφαρμογές της μπορεί επίσης να μην απασχολούνται με το θέμα αντιμετώπισης μεγάλων όγκων δεδομένων, καθώς χειρίζονται συνήθως μικρά σύνολα δεδομένων.

Μηχανική Μάθηση είναι η περιοχή της τεχνητής νοημοσύνης, η οποία εξετάζει πώς γράφονται προγράμματα που να μαθαίνουν. Για τους σκοπούς της εξόρυξης γνώσης, η μηχανική μάθηση χρησιμοποιείται συχνά για πρόβλεψη και κατηγοριοποίηση. Με την μηχανική μάθηση ο υπολογιστής κάνει μία πρόβλεψη και κατόπιν, βασισμένος στην ανάδραση περί της ορθότητας της πρόβλεψης, «μαθαίνει» από την ανάδραση αυτή.

Όταν η μηχανική μάθηση εφαρμόζεται σε εργασίες εξόρυξης γνώσης, χρησιμοποιείται ένα μοντέλο για να αναπαραστήσει τα δεδομένα. Κατά τη διάρκεια της διαδικασίας μάθησης χρησιμοποιείται ένα δείγμα από την βάση δεδομένων, ώστε να εκπαιδεύσει το σύστημα να εκτελέσει σωστά την επιθυμητή εργασία. Κατόπιν, το σύστημα εφαρμόζεται στη γενική βάση δεδομένων για να εκτελέσει στην πραγματικότητα την εργασία. Αυτή η προσέγγιση μοντελοποίησης διαιρείται σε δύο φάσεις. Κατά την φάση εκπαίδευσης χρησιμοποιούνται ιστορικά δεδομένα ή δεδομένα δειγματοληψίας για να δημιουργήσουν ένα μοντέλο που να αναπαριστά τα δεδομένα αυτά. Υποτίθεται ότι το μοντέλο είναι αντιπροσωπευτικό όχι μόνο για τα δεδομένα δειγματοληψίας, αλλά και για την βάση δεδομένων ως σύνολο, ακόμη και για τα μελλοντικά δεδομένα. Κατόπιν, η φάση ελέγχου εφαρμόζει αυτό το μοντέλο στα υπόλοιπα και στα μελλοντικά δεδομένα.

1.5.10 Ταίριασμα Προτύπων

Το ταίριασμα προτύπου (*pattern matching*) ή αναγνώριση προτύπου (*pattern recognition*) βρίσκει εμφανίσεις ενός προκαθορισμένου προτύπου στα δεδομένα και χρησιμοποιείται σε πολλές διαφορετικές εφαρμογές: ένας κειμενογράφος χρησιμοποιεί το ταίριασμα προτύπων για να βρει τις εμφανίσεις μιας γραμματοσειράς στο κείμενο που συντάσσεται, η ανάκτηση πληροφοριών και οι μηχανές αναζήτησης στο διαδίκτυο χρησιμοποιούν το ταίριασμα προτύπων για να

βρίσκουν έγγραφα που περιέχουν ένα προκαθορισμένο πρότυπο (πχ μία λέξη - κλειδί). Η ανάλυση χρονολογικών σειρών εξετάζει τα πρότυπα συμπεριφοράς σε δεδομένα που αποκομίζονται από δύο διαφορετικές χρονολογικές σειρές για να βρει την ομοιότητά τους. Δηλαδή, το ταίριασμα προτύπων μπορεί να θεωρηθεί ως ένας τύπος κατηγοριοποίησης, όπου τα προκαθορισμένα πρότυπα είναι οι υπό εξέταση κατηγορίες. Τα δεδομένα τότε ανατίθενται στην σωστή κατηγορία με βάση την ομοιότητα μεταξύ των δεδομένων και των κατηγοριών. [Dun04]

Κεφάλαιο 2
Εξόρυξη Χωρικών Δεδομένων

2.1 Εισαγωγή – Ορισμοί

Η Εξόρυξη Χωρικής Γνώσης (*Spatial Mining*) είναι εξόρυξη γνώσης που εφαρμόζεται σε βάσεις χωρικών δεδομένων ή χωρικά δεδομένα. Ορισμένες από τις εφαρμογές εξόρυξης χωρικής γνώσης εντάσσονται στα πεδία των γεωγραφικών συστημάτων πληροφοριών, γεωλογίας, περιβαλλοντικής επιστήμης, διαχείρισης πόρων, γεωργίας, ιατρικής και ρομποτικής.

Τα χωρικά δεδομένα είναι δεδομένα, τα οποία έχουν μια χωρική συνιστώσα (ή συνιστώσα θέσης). Μπορούν να θεωρηθούν ως δεδομένα αντικειμένων τα οποία βρίσκονται σε έναν φυσικό χώρο. Αυτό μπορεί να δηλώνεται ρητά με ένα ή περισσότερα γνωρίσματα θέσης, όπως η διεύθυνση ή το γεωγραφικό πλάτος / μήκος ή μπορεί να υπονοείται, όπως με μια διαμέριση της βάσης δεδομένων η οποία βασίζεται στη θέση. Επιπλέον, τα χωρικά δεδομένα μπορούν να προσπελασθούν χρησιμοποιώντας ερωτήσεις που περιέχουν χωρικούς τελεστές όπως οι τελεστές «κοντά», «βόρεια», «νότια», «γειτονικά» και «περιέχεται σε». Τα χωρικά δεδομένα αποθηκεύονται σε βάσεις χωρικών δεδομένων που περιέχουν

τόσο τη χωρική όσο και τη μη χωρική πληροφορία. Εξαιτίας της ενυπάρχουσας πληροφορίας της απόστασης που σχετίζεται με τα χωρικά δεδομένα, οι βάσεις χωρικών δεδομένων πολύ συχνά χρησιμοποιούν ειδικές δομές δεδομένων ή ευρετήρια τα οποία είναι χτισμένα με βάση την πληροφορία απόστασης ή τοπολογίας. Όσον αφορά στην εξόρυξη γνώσης, αυτή η πληροφορία απόστασης παρέχεται στη βάση για τις αναγκαίες μετρήσεις ομοιότητας.

Η προσπέλαση των χωρικών δεδομένων μπορεί να είναι πιο πολύπλοκη από αυτή των μη χωρικών δεδομένων. Υπάρχουν ειδικές λειτουργίες και δομές δεδομένων που χρησιμοποιούνται για την προσπέλαση των χωρικών δεδομένων, οι οποίες αναπτύσσονται πιο κάτω.

2.2 Χωρικές ερωτήσεις

Εξαιτίας της πολυπλοκότητας των χωρικών λειτουργιών, κρίνεται σκόπιμη μία σύντομη αναφορά στο ζήτημα των χωρικών ερωτήσεων.

Μια παραδοσιακή ερώτηση επιλογής που προσπελαύνει μη χωρικά δεδομένα χρησιμοποιεί τις συνήθεις λειτουργίες σύγκρισης: $>,<, \leq,\geq,\neq$. Μια χωρική επιλογή (*spatial selection*) είναι μια επιλογή σε χωρικά δεδομένα, που μπορεί να χρησιμοποιεί διαφορετικές λειτουργίες σύγκρισης. Οι τύποι των τελεστών χωρικής σύγκρισης που θα μπορούσαν να χρησιμοποιηθούν περιλαμβάνουν τους ακόλουθους: «κοντά», «βόρεια», «νότια», «ανατολικά», «δυτικά», «περικλείεται από», «επικαλύπτει», «τέμνει», κ.ά.

Μια ειδική λειτουργία σύνδεσης που εφαρμόζεται σε δυο χωρικές σχέσεις ονομάζεται χωρική σύνδεση (*spatial join).* Κατά κάποιο τρόπο μία χωρική σύνδεση είναι παρόμοια με μια συνηθισμένη σχεσιακή σύνδεση, στην οποία δυο εγγραφές συνδέονται μεταξύ τους εάν έχουν κοινά χαρακτηριστικά. Σε μια παραδοσιακή σύνδεση, δυο εγγραφές πρέπει να έχουν κοινά γνωρίσματα, τα οποία ικανοποιούν μια προκαθορισμένη συσχέτιση (όπως ισότητα σε μια σύνδεση ισότητας). Σε μια χωρική σύνδεση, η συσχέτιση είναι χωρική. Ο τύπος της συσχέτισης βασίζεται στον τύπο του χωρικού χαρακτηριστικού (συσχέτιση «πλησιέστερο»: μπορεί να χρησιμοποιηθεί για σημεία / συσχέτιση «τομή»: χρησιμοποιείται για πολύγωνα).

Ένα χωρικό αντικείμενο συνήθως περιγράφεται από χωρικά και μη χωρικά γνωρίσματα. Σε αυτά μπορεί να περιλαμβάνεται κάποιος τύπος σχετικός με θέση. Το γνώρισμα αυτό της θέσης θα μπορούσε να προσδιορίζει ένα ακριβές σημείο (όπως ένα ζεύγος γεωγραφικού μήκους και πλάτους) ή μπορεί να είναι μια διεύθυνση ή ο ταχυδρομικός κώδικας μιας περιοχής. Συχνά, διαφορετικά χωρικά αντικείμενα αναγνωρίζονται από διαφορετικές θέσεις - και απαιτείται ένα είδος μετάφρασης από το ένα γνώρισμα στο άλλο προκειμένου να εκτελεστούν χωρικές λειτουργίες μεταξύ των διαφορετικών αντικειμένων.

Πολλές βασικές χωρικές ερωτήσεις μπορεί να βοηθήσουν σε εργασίες εξόρυξης γνώσης από δεδομένα:

- Μια ερώτηση περιοχής (*region query*) ή *ερώτηση εύρους* (*range query*) είναι μια ερώτηση που ζητά αντικείμενα που τέμνουν μια δοθείσα περιοχή στην ερώτηση.

- Μια ερώτηση πλησιέστερου γείτονα (*nearest neighbour query*) αναζητά αντικείμενα που είναι κοντά σε ένα, συγκεκριμένο αντικείμενο.

- Μια σάρωση απόστασης (*distance scan*) βρίσκει αντικείμενα εντός μιας προκαθορισμένης απόστασης από ένα συγκεκριμένο αντικείμενο, με την απόσταση να αυξάνεται σταδιακά.

Όλες αυτές οι ερωτήσεις μπορούν να χρησιμοποιηθούν για να βοηθήσουν μια συσταδοποίηση ή κατηγοριοποίηση [Dun04].

2.3 Οργάνωση Χωρικών Δεδομένων

Υπάρχουν πολλές δομές που έχουν σχεδιαστεί ειδικά για την αποθήκευση ή τη δεικτοδότηση των χωρικών δεδομένων. Αυτό συμβαίνει λόγω των μοναδικών γνωρισμάτων που τα χαρακτηρίζουν και προς τούτο στα επόμενα εξετάζονται σύντομα κάποιες από τις πιο γνωστές δομές αυτών.

2.3.1. Δομές Χωρικών Δεδομένων

Οι χωρικές δομές δεδομένων, σε αντίθεση με τις αλφαριθμητικές, διαχειρίζονται αντικείμενα με διαστάσεις, καθώς και σύνθετες σχέσεις που υφίστανται μεταξύ των αντικειμένων αυτών. Διακρίνονται σε **διανυσματικές δομές** και **δομές ψηφιδωτού**. Επίσης, μια σημαντική κατηγορία δομών είναι αυτή των **τοπολογικών δομών**, που έχουν σαν στόχο την ταχεία ανάκτηση των τοπολογικών σχέσεων μεταξύ διανυσμάτων.

Ειδικότερα, οι **δομές ψηφιδωτού** οργανώνουν δεδομένα τύπου ψηφιδωτού και επιτυγχάνουν αφενός τη συμπίεση των δεδομένων και αφετέρου την αποτελεσματική εκτέλεση ορισμένων λειτουργιών χωρικής ανάλυσης. Σε αυτή την κατηγορία ανήκουν η κωδικοποίηση κατά γραμμές και η δομή του τετραδικού δένδρου. Η πρώτη επιτυγχάνει συμπίεση των δεδομένων και . Το τετραδικό δένδρο δομεί το χώρο με αναδρομική διαίρεσή του σε τέσσερις περιοχές (τεταρτημόρια) και είναι πολύ αποτελεσματικό στην εκτέλεση ορισμένων λειτουργιών χωρικής ανάλυσης.

Οι **διανυσματικές δομές** κατηγοριοποιούνται ανάλογα με τον τύπο των διανυσμάτων προς δεικτοδότηση. Συγκεκριμένα διακρίνονται στις:

- **Σημειακές δομές**: αφορούν στη δεικτοδότηση σημειακών αντικειμένων. Αντιπροσωπευτικές δομές δεδομένων για σημειακά αντικείμενα στο χώρο των ν διαστάσεων είναι το $k-d-$ δένδρο, το $K-D-B-$ δένδρο, το αρχείο καννάβου και το σημειακό τετραδικό δένδρο.

- Γραμμικές **δομές**: έχουν προταθεί για να υποστηρίξουν την αποτελεσματική δεικτοδότηση γραμμικών αντικειμένων. Αντιπροσωπευτικές γραμμικές δομές αποτελούν το *strip* – δένδρο και παραλλαγές του τετραδικού δένδρου.
- Πολυ**γωνικές δομές**: έχουν σχεδιαστεί για τη δεικτοδότηση χωρικών αντικειμένων με συντεταγμένες και διαστάσεις στο χώρο των ν διαστάσεων. Συνεπώς, μπορούν να εφαρμοσθούν για την δεικτοδότηση σημειακών, γραμμικών, αλλά και πολυγωνικών αντικειμένων. Η αντιπροσωπευτικότερη δομή για την δεικτοδότηση πολυγωνικών αντικειμένων είναι το R – δένδρο και οι παραλλαγές του (R^+ - δένδρο, R^* - δένδρο, κ.ά.).

Οι ως άνω χωρικές δομές δεδομένων έχουν σχεδιασθεί για να υποστηρίξουν δύο τύπους αναζήτησης: (α) σημείου και (β) περιοχής. Μια αναζήτηση σημείου (πχ επιλογή με τον κέρσορα μίας χώρας σε έναν ψηφιακό χάρτη) στοχεύει στην ανάκτηση των αντικειμένων που τέμνουν ένα σημείο. Μια αναζήτηση περιοχής στοχεύει στην ανάκτηση των αντικειμένων που τέμνουν μια περιοχή (πχ μεγέθυνση μιας περιοχής ενός ψηφιακού χάρτη). Στην πραγματικότητα οι δύο τύποι αναζήτησης αφορούν σε ένα μοναδικό τύπο (αναζήτηση σημείου αποτελεί: αναζήτηση περιοχής με μηδενικό μέγεθος). Όπως προκύπτει από τα παραπάνω, σημαντικές διανυσματικές δομές δεδομένων εν γένει – αλλά και για τους σκοπούς της παρούσας μελέτης – συνιστούν το τετραδικό δένδρο (*quad - tree*), καθώς και το R – δένδρο (*R - tree*), τα οποία αναπτύσσονται στο ΠΑΡΑΡΤΗΜΑ Π1.

Οι **τοπολογικές δομές** δεδομένων, τέλος, συνιστούν κατάλληλες δομές, συνήθως με μορφή πίνακα (συλλογή σχεσιακών πινάκων), όπου οργανώνονται προϋπολογισμένες σχέσεις ή βοηθητικά στοιχεία κατά το κτίσιμο της γεωμετρικής βάσης. [Στεφ03]

Ένα πλεονέκτημα των δομών χωρικών δεδομένων είναι ότι συσταδοποιούν τα αντικείμενα βάσει της θέσης. Αυτό συνεπάγεται ότι αντικείμενα που είναι κοντά στο ν – διάστατο χώρο, τείνουν να αποθηκεύονται κοντά στη δομή δεδομένων και

στο δίσκο. Επομένως, αυτές οι δομές θα μπορούσαν να χρησιμοποιηθούν για να μειώσουν το κόστος εκτέλεσης ενός αλγορίθμου, περιορίζοντας το χώρο αναζήτησης.

2.4 Βασικές Αρχές Εξόρυξης Γνώσης από Χωρικά Δεδομένα

Προκειμένου για την πληρέστερη παρουσίαση του ζητήματος της εξόρυξης γνώσης από χωρικά δεδομένα, παρακάτω γίνεται παρουσίαση ορισμένων από τις βασικές αρχές της.

Έστω ότι τα A και B είναι χωρικά αντικείμενα σε ένα δισδιάστατο χώρο. Μπορεί να θεωρηθεί πως κάθε αντικείμενο αποτελείται από ένα σύνολο σημείων στο χώρο: $\langle x_a, y_a \in A \rangle, \langle x_b, y_b \in B \rangle$. Μπορεί να υπάρχουν πολλές τοπολογικές σχέσεις μεταξύ δυο χωρικών αντικειμένων. Αυτές οι σχέσεις βασίζονται στους τρόπους με τους οποίους δυο αντικείμενα τοποθετούνται γεωγραφικά:

- **Ξένο**: το *A* είναι ξένο (*disjoint*) ως προς το *B*, εάν δεν υπάρχουν σημεία στο *A* που να περιέχονται στο *B*.
- **Έχει επικάλυψη ή τέμνει**: το *A* έχει επικάλυψη με (*overlaps*) ή τέμνει (*disjoints*) το *B*, εάν υπάρχει τουλάχιστον ένα σημείο στο *A* που να ανήκει και στο B.
- **Είναι ίσο**: το A είναι ίσο με (*equals*) το *B*, εάν έχουν όλα τα σημεία τους κοινά.
- **Καλύπτεται από ή βρίσκεται εντός ή περιέχεται σε**: το *A* καλύπτεται από (*covered by*) ή βρίσκεται εντός (*inside*) ή περιέχεται στο (*contained in*) *B*, εάν όλα τα σημεία του *A* ανήκουν στο *B*.
- **Καλύπτει ή περιέχει**: το *A* καλύπτει (*covers*) ή περιέχει (*contains*) το *B* και μόνο εάν το *B* καλύπτεται από ή περιέχεται στο *A*.

Βάσει της τοποθέτησης των αντικειμένων στο χώρο, μπορούν να ορισθούν σχέσεις ως προς κατεύθυνση, με την προσθήκη του προσανατολισμού του χάρτη στο χώρο («βόρεια», «νότια», «ανατολικά» κοκ). [EFKS00]

Τα μέτρα Ευκλείδειας και Μανχάταν απόστασης χρησιμοποιούνται συχνά για την μέτρηση της απόστασης. Η απόσταση μεταξύ δυο χωρικών αντικειμένων μπορεί να οριστεί ως επέκταση των παραδοσιακών ορισμών της Ευκλείδειας και

Μανχάταν απόστασης:

- **Ελάχιστη**: $dis(A,B) = \min_{(x_a,y_a) \in A, (x_b,y_b) \in B} dis((x_a, y_a),(x_b, y_b))$
- **Μέγιστη**: $dis(A,B) = \max_{(x_a,y_a) \in A, (x_b,y_b) \in B} dis((x_a, y_a),(x_b, y_b))$
- **Μέση**: $dis(A,B) = average_{(x_a,y_a) \in A, (x_b,y_b) \in B} dis((x_a, y_a),(x_b, y_b))$
- **Κεντρική**: $dis((x_{ca}, y_{ca}),(x_{cb}, y_b))$,

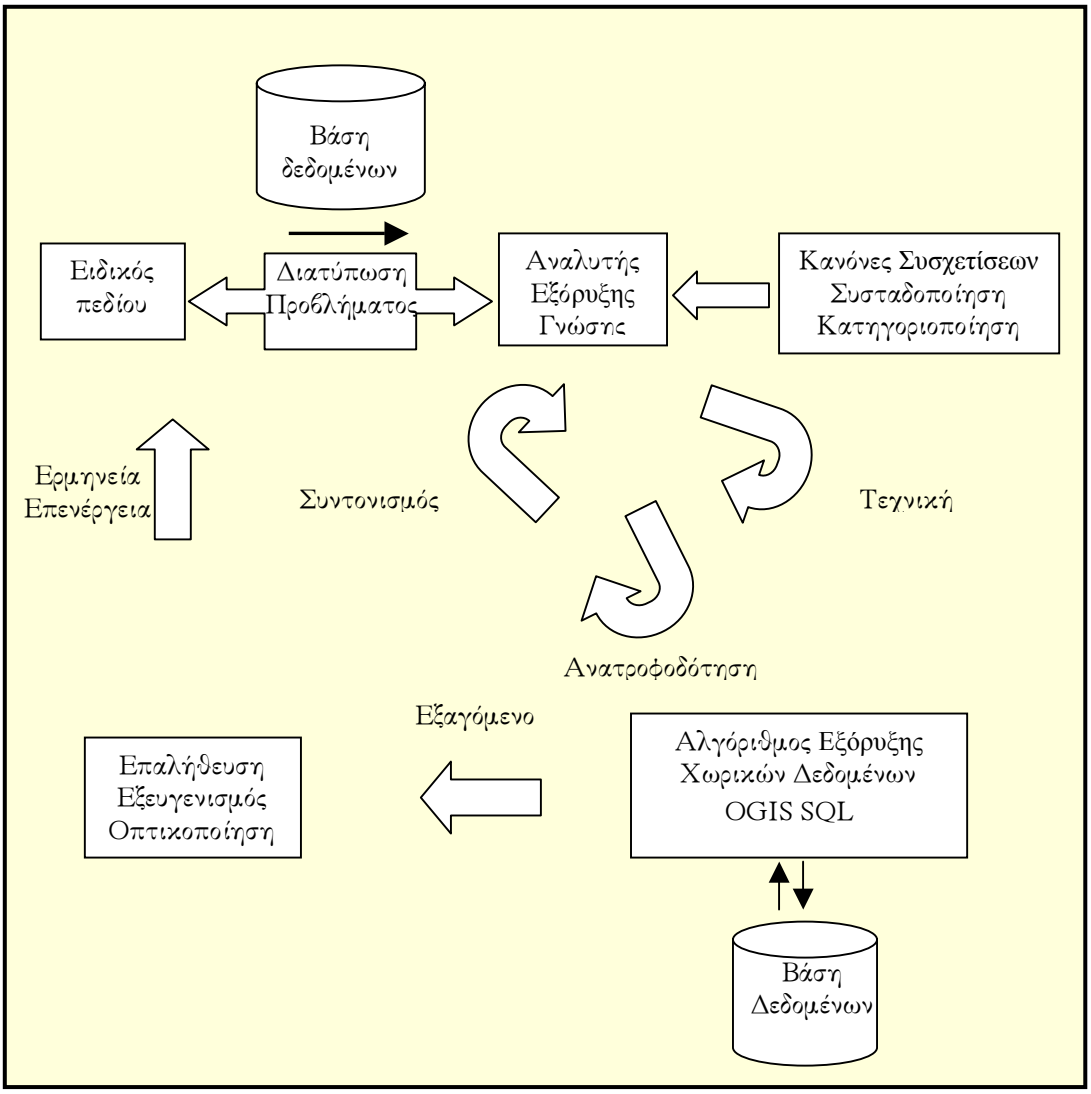
όπου (x_{ca}, y_{ca}) είναι ένα κεντροειδές για το αντικείμενο A και, αντίστοιχα, το (x_{cb}, y_b) για το B.

Για παράδειγμα, το χωρικό αντικείμενο θα μπορούσε να θεωρηθεί ως μια συστάδα των σημείων εντός του. Τα κεντροειδή που χρησιμοποιήθηκαν για τον τελευταίο τύπο μέτρησης της απόστασης, μπορούν να αναγνωρισθούν βρίσκοντας το γεωμετρικό κέντρο του. Έτσι, εάν χρησιμοποιείται ένα ελάχιστο περιβάλλον ορθογώνιο (*Minimum Bounding Rectangle – MBR*), η απόσταση μεταξύ των αντικειμένων μπορεί να βρεθεί χρησιμοποιώντας την Ευκλείδεια απόσταση μεταξύ των κέντρων των ορθογωνίων αυτών για τα δύο αντικείμενα.

Τα χωρικά αντικείμενα μπορούν να ανακτηθούν βάσει λειτουργιών επιλογής, συνάθροισης ή σύνδεσης. Μια επιλογή μπορεί να εφαρμοσθεί πάνω στα χωρικά ή μη χωρικά γνωρίσματα. Η ανάκτηση που βασίζεται στα χωρικά γνωρίσματα μπορεί να εφαρμοσθεί χρησιμοποιώντας έναν από τους χωρικούς τελεστές. Μια χωρική σύνδεση ανακτά τα χωρικά αντικείμενα βάσει της συσχέτισης μεταξύ τους. [Dun04]

2.5 Αλγόριθμοι Εξόρυξης Χωρικών Δεδομένων

Η διαδικασία εξόρυξης χωρικής γνώσης συνοψίζεται πιο κάτω (Σχήμα 2.1).



Σχήμα 2.1: Εξόρυξη Χωρικής Γνώσης από Δεδομένα
Πηγή: [SC03]

Η διαδικασία αυτή εμπεριέχει στενή αλληλεπίδραση ανάμεσα σε έναν ειδικό πεδίου και έναν αναλυτή εξόρυξης γνώσης. Το εξαγόμενο της διαδικασίας είναι ένα σύνολο υποθέσεων (πρότυπα), τα οποία μπορούν να επιβεβαιωθούν με ακρίβεια κάνοντας χρήση στατιστικών εργαλείων και να οπτικοποιηθούν με τη χρήση γεωγραφικών συστημάτων πληροφοριών. Τελικά, ο αναλυτής μπορεί να

ερμηνεύσει τα πρότυπα, να δημιουργήσει και να προτείνει τις κατάλληλες ενέργειες.

Μία ιδανική περίπτωση ενός αλγόριθμου εξόρυξης χωρικών δεδομένων θα έπρεπε να σχεδιασθεί έτσι ώστε να έχει απευθείας πρόσβαση στη βάση δεδομένων. Η επιλογή της τεχνικής και η εκλογή του κατάλληλου αλγορίθμου είναι μια επαναληπτική διαδικασία. Έτσι, οι περισσότεροι αλγόριθμοι απαιτούν την προσαρμογή παραμέτρων που ορίζονται από τον χρήστη, ενώ στην πλειοψηφία των περιπτώσεων δεν υπάρχει τρόπος να προκύψει εκ των προτέρων ποιες παράμετροι είναι κατάλληλες για την συγκεκριμένη βάση δεδομένων. [SC03]

Το ζήτημα της Εξόρυξης Γνώσης από Χωρικά Δεδομένα είναι αρκετά μεγάλο και δεν είναι δυνατόν να εξαντληθεί στα πλαίσια αυτής της μελέτης. Στα επόμενα, ωστόσο γίνεται μία προσπάθεια παρουσίασης αλγορίθμων εξόρυξης χωρικής γνώσης, οι οποίοι αφορούν στην , .

2.5.1 Γενίκευση και Εξειδίκευση

Η χρήση μιας ιεραρχίας εννοιών δείχνει επίπεδα σχέσεων ανάμεσα στα δεδομένα. Όταν εφαρμόζονται σε χαρακτηριστικά χωρικών δεδομένων, οι ιεραρχίες εννοιών επιτρέπουν την ανάπτυξη κανόνων και σχέσεων σε διαφορετικά επίπεδα στην ιεραρχία. Κάτι αντίστοιχο χρησιμοποιείται και στους γενικευμένους κανόνες συσχέτισης, καθώς και στις αρχές γενίκευσης και εξειδίκευσης, οι οποίες βρίσκουν εφαρμογή στην μηχανική μάθηση. Παρόλα αυτά, στις ως άνω περιπτώσεις η ιεραρχία δεν σχετίζεται αναγκαία με χωρικά δεδομένα. Οι τεχνικές εξόρυξης γνώσης σε χωρικά δεδομένα εμπλέκουν και τις δύο προσεγγίσεις τύπου γενίκευσης και εξειδίκευσης.

2.5.2 Προοδευτική Βελτίωση

Εξαιτίας του μεγάλου όγκου δεδομένων που υπάρχουν στις χωρικές εφαρμογές, μπορεί δοθούν προσεγγιστικές απαντήσεις, προτού αναζητηθούν πιο ακριβείς. Η χρήση των ελάχιστων περιβαλλόντων ορθογωνίων είναι μια μέθοδος προσέγγισης

του σχήματος ενός αντικειμένου. Τα τετραδικά δένδρα, τα R – δένδρα και οι περισσότερες τεχνικές χωρικής δεικτοδότησης χρησιμοποιούν ένα είδος προοδευτικής βελτίωσης. Εκτιμούν το σχήμα των αντικειμένων σε υψηλότερα επίπεδα στη δενδρική δομή και οι είσοδοι των χαμηλότερων επιπέδων παρέχουν πιο ακριβείς περιγραφές των χωρικών αντικειμένων. Η προοδευτική βελτίωση (*progressive refinement*) μπορεί να θεωρηθεί ως ένα φιλτράρισμα των δεδομένων που δεν είναι εφαρμόσιμα σε ένα πρόβλημα.

Με την προοδευτική βελτίωση, τα ιεραρχικά επίπεδα βασίζονται σε χωρικές συσχετίσεις.

2.5.3 Γενίκευση

Η γενίκευση (*generalization*) καθοδηγείται από μια ιεραρχία εννοιών και μπορεί να θεωρηθεί ως η διαδικασία εξαγωγής πληροφορίας σε ένα υψηλό επίπεδο, που βασίζεται σε πληροφορία χαμηλότερων επιπέδων. Οι ιεραρχίες εννοιών για χωρικά δεδομένα μπορεί να είναι χωρικές ή μη χωρικές. Μια χωρική ιεραρχία (*spatial hierarchy*) δείχνει τις σχέσεις μεταξύ γεωγραφικών περιοχών. Η γενίκευση μπορεί να εκτελεστεί χρησιμοποιώντας οποιαδήποτε από αυτές τις δύο ιεραρχίες. Όταν γενικεύονται τα χωρικά δεδομένα, τα μη χωρικά πρέπει να τροποποιούνται κατάλληλα, ώστε να αντικατοπτρίζουν τα μη χωρικά δεδομένα που σχετίζονται με τα νέα χωρικά δεδομένα. Παρόμοια, όταν γενικεύονται τα μη χωρικά δεδομένα, τα χωρικά δεδομένα πρέπει να τροποποιούνται κατάλληλα. Χρησιμοποιώντας αυτά τα δύο είδη ιεραρχιών, η γενίκευση, όπως εφαρμόζεται στα χωρικά δεδομένα, μπορεί να διαιρεθεί σε δύο υποκλάσεις: γενίκευση χωρικής τάξης (*spatial data dominant generalization*) και γενίκευση μη χωρικής τάξης (*nonspatial data dominant generalization* [LH093]. Και οι δύο υποκλάσεις μπορεί να θεωρηθούν ως ένα είδος συσταδοποίησης. Η γενίκευση χωρικής τάξης πραγματοποιεί τη συσταδοποίηση που βασίζεται σε χωρικές θέσεις (έτσι ώστε να ομαδοποιούνται κοντινά αντικείμενα), ενώ η γενίκευση μη χωρικής τάξης συσταδοποιεί βάσει της ομοιότητας των τιμών μη χωρικών γνωρισμάτων. Αυτές οι προσεγγίσεις αναφέρονται ως μια επαγωγή προσανατολισμένη σε γνωρίσματα (*attribute*

oriented induction) επειδή η διαδικασία γενίκευσης βασίζεται σε τιμές γνωρισμάτων.

Με την γενίκευση χωρικής τάξης, η γενίκευση αρχικά εφαρμόζεται στα χωρικά δεδομένα και στη συνέχεια τα σχετιζόμενα μη χωρικά γνωρίσματα τροποποιούνται ανάλογα. Η γενίκευση εφαρμόζεται έως έναν αριθμό περιοχών, που θεωρείται κατώφλι. Για παράδειγμα, ο προσδιορισμός της μέσης βροχόπτωσης σε διάφορες περιφέρειες θα μπορούσε να γίνει βρίσκοντας τη μέση βροχόπτωση για όλες τις επιμέρους περιοχές που απεικονίζονται από μια χωρική ιεραρχία. Επομένως, η χωρική ιεραρχία καθορίζει ποιες περιοχές χαμηλού επιπέδου βρίσκονται στην περιοχή υψηλού επιπέδου που εξετάζεται. Ο καθορισμός του τρόπου εφαρμογής της γενίκευσης σε μη χωρικά δεδομένα, δεν είναι, παρόλα αυτά, μια προφανής διαδικασία συνάθροισης. Στην πράξη ο τρόπος καθορισμού της μέσης βροχόπτωσης σε αυτήν την περίπτωση είναι ο ίδιος για κάθε επιμέρους περιοχή. Όμως, απαιτείται μια διαδικασία στάθμισης, η οποία θα παρέχει μια πιο ακριβή μέση τιμή βροχόπτωσης για την περιοχή υψηλότερου επιπέδου.

Μια εναλλακτική προσέγγιση είναι η γενίκευση των τιμών των μη χωρικών γνωρισμάτων. Η γενίκευση βασίζεται στην ομαδοποίηση των δεδομένων. Οι γειτονικές περιοχές συγχωνεύονται εάν έχουν τις ίδιες γενικευμένες τιμές για τα μη χωρικά δεδομένα. Εάν αντί των μέσων τιμών βροχόπτωσης επιστρέφονται απλώς τιμές που αναπαριστούν κάποια συστάδα, θα μπορούσαν να ανατεθούν τιμές όπως ισχυρή, μέτρια, ασθενής κοκ. για να περιγραφεί η βροχόπτωση αντί να δοθούν πραγματικές αριθμητικές τιμές. Μπορεί να δίνεται ένα κατώφλι που καθορίζει το μέγιστο αριθμό περιοχών. Βάσει αυτού του κατωφλίου, επιλέγεται το σωστό επίπεδο στην ιεραρχία και επομένως καθορίζεται ο αριθμός των περιοχών.

Η τεχνική γενίκευσης μη χωρικής *τάξης* λειτουργεί με έναν παρόμοιο τρόπο. Το πρώτο βήμα σε αυτόν τον αλγόριθμο είναι η ανάκτηση των δεδομένων βάσει των κριτηρίων μη χωρικής επιλογής που διατυπώνονται στην ερώτηση. Στη συνέχεια εφαρμόζεται η απαιτούμενη προσανατολισμένη στα γνωρίσματα επαγωγή στα ανακτώμενα μη χωρικά δεδομένα. Για να γίνει αυτό, λαμβάνονται υπόψη οι μη

χωρικές εννοιολογικές ιεραρχίες. Κατά τη διάρκεια αυτού του βήματος, γενικεύονται οι τιμές των μη χωρικών δεδομένων σε τιμές πιο υψηλών επιπέδων. Αυτές οι γενικεύσεις είναι συνοπτικές τιμές υψηλότερων επιπέδων των συγκεκριμένων τιμών χαμηλότερων επιπέδων.

Για παράδειγμα, εάν γενικευόταν η μέση θερμοκρασία, θα μπορούσαν να συνδυαστούν διαφορετικές μέσες θερμοκρασίες (ή διαστήματα) και να τους αποδοθεί η ετικέτα «ζέστη». Στην συνέχεια θα εφαρμοζόταν μια γενίκευση χωρικά προσανατολισμένη, όπου συγχωνεύονται οι γειτονικές περιοχές με τις ίδιες (ή παρόμοιες) γενικευμένες μη χωρικές τιμές. Αυτό αποσκοπεί στην μείωση των περιοχών που επιστρέφονται στην απάντηση της ερώτησης.
Ένα αρνητικό αυτών των προσεγγίσεων είναι ότι η ιεραρχία πρέπει να προκαθορισθεί από ειδικούς του πεδίου και η ποιότητα οποιωνδήποτε αιτημάτων, για διαχείριση δεδομένων εξαρτάται από τη δοθείσα ιεραρχία. Η πολυπλοκότητα δημιουργίας των ιεραρχιών είναι $O(nlogn)$. [Dun04]

2.5.4 Χωρικοί Κανόνες

Μπορούν να παραχθούν χωρικοί κανόνες που να περιγράφουν τη συσχέτιση μεταξύ και τη δομή των χωρικών αντικειμένων. Υπάρχουν τρεις τύποι κανόνων που μπορούν να βρεθούν κατά τη διάρκεια της εξόρυξης γνώσης από χωρικά δεδομένα [KAH96]. Οι κανόνες χωρικών χαρακτηριστικών (*spatial characteristic rules*) περιγράφουν τα δεδομένα. Οι κανόνες χωρικών διαχωρισμών (*spatial discriminant rules*) περιγράφουν τις διαφορές μεταξύ διαφορετικών κλάσεων των δεδομένων (τα χαρακτηριστικά που διαφοροποιούν τις διαφορετικές κλάσεις). Οι κανόνες χωρικών συσχετίσεων (*spatial association rules*) είναι συνεπαγωγές ενός συνόλου δεδομένων από ένα άλλο.

Ειδικότερα, οι κανόνες χωρικών συσχετίσεων συνιστούν κανόνες συσχετίσεων για αντικείμενα χωρικών δεδομένων. Είτε το πρότερο (*antecedent*), είτε το απότοκο (*consequent*) του κανόνα πρέπει να περιέχει κάποια χωρικά κατηγορήματα (π.χ. κοντά).

Βάσει ενός απλού αλγόριθμου δημιουργίας κανόνων χωρικών συσχετίσεων, παράγονται όλοι αυτοί οι κανόνες που ικανοποιούν την ελάχιστη εμπιστοσύνη και υποστήριξη. Εξαιτίας της μεγάλης πιθανότητας για τοπολογικές συσχετίσεις, θεωρείται ότι η αίτηση για εξόρυξη γνώσης από δεδομένα καθορίζει ποιο(-α) χωρικό(-ά) κατηγόρημα(-ατα) θα χρησιμοποιηθεί. Από τη στιγμή που καθοριστεί το σχετικό υποσύνολο της βάσης, αναγνωρίζονται συσχετίσεις αυτού του τύπου. Αρχικά γίνεται η υπόθεση πως χρησιμοποιούνται οι «γενικευμένες» εκδοχές των τοπολογικών συσχετίσεων. Οι γενικευμένες συσχετίσεις ικανοποιούνται εάν κάποια αντικείμενα υψηλότερα στην ιεραρχία εννοιών τις ικανοποιούν. Σε αυτό το επίπεδο, εφαρμόζεται ένα φιλτράρισμα για την απομάκρυνση αντικειμένων που πιθανόν δεν θα μπορούσαν να ικανοποιούν τη συσχέτιση.

Προκειμένου να εξηγηθεί η έννοια της γενίκευσης με τις χωρικές συσχετίσεις, γίνεται η υπόθεση ότι η τοπολογική σχέση που εξετάζεται είναι η «κοντά». Το σύστημα *GIS* ορίζει τι ακριβώς σημαίνει αυτό το κατηγόρημα. Για παράδειγμα, θα μπορούσε να ορίσει τη σχέση αυτή βάσει της Ευκλείδειας απόστασης μεταξύ των δυο χωρικών αντικειμένων. Επιπλέον, μπορεί να ορισθεί διαφορετικά βάσει του τύπου των αντικειμένων στην ερώτηση. Η γενίκευση του «κοντά» που γράφεται «γεν_κοντά» (γενικευμένο κοντά) μπορεί να οριστεί με μια ιεραρχία που δείχνει ότι το «γεν_κοντά» περιέχει το «κοντά» όπως και άλλα κατηγορήματα (όπως το «περιέχει» ή το «ίσο»). Ένα πρώτο βήμα για τον καθορισμό του κατά πόσο ικανοποιείται το κατηγόρημα «κοντά» θα είναι να κοιτάξουμε γενικότερα πόσο ικανοποιείται το «γεν_κοντά». Η γενικευμένη αποτίμηση χρησιμοποιείται ως ένα είδος φίλτρου για τον αποδοτικό αποκλεισμό αντικειμένων που πιθανόν δε θα μπορούσαν να ικανοποιούν το αληθές κατηγόρημα. Το ευρύτερο κατηγόρημα «ευρύτερο__γεν_κοντά» ικανοποιείται από αντικείμενα εάν τα ελάχιστα περιβάλλοντα ορθογώνια αυτών ικανοποιούν το «γεν_κοντά». Μόνο τα αντικείμενα που ικανοποιούν το «ευρύτερο_γεν_κοντά» εξετάζονται ώστε να καταστεί σαφές εάν ικανοποιούν το «γεν_κοντά».

Ο χαρακτηρισμός (*characterization*) είναι η διαδικασία εύρεσης μιας περιγραφής για μια βάση δεδομένων ή για κάποιο τμήμα της. Όλοι αυτοί οι κανόνες μπορεί να θεωρηθούν ως ειδικοί τύποι χαρακτηρισμών. Ο κανόνας χωρικού χαρακτηριστικού είναι ο απλούστερος.

Μια άλλη συνήθης προσέγγιση σύνοψης χωρικών δεδομένων είναι αυτή της εκτέλεσης μιας ανίχνευσης τάσης (*data detection*), η οποία μπορεί να θεωρηθεί ως μια τυπική αλλαγή σε μια ή περισσότερες τιμές μη χωρικών γνωρισμάτων για χωρικά αντικείμενα, καθώς απομακρύνεται κανείς από ένα άλλο χωρικό αντικείμενο [EFKS98]. Για την ανίχνευση μιας τάσης, μπορεί να χρησιμοποιηθεί ανάλυση παλινδρόμησης.

2.5.5 Αλγόριθμοι Χωρικής Κατηγοριοποίησης

Τα προβλήματα χωρικής κατηγοριοποίησης χρησιμοποιούνται για την διαμέριση συνόλων χωρικών αντικειμένων. Τα χωρικά αντικείμενα μπορούν να κατηγοριοποιηθούν με χρήση μη χωρικών γνωρισμάτων, χωρικών κατηγορημάτων (χωρικών γνωρισμάτων), ή χωρικών και μη χωρικών γνωρισμάτων. Μπορεί επίσης να χρησιμοποιηθούν ιεραρχίες εννοιών, όπως και δειγματοληψία. Όπως και με τους άλλους τρόπους εξόρυξης γνώσης από χωρικά δεδομένα, μπορούν να χρησιμοποιηθούν τεχνικές γενίκευσης και προοδευτικής βελτίωσης, για την βελτίωση της αποδοτικότητας. Ένας σημαντικός αλγόριθμος χωρικής κατηγοριοποίησης είναι ο *ID3*, ο οποίος βασίζεται στο «χτίσιμο» ενός δένδρου απόφασης. Λεπτομέρειες για τον αλγόριθμο *ID3* παρατίθενται στο ΠΑΡΑΡΤΗΜΑ Π1.

2.5.5.1 Επέκταση του *ID3*

Για την πραγματοποίηση κατηγοριοποίησης χωρικών αντικειμένων με χρήση μιας επέκτασης του *ID3*, έχει εφαρμοσθεί η έννοια των γράφων γειτνίασης [EKS97]. Ένας γράφος γειτνίασης (*neighborhood grahp*) είναι ένας γράφος που κτίζεται από αντικείμενα στο χώρο. Κάθε αντικείμενο γίνεται ένας κόμβος του γράφου. Οι ακμές κατασκευάζονται από τους γείτονες. Δηλαδή, δύο κόμβοι συνδέονται με μία ακμή στο γράφο γειτνίασης, εάν ο ένας είναι γείτονας του

άλλου. Ο «γείτονας»μπορεί να ορισθεί βάσει οποιασδήποτε συσχέτισης μεταξύ των χωρικών αντικειμένων, όπως απόσταση μικρότερη από κάποιο κατώφλι, ικανοποίηση μιας τοπολογικής σχέσης μεταξύ των αντικειμένων, ή σχέση κατεύθυνσης. Σημειώνεται ότι κάποιες από τις σχέσεις είναι σχέσεις διάταξης και κάποιες άλλες όχι.

Η ιδέα του αλγορίθμου είναι να λάβει υπόψη του τα αντικείμενα που είναι κοντά σε ένα δοθέν αντικείμενο. Ένας δείκτης μέγιστου μήκους δίδεται ως είσοδος που καθορίζει το μέγιστο μήκος ενός μονοπατιού γειτονικότητας που ξεκινά από έναν κόμβο. Αυτό στη συνέχεια προσδιορίζει ένα σύνολο από κόμβους που σχετίζονται με τον κόμβο-στόχο. Στη συνέχεια ο *ID3* θεωρεί για σκοπούς κατηγοριοποίησης όχι μόνο τα μη χωρικά γνωρίσματα του αντικειμένου-στόχου, αλλά και αυτά στα γειτονικά αντικείμενα.

2.5.5.2 Δένδρο χωρικής απόφασης

Μια τεχνική χωρικής κατηγοριοποίησης χτίζει δένδρα αποφάσεων χρησιμοποιώντας μια διαδικασία παρόμοια με αυτή που χρησιμοποιήθηκε για τους κανόνες συσχετίσεων [KHS98]. Η βάση αυτής της προσέγγισης είναι ότι τα χωρικά αντικείμενα μπορούν να περιγραφούν βάσει των αντικειμένου που είναι κοντά σε αυτά. Στη συνέχεια θεωρείται μια περιγραφή των κλάσεων βασισμένη σε μια συνάθροιση των πιο σχετικών κατηγορημάτων για κοντινά αντικείμενα. Για την κατασκευή του δένδρου απόφασης, πρώτα ορίζονται τα πιο σχετικά (χωρικά και μη) κατηγορήματα. Αυτά τα σχετικά κατηγορήματα είναι εκείνα που θα χρησιμοποιηθούν για το χτίσιμο του δένδρου απόφασης. Γίνεται η υπόθεση ότι ένα δείγμα εκπαίδευσης χρησιμοποιείται για να πραγματοποιήσει αυτό το βήμα και ότι ανατίθενται βάρη σε γνωρίσματα και κατηγορήματα. Τα αρχικά βάρη είναι 0. Για κάθε αντικείμενο, εξετάζονται δυο αντίστοιχα αντικείμενα. Η πλησιέστερη αστοχία (*nearest miss*) είναι το κοντινότερο χωρικό αντικείμενο στο αντικείμενο-στόχο, το οποίο ανήκει σε διαφορετική κλάση. Η πλησιέστερη επιτυχία (*nearest hit*) είναι ο κοντινότερος στόχος στην ίδια κλάση.

Για κάθε τιμή κατηγορήματος στο αντικείμενο-στόχο, εάν η πλησιέστερη

επιτυχία έχει την ίδια τιμή, τότε το βάρος του κατηγορήματος αυξάνεται. Εάν έχει διαφορετική τιμή, τότε μειώνεται. Παρόμοια, το βάρος μειώνεται (αυξάνεται) εάν η πλησιέστερη αστοχία έχει την ίδια (διαφορετική) τιμή. Μόνο κατηγορήματα με θετικά βάρη μεγαλύτερα από κάποιο προκαθορισμένο κατώφλι χρησιμοποιούνται στη συνέχεια για την κατασκευή του δένδρου. Προτείνεται, εξαιτίας της πολυπλοκότητας εύρεσης σχετικών κατηγορημάτων να βρίσκονται πρώτα σχετικά κατηγορήματα σε ένα ευρύτερο επίπεδο και στη συνέχεια σε ένα περιορισμένο. Αρχικά χρησιμοποιούνται τα ελάχιστα περιβάλλοντα ορθογώνια αντί των πραγματικών αντικειμένων και μια γενικευμένη ευρύτερη σχέση «κοντά» για την εύρεση των σχετικών κατηγορημάτων. Στη συνέχεια, κατά το δεύτερο πέρασμα, χρησιμοποιούνται αυτά τα σχετικά κατηγορήματα μαζί με τα πραγματικά αντικείμενα.

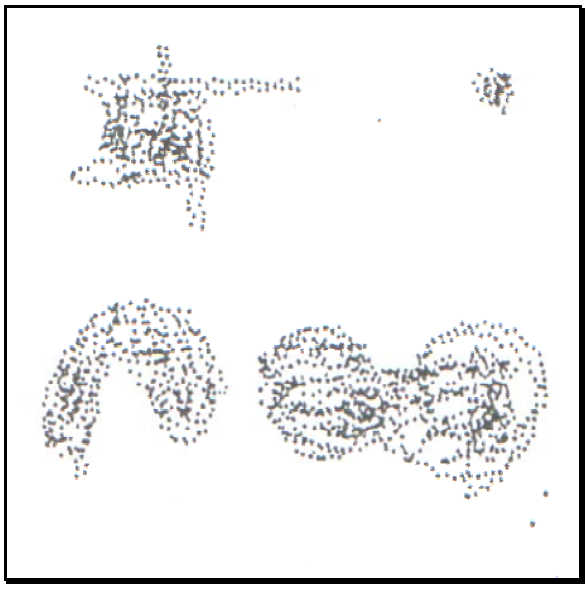
Για κάθε αντικείμενο του δείγματος, εξετάζεται η περιοχή γύρω από αυτό, η οποία καλείται ενδιάμεση ζώνη (*buffer*). Μια περιγραφή αυτής της ενδιάμεσης περιοχής δημιουργείται με την συνάθροιση των τιμών των πιο σχετικών κατηγορημάτων των αντικειμένων στην ενδιάμεση περιοχή. Προφανώς, το μέγεθος και το σχήμα της ενδιάμεσης ζώνης επιδρούν στον προκύπτοντα αλγόριθμο κατηγοριοποίησης. Είναι πιθανό, αν και μη ρεαλιστικό, να πραγματοποιηθεί μια εξαντλητική αναζήτηση σε όλα τα πιθανά μεγέθη και σχήματα ενδιάμεσων περιοχών. Ο αντικειμενικός στόχος θα ήταν να επιλεγεί εκείνη η ενδιάμεση ζώνη που οδηγεί στην καλύτερη διάκριση μεταξύ των κλάσεων στο σύνολο εκπαίδευσης. Αυτό θα υπολογιζόταν χρησιμοποιώντας το κέρδος πληροφορίας (*information gain*). Εξετάστηκαν και άλλες προσεγγίσεις βασισμένες στην επιλογή ενός συγκεκριμένου σχήματος, αλλά τελικά χρησιμοποιήθηκαν κύκλοι (ενδιάμεσες ισαπέχουσες ζώνες).

Για την κατασκευή του δένδρου γίνεται η υπόθεση ότι κάθε αντικείμενο του δείγματος συσχετίζεται με ένα σύνολο από γενικευμένα κατηγορήματα, τα οποία ικανοποιεί. Μπορούν τότε να καθορισθούν οι αριθμοί των αντικειμένων που ικανοποιούν ή δεν ικανοποιούν κάθε κατηγόρημα. Αυτό χρησιμοποιείται στη συνέχεια για να υπολογισθεί το κέρδος της πληροφορίας, όπως γίνεται στον *ID*3.

Αντί να δημιουργείται ένα δένδρο διακλαδώσεων πολλών δρόμων, δημιουργείται ένα δυαδικό δένδρο απόφασης.

2.5.6 Αλγόριθμοι Χωρικής Συσταδοποίησης

Οι αλγόριθμοι χωρικής συσταδοποίησης πρέπει να είναι σε θέση να δουλεύουν αποδοτικά με μεγάλες πολυδιάστατες βάσεις δεδομένων. Επιπλέον, θα πρέπει να μπορούν να εντοπίζουν συστάδες από διαφορετικά σχήματα. Το Σχήμα 2.2 δείχνει συστάδες σε έναν δισδιάστατο χώρο.



Σχήμα 2.2: Διαφορετικά σχήματα χωρικών συστάδων

Κάθε μία από τις συστάδες αυτές έχει ένα ακανόνιστο σχήμα. Ένας αλγόριθμος χωρικής συσταδοποίησης θα πρέπει να μπορεί να εντοπίζει αυτές τις τέσσερις συστάδες, αν και τα σχήματά τους δεν είναι κανονικά και κάποια σημεία σε μια συστάδα μπορεί να είναι πιο κοντά σε κάποια σημεία άλλων συστάδων, παρά σε σημεία της δικής τους συστάδας. Ένας αλγόριθμος που δουλεύει χρησιμοποιώντας κέντρα βάρους και απλές μετρήσεις απόστασης, πιθανόν δεν θα είναι σε θέση να αναγνωρίζει τα ασυνήθιστα σχήματα.

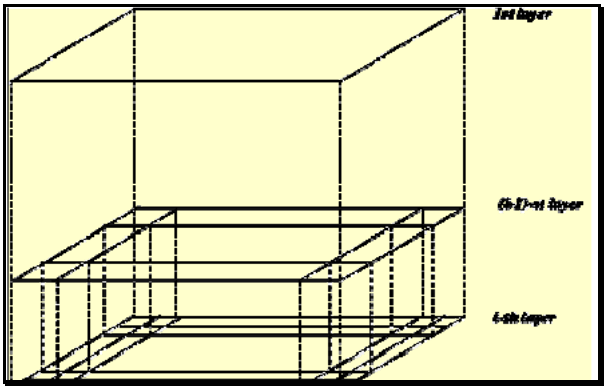
Άλλα επιθυμητά χαρακτηριστικά για την χωρική συσταδοποίηση είναι οι συστάδες που ανακαλύπτονται να είναι ανεξάρτητες της σειράς με την οποία εξετάστηκαν τα σημεία στο χώρο και να μην επηρεάζονται από ακραία σημεία.

Στο Σχήμα 2.2 τα ακραία σημεία στο κάτω δεξιά τμήμα του σχήματος δεν πρέπει να προστεθούν στη μεγάλη συστάδα κοντά σε αυτά.

Στα επόμενα παρουσιάζονται αλγόριθμοι συσταδοποίησης ειδικά σχεδιασμένοι για χωρικά δεδομένα.

2.5.6.1 *STING*

Βάσει του αλγορίθμου *STING (STatistical Information Grid approach)*, ο χώρος διαιρείται σε ορθογωνικά κελιά (Σχήμα 2.3), ενώ υπάρχουν ορισμένα επίπεδα κελιών τα οποία αντιστοιχούν σε διαφορετικά επίπεδα εμπιστοσύνης [WYM97].



Σχήμα 2.3: Διαίρεση του χώρου βάσει αλγορίθμου *STING*
Πηγή: [WYM97]

Πιο συγκεκριμένα, κάθε κελί σε ένα υψηλό επίπεδο διαιρείται σε μικρότερα, στο επόμενο κατώτερο επίπεδο. Η στατιστική πληροφορία για κάθε κελί υπολογίζεται και αποθηκεύεται από πριν και εν συνεχεία χρησιμοποιείται στην απάντηση ερωτημάτων. Οι παράμετροι για τα κελιά υψηλότερου επιπέδου μπορούν εύκολα να υπολογισθούν από τις παραμέτρους των κελιών σε χαμηλότερα επίπεδα (μέσος όρος, τυπική απόκλιση, ελάχιστο / μέγιστο, τύπος κατανομής – κανονική / τυπική), κλπ). Ο αλγόριθμος *STING* χρησιμοποιεί μία προσέγγιση από πάνω προς τα κάτω προκειμένου να δώσει απάντηση σε χωρικά ερωτήματα. Ξεκινά από ένα προεπιλεγμένο θεματικό επίπεδο με μικρό αριθμό κελιών και για κάθε κελί στο τρέχον επίπεδο υπολογίζει το διάστημα εμπιστοσύνης, το οποίο αντικατοπτρίζει την σχέση του κελιού με το δεδομένο

ερώτημα. Μετακινεί τα μη σχετιζόμενα κελιά και όταν η εξέταση του τρέχοντος θεματικού επιπέδου ολοκληρωθεί, προχωρά στο επόμενο κατώτερο επίπεδο. Η διαδικασία επαναλαμβάνεται μέχρι το κατώτατο επίπεδο.
[www.ted.unipi.gr/Uploads/Files/Material/courses/15_1105659031.pdf)]

2.5.6.2 *DBCLASD*

Πρόσφατα προτάθηκε ένας νέος αλγόριθμος χωρικής συσταδοποίησης που βασίζεται στον *DBSCAN* και ο οποίος ονομάζεται *DBCLASD (Distribution Based Clustering of Large Spatial Databases)* - συσταδοποίηση μεγάλων βάσεων χωρικών δεδομένων βασισμένη σε κατανομές. Ο αλγόριθμος *DBSCAN* δεν βρίσκει εφαρμογή αποκλειστικά σε χωρικά δεδομένα· προς τούτο δεν αναπτύσσεται σε αυτήν την ενότητα.

Ο αλγόριθμος *DBCLASD* υποθέτει ότι τα στοιχεία εντός μιας συστάδας είναι ομοιόμορφα κατανεμημένα και ότι σημεία εκτός της συστάδας πιθανόν δεν ικανοποιούν αυτόν τον περιορισμό. Βάσει αυτής της υπόθεσης, ο αλγόριθμος επιχειρεί να προσδιορίσει την κατανομή που ικανοποιείται από τις αποστάσεις μεταξύ πλησιέστερων γειτόνων. Όπως και με τον αλγόριθμο *DBSCAN*, δημιουργείται μια συστάδα γύρω από ένα στοιχείο - στόχο. Στοιχεία προστίθενται στη συστάδα, όσο το σύνολο των πλησιέστερων βάσει της απόστασης γειτόνων ικανοποιεί την υπόθεση της ομοιόμορφης κατανομής. Καθορίζονται τα υποψήφια στοιχεία και στη συνέχεια προστίθενται στην τρέχουσα συστάδα, αν ικανοποιούν ένα κριτήριο μέλους. Τα υποψήφια στοιχεία καθορίζονται με την εκτέλεση μιας ερώτησης περιοχής χρησιμοποιώντας έναν κύκλο ακτίνας *m*, ο οποίος έχει ως κέντρο ένα σημείο *p*, το οποίο μόλις προστέθηκε στη συστάδα. Η παράμετρος *m* επιλέγεται βάσει του ακόλουθου τύπου:

$$\int m \succ \sqrt{\frac{A}{\pi\left(1-\frac{1}{N^{\frac{1}{N}}}\right)}}$$

όπου *N* είναι ο αριθμός των σημείων στη συστάδα και A είναι η περιοχή της. Στη συνέχεια, τα σημεία που προστίθενται, γίνονται νέα υποψήφια.

Η περιοχή της συστάδας εκτιμάται με χρήση πλεγμάτων, τα οποία περιβάλλουν την συστάδα με ένα πολύγωνο. Όταν προστίθεται ένα σημείο σε μια συστάδα, το πλέγμα που περιέχει αυτό το σημείο προστίθεται στο πολύγωνο. Η εγγύτητα του πολυγώνου στο πραγματικό σχήμα της συστάδας εξαρτάται από το μέγεθος των πλεγμάτων. Εάν τα πλέγματα είναι πολύ μεγάλα, το σχήμα μπορεί να μην προσεγγίζει καλά την συστάδα. Εάν είναι πολύ μικρά, η συστάδα μπορεί στην πράξη να εκτιμηθεί από μη συνεκτικά πολύγωνα. Το μήκος του πλέγματος επιλέγεται να είναι η μεγαλύτερη τιμή στο σύνολο των πλησιέστερων βάσει της απόστασης.

Αφού το *test* χ^2 συνήθως απαιτεί τουλάχιστον 30 στοιχεία, γίνεται η υπόθεση ότι αρχικά προστίθενται 29 γειτονικά σημεία σε κάθε συστάδα [XEKS98]. Το τελευταίο βήμα επεκτείνει μια συστάδα βάσει της αναμενόμενης κατανομής του συνόλου των πλησιέστερων – βάσει της απόστασης – γειτόνων. Εάν εξακολουθεί να έχει την επιθυμητή κατανομή, τα σημεία στην γειτονιά αυτού του υποψηφίου προστίθενται στο σύνολο των υποψηφίων. Διαφορετικά, το υποψήφιο απορρίπτεται από το *C*. Αυτή η διαδικασία συνεχίζεται έως ότου αδειάσει το *C*. Τα σημεία στη γειτονιά ενός δοθέντος σημείου καθορίζονται βάσει της τιμής της ακτίνας που αναφέρθηκε προηγουμένως.

Σχετικές μελέτες απόδοσης δείχνουν ότι ο αλγόριθμος *DBCLASD* βρίσκει με επιτυχία συστάδες αυθαίρετων σχημάτων. Μόνο σημεία στα όρια των συστάδων ανατίθενται σε λάθος συστάδα.

2.5.6.3 Επεκτάσεις του *CLARANS*

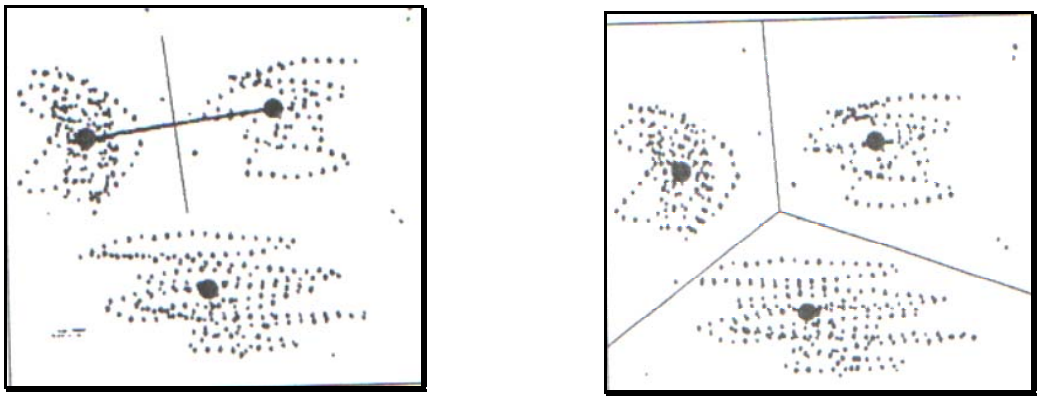
Έχουν προταθεί δύο προσεγγίσεις για την βελτίωση της απόδοσης του *CLARANS*, εκμεταλλευόμενοι τις δομές χωρικής δεικτοδότησης [EKX95].

Η πρώτη προσέγγιση χρησιμοποιεί έναν τύπο δειγματοληψίας βασισμένο στην δομή του *R** - δένδρου (μια παραλλαγή του *R* - δένδρου). Για την εγγύηση της

ποιότητας της δειγματοληψίας, χρησιμοποιείται το *R** - δένδρο για να εξασφαλίσει ότι εξετάζονται αντικείμενα από όλες τις περιοχές του χώρου. Κατά την αναζήτηση, το πιο κεντρικό αντικείμενο που βρίσκεται σε κάθε σελίδα του *R** - δένδρου, χρησιμοποιείται για την αναπαράσταση αυτής της σελίδας. Το πιο κεντρικό αντικείμενο είναι το αντικείμενο με την μικρότερη απόσταση από το κέντρο της σελίδας (από όλα τα αντικείμενα που είναι αποθηκευμένα σε αυτή τη σελίδα). Η σελίδα είναι στην πράξη το *MBR* που περιέχει όλα τα αντικείμενα σε αυτή τη σελίδα. Έτσι, το κέντρο αυτού του *MBR* μπορεί να ορισθεί ως το γεωμετρικό κέντρο του ορθογωνίου που το περικλείει. Ο *CLARANS* χρησιμοποιείται στη συνέχεια για να βρεθούν συστάδες για αυτά τα κεντρικά αντικείμενα. Οι *k* – μέσοι (*k - medoids*) που βρέθηκαν σε αυτό το βήμα αναπαριστούν τις *k* συστάδες που πρέπει να βρεθούν συνολικά για την βάση δεδομένων. Αφού το *R** - δένδρο συσταδοποιεί αντικείμενα που είναι χωρικά κοντά σε έναν κόμβο του δένδρου (και, επομένως, σελίδα δίσκου), είναι λογικό να πιστεύεται ότι αυτή η προσέγγιση στη δειγματοληψία βρίσκει καλούς μέσους.

Με την δεύτερη προσέγγιση, αντί να εξετάζεται όλη η βάση, εξετάζονται μόνο τα αντικείμενα στις δύο συστάδες που επηρεάζονται. Μια ερώτηση περιοχής μπορεί να χρησιμοποιηθεί για την ανάκτηση των απαραίτητων αντικειμένων. Μια αποδοτική τεχνική για την ανάκτηση μόνο των αντικειμένων σε μια δοθείσα συστάδα, βασίζεται στην κατασκευή ενός πολύεδρου γύρω από τον μέσο της συστάδας. Το πολύεδρο που κατασκευάζεται λέγεται πολύεδρο *Voronoi* ή διάγραμμα *Voronoi*.Αυτό το πολύεδρο δημιουργείται με την κατασκευή κάθετων διχοτόμων ανάμεσα σε ζεύγη από *medoids*. Η διαδικασία επεξηγείται στο Σχήμα 2.4.

Αυτό το πολύεδρο στη συνέχεια ορίζει την συστάδα. Τα αντικείμενα ενός διαγράμματος *Voronoi* είναι πιο κοντά στο μέσο του πολυέδρου που ανήκουν σε σχέση με οποιοδήποτε άλλο.



(α) Κάθετη διχοτόμος (β) Πολύεδρα *Voronoi*

Σχήμα 2.4: Πολύεδρο *Voronoi*

2.5.6.4 *SD(CLARANS)*

Ο αλγόριθμος *CLARANS* χωρικής τάξης (*spatial dominant - SD(CLARANS)*) υποθέτει ότι τα στοιχεία που πρόκειται να συσταδοποιηθούν περιέχουν χωρικές και μη χωρικές συνιστώσες. Πρώτα συσταδοποιεί τις χωρικές συνιστώσες χρησιμοποιώντας τον *CLARANS* και στη συνέχεια εξετάζει τα μη χωρικά γνωρίσματα εντός κάθε συστάδας για να εξάγει μια περιγραφή αυτής της συστάδας. Για παράδειγμα, η συσταδοποίηση της βλάστησης σε απομακρυσμένες περιοχές μπορεί να βρει ότι σε μια περιοχή (συστάδα) κυριαρχεί ένα δάσος από πεύκα, ενώ μία άλλη περιέχει μεγάλες ανοιχτές πεδιάδες και χορτολιβαδικές εκτάσεις. Ο *SD(CLARANS)* υποθέτει ότι κάποιο εργαλείο μάθησης, όπως το *DBLEARN* [HCC92] χρησιμοποιείται για να εξάγει την περιγραφή της συστάδας. Αυτή η περιγραφή μπορεί να θεωρηθεί ως μια γενικευμένη πλειάδα. Δηλαδή, χρησιμοποιώντας μια ιεραρχία εννοιών, οι τιμές των γνωρισμάτων για το σύνολο των πλειάδων σε μια συστάδα μπορούν να γενικευθούν, για να παρέχουν συνοπτικές τιμές σε ένα μεγαλύτερο επίπεδο στην ιεραρχία. Το εργαλείο μάθησης πραγματοποιεί αυτήν την πράξη. Ο αλγόριθμος *SD(CLARANS)* υποθέτει ότι στο πρώτο βήμα πραγματοποιείται ένα αρχικό φιλτράρισμα των δεδομένων, χρησιμοποιώντας μία σχέση βασισμένη στα μη χωρικά δεδομένα. Οποιοσδήποτε αλγόριθμος συσταδοποίησης θα μπορούσε να χρησιμοποιηθεί στην

θέση του *CLARANS* σε αυτόν τον αλγόριθμο. Στον αλγόριθμο που ακολουθεί, δίδεται ως είσοδος ο αριθμός των επιθυμητών συστάδων.

Σε αντίθεση με τον *SD(CLARANS)*, ο αλγόριθμος *CLARANS* μη χωρικής τάξης (*non – spatial dominant – NSD*(*CLARANS*)), εξετάζει πρώτα τα μη χωρικά γνωρίσματα. Εφαρμόζοντας μια γενίκευση σε αυτά τα γνωρίσματα, μπορεί να βρεθεί ένα σύνολο από αντιπροσωπευτικές πλειάδες, στο οποίο μια πλειάδα αναπαριστά κάθε συστάδα. Στη συνέχεια ο αλγόριθμος καθορίζει ποια χωρικά αντικείμενα ταιριάζουν με ποια αντιπροσωπευτική πλειάδα, για να ολοκληρώσει την διαδικασία συσταδοποίησης.

2.5.6.5 *BANG*

Η προσέγγιση *BANG* χρησιμοποιεί μια δομή πλέγματος όμοια με ένα $k - D$ δένδρο. Η δομή προσαρμόζεται στην κατανομή των στοιχείων, έτσι ώστε οι πιο πυκνές περιοχές να έχουν μεγαλύτερο αριθμό από μικρότερα πλέγματα, ενώ οι λιγότερο πυκνές να έχουν λίγα μεγάλα πλέγματα. Τα τελευταία στη συνέχεια ταξινομούνται βάσει της πυκνότητάς τους, που είναι ο αριθμός των στοιχείων στην περιοχή διαιρεμένος με το εμβαδόν. Βάσει του αριθμού των επιθυμητών συστάδων, αυτά τα πλέγματα με τις μεγαλύτερες πυκνότητες επιλέγονται ως τα κέντρα των συστάδων. Για κάθε επιλεγμένο πλέγμα, προστίθενται γειτονικά πλέγματα, όσο οι πυκνότητές τους είναι μικρότερες ή ίσες από αυτήν του κέντρου της τρέχουσας συστάδας.

2.5.6.6 *CLIQUE*

Ο αλγόριθμος *CLIQUE (CLustering In QUEst)* διαμερίζει έναν χώρο ν – διαστάσεων σε μη επικαλυπτόμενες ορθογωνικές μονάδες, ίσου μήκους. Ταυτοποιεί τις πυκνές περιοχές ως εξής:

- Αρχικά εντοπίζει τους υποχώρους που περιέχουν συστάδες, χρησιμοποιώντας την *Apriori* αρχή (εάν μία k – μονάδα είναι πυκνή, το ίδιο συμβαίνει και με τις προβολές της στον $(k - 1)$ – διάστατο χώρο).

- Εάν οι προβολές δεν είναι πυκνές, τότε η k – διάστατη μονάδα δεν είναι επίσης πυκνή.
- Γενικεύει τις υποψήφιες πυκνές περιοχές μέσω των $(k - 1)$ διαστάσεων προβολών αυτών.

Στη συνέχεια, για κάθε συστάδα, προσδιορίζει τις μέγιστες περιοχές οι οποίες καλύπτουν μία συστάδα συνδεόμενων πυκνών περιοχών.
[www.ted.unipi.gr/Uploads/Files/Material/courses/15_1105659031.pdf]

Ειδικότερα, ο εν λόγω αλγόριθμος ταυτοποιεί υποχώρους ενός χώρου υψηλών διαστάσεων οι οποίοι επιτρέπουν καλύτερη συσταδοποίηση από ότι ο αρχικός χώρος. Βασίζεται στην ταυτοποίηση των «αραιών» και «κοσμοβριθών» περιοχών στο χώρο, ανακαλύπτοντας το συνολικό πρότυπο κατανομής του συνόλου δεδομένων. Εάν το κλάσμα του συνόλου των δεδομένων σημείων που περιέχεται σε μία μονάδα χώρου υπερβαίνει την παράμετρο εισόδου, τότε η μονάδα αυτή χαρακτηρίζεται ως «πυκνή», ενώ η συστάδα ορίζεται ως το μέγιστο σύνολο συνδεόμενων πυκνών μονάδων εντός ενός υποχώρου.

2.5.6.7 *WaveCluster*

Η προσέγγιση συσταδοποίησης με χρήση μετασχηματισμού κυματιδίων (*WaveCluster)* για την παραγωγή χωρικών συστάδων εξετάζει τα δεδομένα σαν να ήταν σήματα. Όπως ο *STING*, έτσι και ο *WaveCluster* χρησιμοποιεί πλέγματα. Η πολυπλοκότητα παραγωγής συστάδων είναι $O(n)$ και δεν επηρεάζεται από ακραία σημεία. Αντίθετα με κάποιες προσεγγίσεις, ο *WaveCluster* μπορεί να βρει συστάδες τυχαίου σχήματος και δεν χρειάζεται να γνωρίζει τον επιθυμητό αριθμό από συστάδες. Ένα σύνολο από χωρικά αντικείμενα σε έναν ν – διάστατο χώρο θεωρούνται ως ένα σήμα. Τα όρια των συστάδων αντιστοιχούν στις υψηλές συχνότητες. Οι συστάδες από μόνες τους είναι χαμηλής συχνότητας με μεγάλο πλάτος. Μπορούν να χρησιμοποιηθούν τεχνικές επεξεργασίας σήματος για να βρουν τα χαμηλής συχνότητας τμήματα του χώρου. Προτείνεται η χρήση ενός μετασχηματισμού κυματιδίων (*wavelet transformation*) για να βρεθούν οι συστάδες. Ένας μετασχηματισμός κυματιδίων

χρησιμοποιείται ως φίλτρο για τον καθορισμό της αναλογίας συχνότητας του σήματος. Ένας μετασχηματισμός κυματιδίων ενός χωρικού αντικειμένου το αποσυνθέτει σε μια ιεραρχία από χωρικές εικόνες. Αυτές μπορούν να χρησιμοποιηθούν για κλιμάκωση μιας εικόνας σε διαφορετικά μεγέθη.

2.5.6.8 Προσέγγιση

Μόλις βρεθούν οι χωρικές συστάδες είναι επωφελές να προσδιορισθεί γιατί υπάρχουν οι συστάδες, με άλλα λόγια, ποια είναι τα μοναδικά χαρακτηριστικά των συστάδων. Για τον προσδιορισμό των χαρακτηριστικών των συστάδων, μπορεί να χρησιμοποιηθεί η έννοια της προσέγγισης (*approximation*). Αυτό γίνεται καθορίζοντας τα χαρακτηριστικά που είναι κοντά στις συστάδες. Οι συστάδες μπορούν να διακρίνονται βάσει χαρακτηριστικών μοναδικών σε αυτές ή κοινών σε πολλές συστάδες. Συνήθως γίνεται η υπόθεση ότι τα χαρακτηριστικά και οι συστάδες αναπαρίστανται από πιο πολύπλοκα κλειστά πολύγωνα παρά από απλά *MBR*.

Η συναθροιστική εγγύτητα (*aggregate proximity*) ορίζεται ως μέτρο του πόσο κοντά είναι μια συστάδα (ή ομάδα από στοιχεία) σε ένα χαρακτηριστικό (ή σε ένα αντικείμενο στο χώρο). Αυτό δεν είναι ένα μέτρο της απόστασης από τα όρια της συστάδας, αλλά μάλλον προς τα σημεία της συστάδας. Οι παραδοσιακές δομές δεδομένων, όπως τα R – δένδρα και τα k – D – δένδρα, δεν μπορούν να χρησιμοποιηθούν για την αποδοτική εύρεση αυτών των συσχετίσεων συναθροιστικής εγγύτητας, επειδή εστιάζουν σε ένα όριο συστάδας, αντί των αντικειμένων στην συστάδα. Η απόσταση συναθροιστικής εγγύτητας μπορεί να μετρηθεί από το άθροισμα των αποστάσεων σε όλα τα σημεία στη συστάδα.

Η σχέση συναθροιστικής εγγύτητας (*aggregate proximity relationship*) βρίσκει τα k κοντινότερα χαρακτηριστικά σε μία συστάδα. Ο αλγόριθμος *CRH* έχει προταθεί για την αναγνώριση αυτών των σχέσεων [KN96]. Το C αναπαριστά τον περικλείοντα κύκλο (*encompassing circle*), το R το ισοθετικό ορθογώνιο (*isothetic rectangle*) και το H το πολύγωνο (*convex hull*). Αυτά ορίζονται ως ακολούθως:

- **Ισοθετικό τρίγωνο**: Είναι το *MBR* που περιέχει ένα σύνολο σημείων, τέτοιο ώστε οι πλευρές του να είναι παράλληλες στους άξονες συντεταγμένων.
- **Περικλείων κύκλος**: Είναι ένας κύκλος που περιέχει ένα σύνολο σημείων και του οποίου η διάμετρος ισούται με την διαγώνιο του ισοθετικού τριγώνου.
- **Κυρτό περίβλημα**: Είναι το ελάχιστο περιβάλλον κλειστό σχήμα που περιέχει ένα σύνολο σημείων.

Αυτό που κάνει αυτά τα σχήματα αποδοτικά είναι ότι, δοθέντος ενός συνόλου από n σημεία, τα δύο πρώτα σχήματα μπορούν να βρεθούν σε $O(n)$ και το τελευταίο σε $O(nlogn)$. Αυτά τα γεωμετρικά σχήματα μπορούν να θεωρηθούν ως περιβάλλουσες δομές και μπορούν να χρησιμοποιηθούν ως πολλαπλά επίπεδα φιλτραρίσματος από τα πιθανά κοντινά χαρακτηριστικά. Ο στόχος είναι να εξασφαλισθεί ισορροπία ανάμεσα στην ακρίβεια και την αποδοτικότητα στην αναγνώριση των σχέσεων.

Το πρώτο βήμα του αλγορίθμου *CRH* είναι να εφαρμόσει τον περικλείοντα κύκλο. Τα χαρακτηριστικά (χρησιμοποιώντας την κυκλική προσέγγιση) που κατατάσσονται ως τα μεγαλύτερα (τα κοντινότερα) σε μια δοθείσα συστάδα στέλνονται στη συνέχεια στο φίλτρο στο επόμενο επίπεδο. Σε αυτό το επίπεδο το ισοθετικό ορθογώνιο χρησιμοποιείται για να αναπαραστήσει τα χαρακτηριστικά, τα οποία κατατάσσονται εκ νέου βάσει της εγγύτητας στην συστάδα. Τα υψηλότερα κατατασσόμενα χαρακτηριστικά σε αυτό το επίπεδο εξετάζονται στο τελικό επίπεδο, όπου χρησιμοποιείται ένα κυρτό περίβλημα για την εκτίμηση κάθε χαρακτηριστικού. Αυτή η προσέγγιση χρησιμοποιείται για κάθε συστάδα. Ο επιθυμητός αριθμός από γνωρίσματα που αναγνωρίζονται σε κάθε επίπεδο καθορίζεται ως είσοδος στον αλγόριθμο. Παρόλο που μπορούν να χρησιμοποιηθούν διαφορετικές τεχνικές για την κατάταξη των χαρακτηριστικών, συνήθως χρησιμοποιείται η τομή, ή υπολογίζονται οι πραγματικές αποστάσεις. Ο αλγόριθμος *CRH* χρησιμοποιεί διαφορετικά χαρακτηριστικά βελτιστοποίησης για την μείωση της συνολικής πολυπλοκότητας και της εξάλειψης επιπρόσθετου υπολογισμού των στιγμιοτύπων.

Κεφάλαιο 3

Συσταδοποίηση Χωρικών Δεδομένων

3.1 Αρχές Συσταδοποίησης

Η συσταδοποίηση (*clustering*) είναι παρόμοια με την κατηγοριοποίηση καθώς και στις δύο περιπτώσεις τα δεδομένα οργανώνονται σε ομάδες. Στην συσταδοποίηση, ωστόσο, σε αντίθεση με την κατηγοριοποίηση, οι ομάδες δεν είναι προκαθορισμένες. Η συσταδοποίηση επιτυγχάνεται βρίσκοντας ομοιότητες μεταξύ των δεδομένων βάσει των χαρακτηριστικών που υπάρχουν σε αυτά. Οι ομάδες αυτές ονομάζονται συστάδες (*clusters*). Κατά πολλούς, η συσταδοποίηση είναι μια ειδική μορφή κατηγοριοποίησης. Στα επόμενα οι δύο αυτές έννοιες λογίζονται διαφορετικές. Για τη συσταδοποίηση έχουν προταθεί πολλοί ορισμοί:

- Σύνολο όμοιων στοιχείων. Στοιχεία διαφορετικών συστάδων δεν είναι όμοια.
- Η απόσταση μεταξύ των σημείων κάποιας συστάδας είναι μικρότερη από την απόσταση μεταξύ ενός σημείου της συστάδας και οποιουδήποτε σημείου εκτός της συστάδας. [Dun04]
- Η διαδικασία ομαδοποίησης συνόλου απτών ή αφηρημένων αντικειμένων σε κατηγορίες παρεμφερών αντικειμένων. Μία συστάδα ορίζεται ως η συλλογή δεδομένων τα οποία είναι όμοια για την ίδια συστάδα και ανόμοια σε σχέση

με αντικείμενα διαφορετικών συστάδων. Μία συστάδα αντικειμένων / δεδομένων μπορεί να αντιμετωπισθεί συνολικά σαν μία ομάδα σε πολλές εφαρμογές. [HK01]

Το πρόβλημα της συσταδοποίησης ορίζεται αναλυτικά ως εξής: Γίνεται η υπόθεση ότι ο αριθμός των συστάδων που πρόκειται να δημιουργηθούν είναι μια τιμή εισόδου *k*. Το πραγματικό περιεχόμενο (και η ερμηνεία) κάθε συστάδας $K_j, 1 \le j \le k$ καθορίζεται ως αποτέλεσμα του ορισμού μιας συνάρτησης αντιστοίχισης. Χωρίς βλάβη της γενικότητας, θεωρείται ότι το αποτέλεσμα της επίλυσης ενός προβλήματος συσταδοποίησης είναι η δημιουργία ενός συνόλου συστάδων: $k = \{k_1, k_2, ..., k_k\}$.

ΟΡΙΣΜΟΣ: Δοθείσης μιας βάσης δεδομένων $D = \{ t_1, t_2, ..., t_n \}$ που αποτελείται από πλειάδες και μιας ακέραιας τιμής *k*, το πρόβλημα της συσταδοποίησης είναι να οριστεί μια αντιστοίχιση $f: D \rightarrow \{1, ..., k\}$ όπου κάθε t_1 ανατίθεται σε μία πλειάδα $K_j, 1 \le j \le k$. Μία ***συστάδα,*** *K.,* περιέχει ακριβώς εκείνες τις πλειάδες που της

ανατέθηκαν δηλαδή, $K_j = \{t_i \mid f(ti) = K_j, 1 \le i \le n, \text{και } t_i \in D\}$.

Τα βασικά χαρακτηριστικά της συσταδοποίησης που έρχονται σε αντίθεση με την κατηγοριοποίηση συνοψίζονται ως εξής:

- Ο (βέλτιστος) αριθμός συστάδων δεν είναι γνωστός.
- Μπορεί να μην υπάρχει καμία εκ των προτέρων γνώση σχετικά με τις συστάδες.
- Τα αποτελέσματα των συστάδων είναι δυναμικά.

[Dun04]

Η συσταδοποίηση συνιστά ένα πολλά υποσχόμενο ερευνητικό πεδίο της οποίας οι δυνητικές εφαρμογές θέτουν τις δικές τους απαιτήσεις. Τυπικές απαιτήσεις συσταδοποίησης στην εξόρυξη δεδομένων συνιστούν τα εξής:

- Εύρος κλίμακας αναφορικά με την διαχείριση στοιχείων στη βάση δεδομένων.

- Ικανότητα διαχείρισης διαφορετικών τύπων δεδομένων.
- Εύρεση κατηγοριών με ακανόνιστο σχήμα.
- Ικανότητα αντιμετώπισης «θορύβου».
- Έλλειψη ευαισθησίας αναφορικά με την σειρά εισαγωγής των στοιχείων
- Διαχείριση δεδομένων στο χώρο πολλών διαστάσεων
- Συσταδοποίηση υπό περιορισμούς
- Μεταφράσιμα και χρηστικά αποτελέσματα

[HK01]

3.2 Εφαρμογές Συσταδοποίησης

Οι βασικοί άξονες στους οποίους χρησιμοποιείται η συσταδοποίηση είναι οι εξής:

- Μείωση των δεδομένων: η ανάλυση συσταδοποίησης μπορεί να χρησιμοποιηθεί προκειμένου να διαμερίσει τα δεδομένα σε ένα σύνολο από «ενδιαφέρουσες συστάδες». Στη συνέχεια, αντί να γίνει διαχείριση των δεδομένων σαν μία ολότητα, υιοθετούνται αντιπροσωπευτικά δείγματα των συστάδων που έχουν καθορισθεί. Με αυτόν τον τρόπο επιτυγχάνεται συμπίεση των δεδομένων.
- Γενίκευση υποθέσεων: αφορά συμπερασμό σε σχέση με κάποιες υποθέσεις οι οποίες αφορούν στα δεδομένα.
- Δοκιμασία υποθέσεων: αφορά στην επιβεβαίωση της εγκυρότητας μιας συγκεκριμένης υπόθεσης.
- Πρόβλεψη που βασίζεται σε ομάδες: Οι συστάδες που προκύπτουν από την εφαρμογή της συσταδοποίησης σε ένα σύνολο δεδομένων, χαρακτηρίζονται από τα χαρακτηριστικά των προτύπων που ανήκουν σε αυτές. Εν συνεχεία, τα άγνωστα πρότυπα μπορούν να ταξινομηθούν σε συγκεκριμένες κατηγορίες, βάσει της ομοιότητάς τους με τα χαρακτηριστικά των συστάδων. Έτσι μπορεί να εξαχθεί χρήσιμη γνώση που σχετίζεται με τα δεδομένα.

Μερικές τυπικές εφαρμογές της συσταδοποίησης παρατηρούνται στα πεδία μάρκετινγκ και οικονομίας, ιατρικής, ανθρωπολογίας, βιολογίας (*taxonomy*), επεξεργασία εικόνας, ανάκτηση κειμένων. 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. Τέλος, σημαντικότατη εφαρμογή της συσταδοποίησης συνιστά η ανάλυση χωρικών δεδομένων. Λόγω του μεγάλου όγκου αυτών των δεδομένων, καθίσταται αντιοικονομική και δυσχερής η λεπτομερής εξέταση αυτών των δεδομένων από τους χρήστες. Η συσταδοποίηση βοηθά στην αυτοματοποίηση της διαδικασίας ανάλυσης και κατανόησης των χωρικών δεδομένων. Χρησιμοποιείται προκειμένου να ταυτοποιήσει και να εξάγει

ενδιαφέροντα χαρακτηριστικά και πρότυπα που ενδέχεται να υπάρχουν σε μεγάλες βάσεις χωρικών δεδομένων.

Όταν εφαρμόζεται συσταδοποίηση σε πραγματικές βάσεις δεδομένων, προκύπτουν ζητήματα όπως:

- Ο χειρισμός των *ακραίων σημείων* (*outliers*). Τα στοιχεία αυτά δεν ανήκουν στην πράξη σε καμία συστάδα· μπορούν να θεωρηθούν σαν μεμονωμένες συστάδες. Ωστόσο, αν ένας αλγόριθμος συσταδοποίησης επιχειρήσει να βρει μεγαλύτερες συστάδες, αυτά τα στοιχεία αναγκαστικά θα τοποθετούν σε κάποια ευρύτερη συστάδα. Καθώς αυτή η διαδικασία μπορεί να συνδυάσει δύο υπάρχουσες συστάδες και να αφήσει το απομονωμένο σημείο στη δική του συστάδα, μπορεί να οδηγήσει σε φτωχή συσταδοποίηση.
- Τα δυναμικά δεδομένα που υπάρχουν στη βάση δεδομένων υποδηλώνουν ότι η σύσταση των συστάδων μπορεί να αλλάξει στην πορεία του χρόνου.
- Η ερμηνεία της σημασιολογίας κάθε συστάδας ενδέχεται να είναι δύσκολη. Στην περίπτωση της κατηγοριοποίησης, η περιγραφή των κλάσεων είναι γνωστή εκ των προτέρων. Αυτό όμως δεν ισχύει στη συσταδοποίηση. Συνεπώς, όταν ολοκληρωθεί η διαδικασία συσταδοποίησης δημιουργώντας ένα σύνολο συστάδων, μπορεί να μην είναι προφανής η ακριβής σημασία της κάθε συστάδας. Στο σημείο αυτό χρειάζεται ένας ειδικός του πεδίου προκειμένου να αναθέσει «ετικέτες» (προσδιορισμούς) στις συστάδες.
- Δεν υπάρχει μία και μόνη σωστή λύση σε ένα πρόβλημα συσταδοποίησης. Στην πραγματικότητα, μπορούν να βρεθούν πολλές απαντήσεις. Το ακριβές πλήθος των συστάδων που απαιτούνται δεν είναι τόσο εύκολο να προσδιοριστεί.
- Ένα άλλο σχετικό θέμα είναι τι δεδομένα θα πρέπει να χρησιμοποιηθούν για τη συσταδοποίηση. Σε αντίθεση με τη μάθηση κατά τη διάρκεια της διαδικασίας κατηγοριοποίησης, όπου υπάρχει εκ των προτέρων κάποια γνώση σχετικά με το ποια πρέπει να είναι τα γνωρίσματα της κατηγοριοποίησης, στη συσταδοποίηση δεν υπάρχει επιβλεπόμενη μάθηση για να βοηθήσει τη

διαδικασία. Πράγματι, η συσταδοποίηση μπορεί να θεωρηθεί παρόμοια με τη μη επιβλεπόμενη μάθηση.

(www.dblab.aueb.gr/index.php/corporate/conbnt/download/232/868/file/HBV_SSDBMOI.pdf)

3.3 Μέθοδοι Συσταδοποίησης

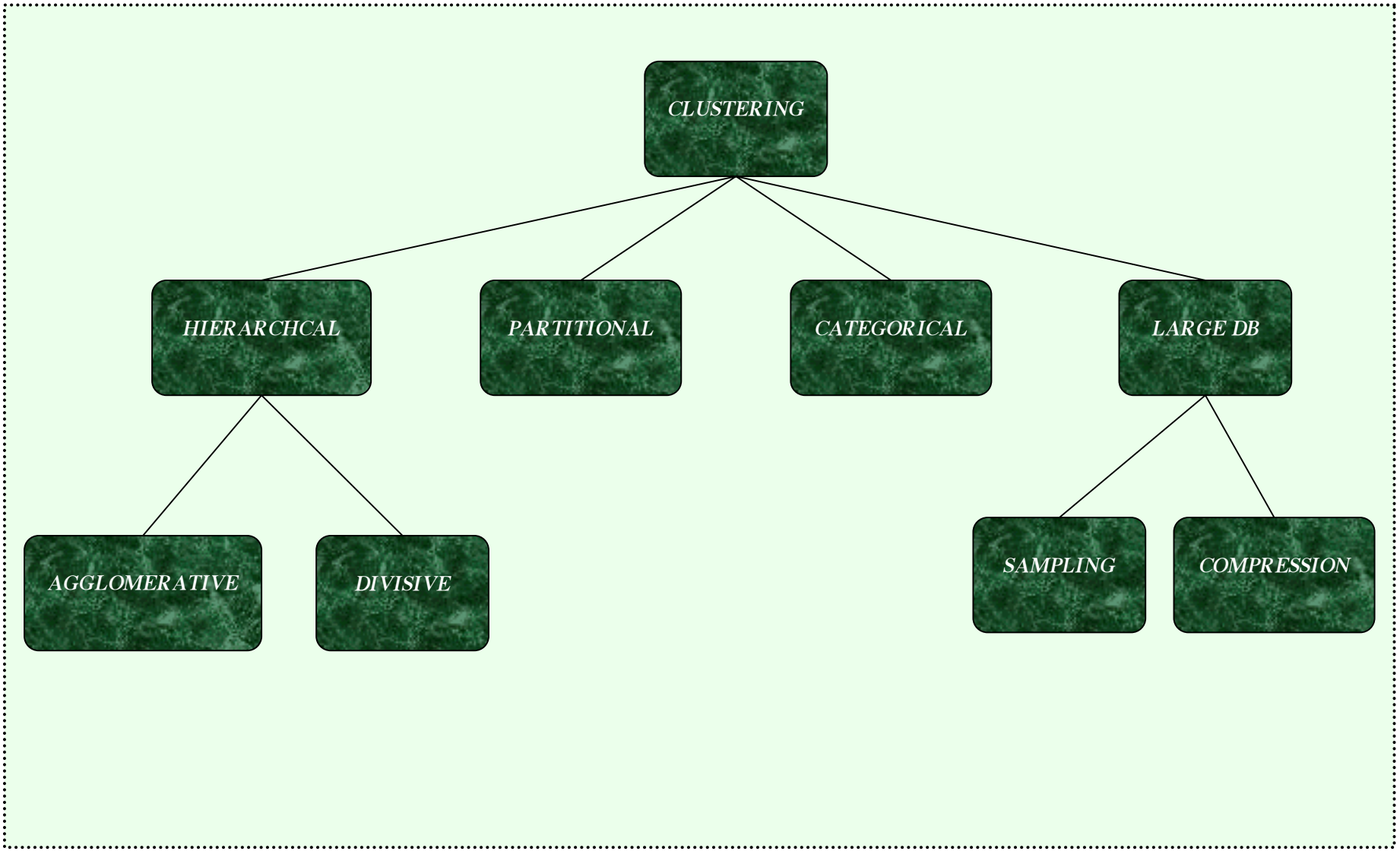
Από όσα αναπτύχθηκαν στην Ενότητα 3.2, γίνεται σαφής η σημασία της συσταδοποίησης στα πλαίσια της εξόρυξης (χωρικής) γνώσης. Αναφορικά με τους διαφορετικούς τύπους των αλγορίθμων συσταδοποίησης, ένας τρόπος κατηγοριοποίησής τους παρουσιάζεται στο Σχήμα 3.1.

Βάσει αυτής της κατηγοριοποίησης, στην **ιεραρχική συσταδοποίηση** (*hierarchical clustering*) δημιουργείται ένα εμφωλιασμένο σύνολο από συστάδες. Κάθε επίπεδο της ιεραρχίας έχει ένα ξεχωριστό σύνολο συστάδων. Στο κατώτατο επίπεδο, κάθε αντικείμενο βρίσκεται στη δική του συστάδα. Στο ανώτατο επίπεδο, όλα τα αντικείμενα ανήκουν στην ίδια συστάδα. Στην ιεραρχική συσταδοποίηση, ο επιθυμητός αριθμός των συστάδων δεν αποτελεί είσοδο. Στην **διαμεριστική συσταδοποίηση** (*partitional clustering*) ο αλγόριθμος δημιουργεί μόνο ένα σύνολο συστάδων. Οι προσεγγίσεις που ακολουθούνται στην περίπτωση αυτή χρησιμοποιούν τον επιθυμητό αριθμό συστάδων για να καθοδηγήσουν τη δημιουργία του τελικού συνόλου αυτών. Οι παραδοσιακοί αλγόριθμοι συσταδοποίησης προορίζονται για μικρές αριθμητικές βάσεις δεδομένων που χωράνε στη μνήμη. Υπάρχουν, ωστόσο, πιο πρόσφατοι αλγόριθμοι συσταδοποίησης που αφορούν σε μη αριθμητικά δεδομένα και προορίζονται για μεγαλύτερες, πιθανόν δυναμικές, βάσεις δεδομένων. Οι αλγόριθμοι που αφορούν σε μεγάλες βάσεις δεδομένων μπορούν να προσαρμοστούν στους εκάστοτε περιορισμούς μνήμης, είτε εφαρμόζοντας δειγματοληψία στη βάση δεδομένων, είτε χρησιμοποιώντας δομές δεδομένων που μπορούν να συμπιεστούν ή να υποστούν περικοπές προκείμενου να χωρέσουν στη μνήμη ανεξάρτητα από το μέγεθος της βάσης.

Οι αλγόριθμοι συσταδοποίησης διαφοροποιούνται επίσης και ως προς το εάν παράγουν επικαλυπτόμενες ή μη επικαλυπτόμενες συστάδες. Οι μη επικαλυπτόμενες συστάδες μπορούν να θεωρηθούν ως εξωγενείς (*extrinsic*) ή εγγενείς (*intrinsic*). Οι εξωγενείς τεχνικές χρησιμοποιούν ετικέτες πάνω στα στοιχεία για να βοηθήσουν την διαδικασία κατηγοριοποίησης. Περιλαμβάνουν

τους κλασσικούς αλγορίθμους επιβλεπόμενης μάθησης για την ταξινόμηση, οι οποίοι έχουν ως είσοδο ένα ειδικό σύνολο εκπαίδευσης. Οι εγγενείς (*intrinsic*) τεχνικές δεν χρησιμοποιούν κανέναν εκ των προτέρων προσδιορισμό των κατηγοριών, αλλά βασίζονται αποκλειστικά στην μήτρα γειτνίασης που περιέχει τις αποστάσεις μεταξύ των αντικειμένων.

Οι τύποι των αλγορίθμων συσταδοποίησης μπορούν να κατηγοριοποιηθούν περαιτέρω με βάση την τεχνική υλοποίησης που υιοθετούν. Οι ιεραρχικοί αλγόριθμοι μπορούν να διαιρεθούν σε **συσσωρευτικούς** (*agglomerative*) και **διαιρετικούς** (*divisive*). Στους συσσωρευτικούς αλγορίθμους οι συστάδες δημιουργούνται με διαδικασία από κάτω προς τα πάνω (*bottom - up*) ενώ στους διαιρετικούς με σχεδιασμό από πάνω προς τα κάτω (*top - down*). Παρά το γεγονός ότι τόσο οι ιεραρχικοί όσο και οι διαμεριστικοί αλγόριθμοι θα μπορούσαν να χωριστούν σε συσσωρευτικούς και διαιρετικούς, τυπικά αυτή η διάκριση σχετίζεται περισσότερο με τους ιεραρχικούς αλγορίθμους. Ένας άλλος περιγραφικός προσδιορισμός , η σειριακή (μερικές φορές, ονομάζεται και αυξητική) προσέγγιση, δείχνει αν γίνεται ξεχωριστός χειρισμός κάθε επιμέρους στοιχείου, ενώ η ταυτόχρονη προσέγγιση αν εξετάζονται μαζί όλα τα στοιχεία. Αν θεωρηθεί ότι κάθε πλειάδα έχει τιμές για όλα τα γνωρίσματα του σχήματος της βάσης, τότε θα διαφοροποιούνταν οι αλγόριθμοι συσταδοποίησης και ως προς το πώς εξετάζουν τις τιμές κάθε γνωρίσματος. Όπως συμβαίνει συνήθως με τις τεχνικές κατηγοριοποίησης με δένδρα αποφάσεων, οι μονοθετικοί (*monothetic*)) αλγόριθμοι εξετάζουν μία τιμή γνωρίσματος τη φορά. Αντιθέτως, οι πολυθετικοί (*polythetic*) αλγόριθμοι εξετάζουν όλες τις τιμές του γνωρίσματος μαζί. [Dun04]



Σχήμα 3.1: Κατηγοριοποίηση των Αλγορίθμων Συσταδοποίησης
Πηγή: [Dun04]

3.4 Συγκριτική Θεώρηση των Αλγορίθμων Συσταδοποίησης

Στον Πίνακα 1.2 της ενότητας 1.4 έγινε αναφορά σε ένα υποσύνολο της πληθώρας των αλγορίθμων συσταδοποίησης που έχουν προταθεί στην βιβλιογραφία και ακολουθούν την ως άνω κατηγοριοποίηση. Στον Πίνακα 3.1 παρουσιάζεται μια σύγκριση των αλγορίθμων συσταδοποίησης που παρουσιάσθηκαν στην ενότητα 1.4. Τα κριτήρια που χρησιμοποιούνται για τη σύγκριση είναι ο τύπος του αλγορίθμου, η πολυπλοκότητα χώρου, η πολυπλοκότητα χρόνου, και κάποιες παρατηρήσεις σχετικά με την καταλληλότητα του αλγορίθμου. [Dun04]

Οι τεχνικές του απλού συνδέσμου, πλήρους συνδέσμου και μέσου συνδέσμου είναι ιεραρχικές τεχνικές, συσσωρευτικές και παράλληλα διαμεριστικές. Δημιουργούν τις συστάδες με φορά από πάνω προς τα κάτω (*top - down*). Επίσης, οι τεχνικές αυτές υποθέτουν ότι υπάρχουν όλα τα δεδομένα ταυτόχρονα και συνεπώς δεν είναι αυξητικές. Υπάρχουν αρκετοί αλγόριθμοι που στηρίζονται στην δημιουργία ενός *MST* δένδρου, τόσο σε ιεραρχικές, όσο και διαμεριστικές εκδόσεις. Η πολυπλοκότητα τους είναι ίδια με αυτή των άλλων ιεραρχικών τεχνικών και, δεδομένου ότι βασίζονται στην δημιουργία ενός *MST*, οι αλγόριθμοι αυτοί δεν είναι αυξητικοί. Τόσο ο *k – means*, όσο και οι τεχνικές τετραγωνικού σφάλματος είναι επαναληπτικές τεχνικές και απαιτούν *O(tkn)* χρόνο. Η τεχνική πλησιέστερου γείτονα δεν είναι επαναληπτική, αλλά στην περίπτωση αυτή ο αριθμός των συστάδων δεν είναι προκαθορισμένος. Συνεπώς, η πολυπλοκότητα χειρότερης περίπτωσης μπορεί να είναι $O(n^2)$. Ο αλγόριθμος *BIRCH* φαίνεται να είναι αρκετά αποδοτικός,. Η πολυπλοκότητα χρόνου που δίνεται στον πίνακα αναφέρεται στην περίπτωση που το δένδρο δεν χρειάζεται να ανακατασκευαστεί. Ο αλγόριθμος *CURE* αποτελεί βελτίωση των ανωτέρω, επιτυγχάνει καλύτερη κλιμάκωση μέσω δειγματοληψίας και διαμερισμού και αναπαριστά μια συστάδα με πολλαπλά σημεία αντί ενός.

Η χρήση πολλαπλών σημείων επιτρέπει στη συγκεκριμένη προσέγγιση να εντοπίζει μη σφαιρικές συστάδες. ακραίων σημείων. Με τη δειγματοληψία, ο

ΑΛΓΟΡΙΘΜΟΣ	ΤΥΠΟΣ	ΧΩΡΟΣ	ΧΡΟΝΟΣ	ΣΧΟΛΙΑ
Απλού συνδέσμου	Ιεραρχικός	$O(n^2)$	$O(kn^2)$	Μη αυξητικός
Μέσου συνδέσμου	Ιεραρχικός	$O(n^2)$	$O(kn^2)$	Μη αυξητικός
Πλήρους συνδέσμου	Ιεραρχικός	$O(n^2)$	$O(kn^2)$	Μη αυξητικός
MST	Ιεραρχικός/ Διαμεριστικός	$O(n^2)$	$O(kn^2)$	Μη αυξητικός
Τετραγωνικού σφάλματος	Διαμεριστικός	$O(n)$	$O(tkn)$	Επαναληπτικός
K - means	Διαμεριστικός	$O(n)$	$O(tkn)$	Επαναληπτικός, αριθμητικά δεδομένα μόνο
Πλησιέστερου γείτονα	Διαμεριστικός	$O(n^2)$	$O(n^2)$	Επαναληπτικός
PAM	Διαμεριστικός	$O(n^2)$	$O(tk\text{-}(n\text{-}k)^2)$	Επαναληπτικός, προσαρμοσμένος, συσσωρευτικός, ακραία σημεία
BIRCH	Διαμεριστικός	$O(n)$	$O(n)$ (χωρίς ανακατασκευή)	*CF* - δένδρο, αυξητικός, ακραία σημεία
CURE	Μεικτός	$O(n)$	$O(n)$	Σωρός, $k-D$ δένδρο, αυξητικός, ακραία σημεία, δειγματοληψία
ROCK	Συσσωρευτικός	$O(n^2)$	$O(n^2 lgn)$	Δειγματοληψία, κατηγορικά δεδομένα, σύνδεσμοι
DBSCAN	Μεικτός	$O(n^2)$	$O(n^2)$	Δειγματοληψία, ακραία σημεία

Πίνακας 3.1: Σύγκριση των Αλγορίθμων Συσταδοποίησης

CURE επιτυγχάνει πολυπλοκότητα χρόνου $O(n)$. Ωστόσο, ο *CURE* δεν χειρίζεται αποδοτικά τα κατηγορικά δεδομένα. Αυτό βέβαια του επιτρέπει να είναι πιο ανθεκτικός στις αρνητικές συνέπειες των ακραίων σημείων.

Οι αλγόριθμοι k – *means* και *PAM* στηρίζονται στην επαναπροσδιορισμό των στοιχείων στις συστάδες, ο οποίος δεν οδηγεί πάντα στον εντοπισμό μιας καθολικά βέλτιστης ανάθεσης. Τα αποτελέσματα του k – *means* είναι αρκετά ευαίσθητα στην ύπαρξη ακραίων σημείων.

Ο αλγόριθμος *BIRCH* είναι ταυτόχρονα δυναμικός και κλιμακούμενος. Ωστόσο, εντοπίζει μόνο σφαιρικές συστάδες. Ο αλγόριθμος *DBSCAN* στηρίζεται στην πυκνότητα. Η πολυπλοκότητα χρόνου του *DBSCAN* μπορεί να βελτιωθεί στο *O(nlogn)* με κατάλληλα χωρικά ευρετήρια. Οι γενετικοί αλγόριθμοι δεν έχουν συμπεριληφθεί στον παραπάνω πίνακα επειδή η απόδοση τους εξαρτάται εξ' ολοκλήρου από την τεχνική που επιλέγεται για την αναπαράσταση των επιμέρους στοιχείων, από το πώς γίνεται η διασταύρωση και από το κριτήριο τερματισμού που χρησιμοποιείται.

Γενικά, κατηγοριοποίηση των μεθόδων συσταδοποίησης μπορεί να γίνει βάσει:

- Του τύπου των μεταβλητών που επιτρέπουν να συμμετέχουν στην βάση δεδομένων ενδιαφέροντος
- Του τρόπου που απεικονίζουν τις συστάδες
- Του τρόπου που οργανώνουν τις συστάδες (ιεραρχικά, σε επίπεδες λίστες, κλπ)
- Των αλγορίθμων που χρησιμοποιούν

Έτσι, ανάλογα με τον τύπο των μεταβλητών και την θεωρία που αποτελεί την βάση των αλγορίθμων που εφαρμόζονται για την συσταδοποίηση (*clustering*), διακρίνονται τα εξής είδη της τελευταίας:

- Στατιστική Συσταδοποίηση (*Statistical Clustering)*
- Εννοιολογική Συσταδοποίηση (*Conceptual Clustering)*
- Συσταδοποίηση Βασιζόμενη σε δίκτυο *Kohonen - Kohonen Net Clustering*
- Ασαφής Συσταδοποίηση - *Fuzzy Clustering*

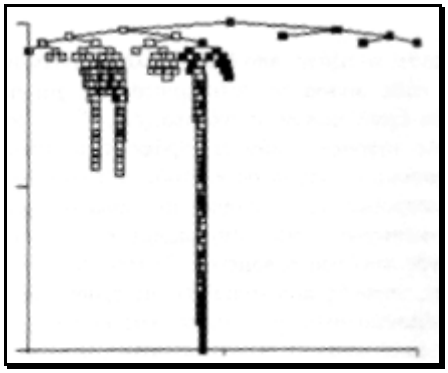
Περαιτέρω ανάλυση αυτών των ειδών συσταδοποίησης δεν κρίνεται σκόπιμη. Επεξήγηση αυτών παρατίθεται στο ΠΑΡΑΡΤΗΜΑ Π1.

Ανάλογα με τον τρόπο που οι αλγόριθμοι συσταδοποίησης οργανώνουν τις συστάδες, διακρίνονται τα εξής είδη συσταδοποίησης:

- *Hierarchical Clustering* (ιεραρχική)

- *Partitional Clustering* (διαμεριστική)
- *Density – based Clustering* (βασιζόμενη στην πυκνότητα)
- *Grid – based Clustering* (βασιζόμενη σε κάνναβο)

Στα δύο πρώτα είδη συσταδοποίησης έχει γίνει αναφορά στα προηγούμενα. Εν συντομία, η ιεραρχική συσταδοποίηση προχωρά διαδοχικά είτε συνδυάζοντας μικρότερες συστάδες σε μεγαλύτερες, ή διασπώντας μεγαλύτερες συστάδες. Οι μέθοδοι συσταδοποίησης διαφέρουν στον κανόνα με βάση τον οποίο αποφασίζεται ποιες από τις μικρότερες συστάδες θα συγχωνευτούν για την δημιουργία κάποιας μεγαλύτερης, ή ποια μεγάλη συστάδα θα διασπαστεί. Το τελικό αποτέλεσμα του αλγορίθμου είναι ένα δένδρο από συστάδες το οποί καλείται δενδρογράφημα (Σχήμα 3.2) και το οποίο παρουσιάζει πού οι συστάδες συνδέονται μεταξύ τους. Εάν το δενδρογράφημα κοπεί σε κάποιο επιθυμητό επίπεδο, προκύπτει η συσταδοποίηση των δεδομένων σε ομάδες μη σχετιζόμενες.



Σχήμα 3.2: Κατηγοριοποίηση των Αλγορίθμων Συσταδοποίησης
Πηγή: [BX02]

Η διαμεριστική συσταδοποίηση βασίζεται στην άμεση αποσύνθεση του συνόλου των δεδομένων σε ένα σύνολο μη σχετιζόμενων συστάδων. Η συνάρτηση που ο αλγόριθμος συσταδοποίησης προσπαθεί να ελαχιστοποιήσει, μπορεί να δώσει έμφαση στην τοπική δομή των δεδομένων, αναθέτοντας συστάδες στα άκρα της συνάρτησης (ελάχιστο, μέγιστο), ή στην γενική δομή των δεδομένων. Τυπικά το γενικό κριτήριο είναι η ελαχιστοποίηση κάποιων μέτρων ανομοιότητας μεταξύ των δειγμάτων μέσα σε κάθε μία από τις συστάδες, καθώς και η μεγιστοποίηση της ανομοιότητας μεταξύ διαφορετικών συστάδων. [BX02]

Η συσταδοποίηση η οποία βασίζεται στην πυκνότητα θεωρεί τις συστάδες ως πυκνές συλλογές αντικειμένων στο χώρο, οι οποίες διαχωρίζονται από περιοχές χαμηλής πυκνότητας (οι οποίες αντιπροσωπεύουν «θόρυβο»). Οι μέθοδοι αυτές χρησιμοποιούνται για να «φιλτράρουν» τα ακραία σημεία, τα οποία συνιστούν θόρυβο, και να ανακαλύψουν συστάδες αυθαίρετου σχήματος.

Η μέθοδος συσταδοποίησης η οποία βασίζεται σε κάνναβο ποσοτικοποιεί τον χώρο σε πεπερασμένο αριθμό κελιών, τα οποία διαμορφώνουν μία μορφή καννάβου, όπου εκτελούνται όλες οι λειτουργίες που σχετίζονται με αυτήν (την συσταδοποίηση). [HKT01]

Οι πίνακες 3.2, 3.3, 3.4 και 3.5 συνοψίζουν τις βασικές απόψεις και χαρακτηριστικά των πιο αντιπροσωπευτικών αλγορίθμων των παραπάνω κατηγοριών συσταδοποίησης. Η σύνοψη αυτή εστιάζει στα ακόλουθα χαρακτηριστικά των αλγορίθμων: i) τύπος δεδομένων που υποστηρίζει ο αλγόριθμος (αριθμητικά, κατηγορικά), ii) σχήμα συστάδων, iii) ικανότητα να διαχειρισθούν θόρυβο και ακραία σημεία, iv) κριτήριο συσταδοποίησης, v) πολυπλοκότητα. Επίσης παρουσιάζονται οι παράμετροι εισόδου των αλγορίθμων, ενώ μελετάται η επιρροή αυτών των παραμέτρων στα αποτελέσματα της συσταδοποίησης. Τέλος, περιγράφεται ο τύπος των αποτελεσμάτων της εκτέλεσης των αλγορίθμων, δηλαδή η πληροφορία που δίνει ένας αλγόριθμος, ώστε να αντιπροσωπεύσει τις ανακαλυφθείσες συστάδες στο σύνολο των δεδομένων.

Στα επόμενα η μελέτη εστιάζεται σε τρεις κυρίως αλγόριθμους συσταδοποίησης: Τον αλγόριθμο *DBSCAN*, ο οποίος συνιστά αλγόριθμο συσταδοποίησης βασιζόμενης στην πυκνότητα, τον *STING* και τον *K – means*.

ΣΗΜΑΝΤΙΚΟΙ ΔΙΑΜΕΡΙΣΤΙΚΟΙ ΑΛΓΟΡΙΘΜΟΙ ΣΥΣΤΑΔΟΠΟΙΗΣΗΣ							
ΟΝΟΜΑ	**ΤΥΠΟΣ ΔΕΔΟΜΕΝΩΝ**	**ΠΟΛΥΠΛΟΚΟΤΗΤΑ***	**ΓΕΩΜΕΤΡΙΑ**	**ΘΟΡΥΒΟΣ, ΑΚΡΑΙΑ ΣΗΜΕΙΑ**	**ΠΑΡΑΜΕΤΡΟΙ ΕΙΣΟΔΟΥ**	**ΑΠΟΤΕΛΕΣΜΑΤΑ**	**ΚΡΙΤΗΡΙΟ ΣΥΣΤΑΔΟΠΟΙΗΣΗΣ**
K - MEANS	Αριθμητικά	$O(n)$	Μη κυρτά σχήματα	Όχι	Αριθμός συστάδων	Κέντρο συστάδων	$\min_{v1,v2,\ldots,vk}(E_k)$ $E_k = \sum_{i=1}^{k}\sum_{k=i}^{N} d^2(x_k, v_i)$
PAM	Κατηγορικά	$O(k(n-k)^2)$	Μη κυρτά σχήματα	Όχι	Αριθμός συστάδων	Διάμεσος συστάδων	$\min(TC_{ih})$ $TC_{ih} = \Sigma_j C_{jih}$
CLARA	Αριθμητικά	$O(k(40+k)^2 + k(n-k))$	Μη κυρτά σχήματα	Όχι	Αριθμός συστάδων	Διάμεσος συστάδων	$\min(TC_{ih})$ $TC_{ih} = \Sigma_j C_{jih}$ (C_{jih} = the cost of replacing center i with h as far as Oj is concerned)
CLARANS	Αριθμητικά	$O(kn^2)$	Μη κυρτά σχήματα	Όχι	Αριθμός συστάδων, μέγιστος αριθμός εξεταζόμενων «γειτόνων»	Διάμεσος συστάδων	$\min(TC_{ih})$ $TC_{ih} = \Sigma_j C_{jih}$

Πίνακας 3.2: Τα κύρια χαρακτηριστικά των διαμεριστικών αλγόριθμων συσταδοποίησης
Πηγή: [www.dblab.aueb.gr/index.php/corporate/conbnt/download/232/868/file/HBV_SSDBMOI.pdf]

* *n* είναι ο αριθμός των σημείων του συνόλου δεδομένων και *k* αριθμός των ορισθεισών συστάδων

ΣΗΜΑΝΤΙΚΟΙ ΙΕΡΑΡΧΙΚΟΙ ΑΛΓΟΡΙΘΜΟΙ ΣΥΣΤΑΔΟΠΟΙΗΣΗΣ							
ΟΝΟΜΑ	**ΤΥΠΟΣ ΔΕΔΟΜΕΝΩΝ**	**ΠΟΛΥΠΛΟΚΟΤΗΤΑ***	**ΓΕΩΜΕΤΡΙΑ**	**ΘΟΡΥΒΟΣ, ΑΚΡΑΙΑ ΣΗΜΕΙΑ**	**ΠΑΡΑΜΕΤΡΟΙ ΕΙΣΟΔΟΥ**	**ΑΠΟΤΕΛΕΣΜΑΤΑ**	**ΚΡΙΤΗΡΙΟ ΣΥΣΤΑΔΟΠΟΙΗΣΗΣ**
BIRCH	Αριθμητικά	$O(n)$	Μη κυρτά σχήματα	Ναι	Ακτίνα συστάδων, *branching factor*	*CF = (αριθμός σημείων στην συστάδα N, γραμμικό άθροισμα των σημείων στην συστάδα LS, το τετραγωνικό άθροισμα N σημείων SS)*	Ενα σημείο ανατίθεται στον κοντινότερο κόμβο (συστάδα) βάσει ενός επιλεγέντος μέτρου απόστασης. Ακόμη, ο ορισμός των συστάδων βασίζεται στην απαίτηση ότι ο αριθμός των σημείων σε κάθε συστάδα πρέπει να ικανοποιεί ένα κατώφλι.
CURE	Αριθμητικά	$O(n^2 \log n)$, $O(n)$	Ακανόνιστα σχήματα	Ναι	Αριθμός συστάδων, μέλη αντιπροσώπων συστάδων	Ανάθεση τιμών δεδομένων σε συστάδες	Οι συστάδες με το κοντινότερο ζευγάρι αντιπροσώπων (καλά διεσπαρμένων σημείων) συνενώνονται σε κάθε βήμα.
ROCK	Κατηγορικά	$O(n^3 + n m_m m_a + n^2 \log n)$, $O(n^2, n m_m m_a)$	Ακανόνιστα σχήματα	Ναι	Αριθμός συστάδων	Ανάθεση τιμών δεδομένων σε συστάδες	$\max(E_l)$ $E_l = \sum_{i=1}^{k} n_i \sum_{p_q, p_r \in v_i} \frac{link(p_q, p_r)}{n_i^{(1+2f(\theta))}}$ - v_i center of cluster i - $link(p_q, p_r)$ = the number of common neighbors between p_q and p_r.

Πίνακας 3.3: Τα κύρια χαρακτηριστικά των ιεραρχικών αλγόριθμων συσταδοποίησης
Πηγή: [www.dblab.aueb.gr/index.php/corporate/conbnt/download/232/868/file/HBV_SSDBMOI.pdf]

* *n* είναι ο αριθμός των σημείων στο σύνολο των δεδομένων που λαμβάνεται υπόψη

ΣΗΜΑΝΤΙΚΟΙ ΒΑΣΙΖΟΜΕΝΟΙ ΣΤΗΝ ΠΥΚΝΟΤΗΤΑ ΑΛΓΟΡΙΘΜΟΙ ΣΥΣΤΑΔΟΠΟΙΗΣΗΣ							
ΟΝΟΜΑ	**ΤΥΠΟΣ ΔΕΔΟΜΕΝΩΝ**	**ΠΟΛΥΠΛΟΚΟΤΗΤΑ***	**ΓΕΩΜΕΤΡΙΑ**	**ΘΟΡΥΒΟΣ, ΑΚΡΑΙΑ ΣΗΜΕΙΑ**	**ΠΑΡΑΜΕΤΡΟΙ ΕΙΣΟΔΟΥ**	**ΑΠΟΤΕΛΕΣΜΑΤΑ**	**ΚΡΙΤΗΡΙΟ ΣΥΣΤΑΔΟΠΟΙΗΣΗΣ**
DBSCAN	Αριθμητικά	*O(logn)*	Ακανόνιστα σχήματα	Ναι	Ακτίνα συστάδας, ελάχιστος αριθμός αντικειμένων	Ανάθεση των τιμών των δεδομένων σε συστάδες	$f^{D}_{Gauss}(x^*) = \sum_{x_i \in near(x^*)} e^{\frac{d(x^*,x_i)^2}{2\sigma^2}}$ *x^* density attractor for a point x if $F_{Gauss} > \xi$ then x attached to the cluster belonging to x^*.*

Πίνακας 3.4: Τα κύρια χαρακτηριστικά των βασιζόμενων στην πυκνότητα αλγόριθμων συσταδοποίησης

Πηγή: [www.dblab.aueb.gr/index.php/corporate/conbnt/download/232/868/file/HBV_SSDBMOI.pdf]

* *n* είναι ο αριθμός των σημείων στο σύνολο των δεδομένων που λαμβάνεται υπόψη

ΣΗΜΑΝΤΙΚΟΙ ΒΑΣΙΖΟΜΕΝΟΙ ΣΕ ΚΑΝΝΑΒΟ ΑΛΓΟΡΙΘΜΟΙ ΣΥΣΤΑΔΟΠΟΙΗΣΗΣ							
ΟΝΟΜΑ	**ΤΥΠΟΣ ΔΕΔΟΜΕΝΩΝ**	**ΠΟΛΥΠΛΟΚΟΤΗΤΑ***	**ΓΕΩΜΕΤΡΙΑ**	**ΑΚΡΑΙΑ ΣΗΜΕΙΑ**	**ΠΑΡΑΜΕΤΡΟΙ ΕΙΣΟΔΟΥ**	**ΑΠΟΤΕΛΕΣΜΑΤΑ**	**ΚΡΙΤΗΡΙΟ ΣΥΣΤΑΔΟΠΟΙΗΣΗΣ**
WAVECLUSTER	Χωρικά δεδομένα	*O(n)*	Ακανόνιστα σχήματα	Ναι	Αντικείμενα σε συστάδες	Αποσύνθεση χαρακτηριστικών του χώρου με την εφαρμογή κυματοειδούς μετασχηματισμού	Ένα σημείο ανατίθεται στον κοντινότερο κόμβο (συστάδα) βάσει ενός επιλεγέντος μέτρου απόστασης. Ακόμη, ο ορισμός των συστάδων βασίζεται στην απαίτηση ότι ο αριθμός των σημείων σε κάθε συστάδα πρέπει να ικανοποιεί ένα κατώφλι.
STING	Αριθμητικά	*O(K), K* είναι ο αριθμός των συστάδων στο κατώτατο επίπεδο	Ακανόνιστα σχήματα	Ναι	Αριθμός αντικειμένων σε ένα κελί	Αντικείμενα σε συστάδες	Διαιρεί τον χώρο σε ορθογωνικά κελιά και χρησιμοποιεί ιεραρχική δομή. Κάθε κελί σε υψηλότερο επίπεδο διαμερίζεται σε έναν αριθμό μικρότερων κελιών στο επόμενο κατώτερο επίπεδο

Πίνακας 3.5: Τα κύρια χαρακτηριστικά των βασιζόμενων σε κάνναβο αλγόριθμων συσταδοποίησης

Πηγή: [www.dblab.aueb.gr/index.php/corporate/conbnt/download/232/868/file/HBV_SSDBMOI.pdf]

* *n* είναι ο αριθμός των σημείων στο σύνολο των δεδομένων που λαμβάνεται υπόψη

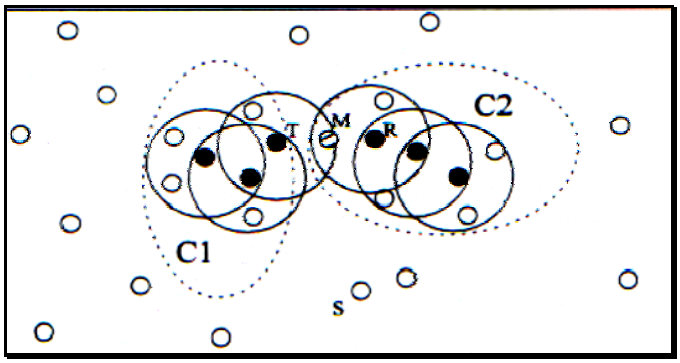
3.4.1 Αλγόριθμος *DBSCAN*

Ο *DBSCAN* είναι ένα αλγόριθμος ο οποίος βασίζεται στην πυκνότητα. Ο αλγόριθμος εξελίσσει περιοχές με ικανοποιητικά υψηλή πυκνότητα σε συστάδες και ανακαλύπτει συστάδες ακανόνιστου σχήματος σε χωρικές βάσεις δεδομένων. Απαιτεί την εισαγωγή δύο παραμέτρων, της ελάχιστης ακτίνας (*Eps)* και του ελάχιστου αριθμού σημείων (*MinPts).* Η γειτονιά εντός μίας ακτίνας *Eps* ενός δεδομένου αντικειμένου καλείται ***Eps – γειτονιά*** του αντικειμένου και ένα αντικείμενο με αριθμό αντικειμένων τουλάχιστον ίσο με *MinPts* εντός της *Eps* – γειτονιάς του καλείται **αντικείμενο – πυρήνας**. Η συσταδοποίηση που εκτελείται από τον αλγόριθμο *DBSCAN* ακολουθεί τους παρακάτω κανόνες:

- Ένα αντικείμενο μπορεί να ανήκει σε μία συστάδα, αν και μόνο αν κείται εντός της *Eps* – γειτονιάς κάποιων αντικειμένων – πυρήνα εντός της συστάδας.
- Ένα αντικείμενο – πυρήνας *o*, εντός της *Eps* – γειτονιάς ενός άλλου αντικειμένου – πυρήνα *p* οφείλει να ανήκει στην ίδια συστάδα με το *p*.
- Ένα αντικείμενο *q* το οποίο δεν συνιστά πυρήνα και κείται εντός της *Eps* – γειτονιάς ορισμένων αντικειμένων – πυρήνα p_1, p_2, ..., p_i, i >0, οφείλει να ανήκει στην ίδια συστάδα με τουλάχιστον ένα από τα αντικείμενα – πυρήνες p_1, p_2, ..., p_i.
- Ένα αντικείμενο το οποίο δεν συνιστά πυρήνα και δεν κείται εντός της *Eps* – γειτονιάς ενός οποιουδήποτε αντικειμένου *b*- πυρήνα λογίζεται θόρυβος. [HKT01]

Προκειμένου να ανακαλύψει συστάδες στα δεδομένα, ο *DBSCAN*, ελέγχει την *Eps* – γειτονιά κάθε σημείου στην βάση δεδομένων. Εάν η *Eps* – γειτονιά ενός σημείου *p* περιέχει περισσότερα από *MinPts*, δημιουργείται μία νέα συστάδα με πυρήνα το *p*. Τα αντικείμενα – πυρήνες εντός των νεοπροστιθέμενων μελών θα υποστούν την ίδια διαδικασία με το *p* , προκειμένου να εξελίξουν την συστάδα. Όταν σταματήσουν να εντοπίζονται πυρήνες με αυτήν την διαδικασία, ένα άλλο αντικείμενο πυρήνας θα ληφθεί από την βάση δεδομένων προκειμένου να εξελίξει μία νέα συστάδα. Σημειώνεται ότι κατά την διάρκεια της αυξητικής διαδικασίας,

ένα αντικείμενο – πυρήνας το οποίο ήδη ανήκει σε μία άλλη συστάδα μπορεί να απαντηθεί, κάτι το οποίο θα οδηγήσει στην συνένωση των δύο συστάδων. Η διαδικασία ολοκληρώνεται όταν κανένα νέο σημείο δεν μπορεί να προστεθεί στην συστάδα. Τα παραπάνω φαίνονται στο παράδειγμα που ακολουθεί (Σχήμα 3.3)



Σχήμα 3.3: Εύρεση δύο συστάδων από τον *DBSCAN*. Τα αντικείμενα – πυρήνες είναι τα συμπαγή σημεία, ενώ τα υπόλοιπα αντικείμενα είναι τα μη συμπαγή σημεία.

Πηγή: [HKT01]

Για μία δεδομένη γειτονιά *Eps* η οποία αντιπροσωπεύεται από την ακτίνα των κύκλων, και για $MinPts = 3$, είναι:

Οι δύο συστάδες C_1 και C_2 ορίζονται με την βοήθεια του αλγορίθμου *DBSCAN*. Τα αντικείμενα που ανήκουν είτε στην C_1, είτε στην C_2 κείνται εντός της *Eps* – γειτονιάς ενός τουλάχιστον πυρήνα είτε της μιας, είτε της άλλης από τις εξεταζόμενες συστάδες, ενώ δεν υπάρχουν δύο αντικείμενα – πυρήνες που να κείνται το ένα εντός της *Eps* – γειτονιάς του άλλου και να ανήκουν ταυτόχρονα σε διαφορετικές κατηγορίες. Ένα αντικείμενο το οποίο δεν συνιστά πυρήνα, όπως το *M* κείται εντός της *Eps* – γειτονιάς των *T* και *R*, τα οποία συνιστούν πυρήνες για τις συστάδες C_1 και C_2 αντιστοίχως. Συνεπώς, μπορεί να ανατεθεί σε μία από αυτές τις δύο συστάδες, δεδομένου ότι εντοπίζεται στο «σύνορο» αυτών. Τελικά το αντικείμενο *S* χαρακτηρίζεται θόρυβος, δεδομένου ότι δεν συνιστά πυρήνα και δεν βρίσκεται στην *Eps* – γειτονιά κάποιου αντικειμένου – πυρήνα. [HK03]

Ο αλγόριθμος *DBSCAN* μπορεί να εντοπίσει συστάδες με αυθαίρετα σχήματα. Παρά το γεγονός αυτό, όμως, παρουσιάζει ορισμένες αδυναμίες:

- Επηρεάζεται από τις τιμές των παραμέτρων *Eps* και *MinPts*, οι οποίες είναι δύσκολο να προσδιορισθούν.
- Στην περίπτωση που υπάρχει πυκνή σειρά σημείων που συνδέει δύο συστάδες, ο *DBSCAN* δεν μπορεί να τελειώσει συγχωνεύοντάς τες.
- Δεν χρησιμοποιεί κάποια μορφή προσυσταδοποίησης, αλλά εφαρμόζεται απευθείας στο σύνολο των δεδομένων, με αποτέλεσμα να καθίσταται ασύμφορος για μεγάλες βάσεις δεδομένων, λόγω του κόστους *I / O*.
- Η χρήση δείγματος προκειμένου να περιορισθεί το μέγεθος της εισόδου κατά την εφαρμογή των αλγορίθμων που βασίζονται στην πυκνότητα, δεν είναι εφικτή. Ο λόγος είναι ότι ακόμη και αν το δείγμα είναι μεγάλο, μπορεί να υπάρχουν μεγάλες διακυμάνσεις στην πυκνότητα των σημείων μέσα σε κάθε συστάδα στο τυχαίο δείγμα [BX02]

Ο αλγόριθμος *DBSCAN* μπορεί να περιγραφεί με την μορφή ψευδοκώδικα ως εξής:

Algorithm *DSCAN* (*D, Eps, MinPts*)

//Προϋπόθεση: όλα τα αντικείμενα σε σύνολο δεδομένων *D* δεν έχουν τοποθετηθεί σε *clusters*

FOR ALL objects o in D DO:

IF o δεν έχει ταξινομηθεί

Κάλεσε την συνάρτηση *expand_cluster* προκειμένου να κατασκευασθεί ένα *cluster* με ακτίνα *Eps* και ελάχιστο αριθμό στοιχείων *MinPts*, το οποίο θα περιέχει το *o*.

Function *expand_cluster* (*o, D, Eps, MinPts*):

Ανάκτηση της *Eps* – γειτονιάς $N_{EPS(o)}$ του *o*;

if $|N_{EPS(o)}| < MinPts$ //δηλαδή *o* δεν είναι αντικείμενο – πυρήνας

Σημείωσε το *o* σαν θόρυβο, *RETURN*;

else //το *o* είναι αντικείμενο – πυρήνας

Επίλεξε ένα νέο *cluster_id* και σημείωσε όλα τα αντικείμενα στο $N_{EPS(o)}$

με το τρέχον *cluster_id;*

ώθησε όλα τα αντικείμενα από το $N_{EPS(o)} - \{o\}$ στην στοίβα *seeds*;

while *not seeds.empty () do*

currentObject:=seed.top();

ανάκτηση της *Eps* – γειτονιάς του τρέχοντος αντικειμένου;

if $|N_{EPS(currentObject)}| \geq MinPts$

Επίλεξε όλα τα αντικείμενα $N_{EPS(currentObject)}$ που δεν έχουν ταξινομηθεί ακόμη, ή έχουν σημειωθεί ως θόρυβος,

τοποθέτησε τα μη ταξινομημένα αντικείμενα στην στοίβα *seeds* και

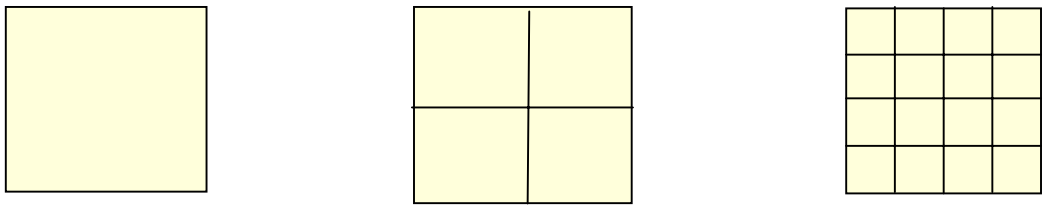
σημείωσε όλα τα αντικείμενα με το τρέχον *cluster_id;seeds.pop();*

RETURN

3.4.2 Αλγόριθμος *STING*

Η μέθοδος *STING* χρησιμοποιεί μια ιεραρχική τεχνική για την διαίρεση των χωρικών περιοχών σε ορθογώνια κελιά, παρόμοια με ένα τετραδικό δένδρο. Η βάση χωρικών δεδομένων σαρώνεται μία φορά και για κάθε κελί καθορίζονται στατιστικές παράμετροι (μέση τιμή, διασπορά, τύπος κατανομής). Κάθε κόμβος στη δομή πλέγματος συνοψίζει την πληροφορία για τα στοιχεία εντός της. Με τη λήψη αυτής της πληροφορίας μπορούν να απαντηθούν πολλά αιτήματα για εξόρυξη γνώσης από δεδομένα, συμπεριλαμβανομένης της συσταδοποίησης, εξετάζοντας τα στατιστικά που δημιουργήθηκαν για τα κελιά. Έτσι παράγονται μόνο συστάδες με κάθετα και οριζόντια όρια. Παρόλα αυτά, μπορεί να μη χρειάζεται να σαρωθεί ολόκληρη η βάση δεδομένων, αφού ληφθεί αυτή η στατιστική πληροφορία. Αυτό μπορεί να είναι πολύ αποδοτικό όταν γίνονται πολλαπλές αιτήσεις για εξόρυξη γνώσης από τα δεδομένα. Σε αντίθεση με τις τεχνικές γενίκευσης και προοδευτικής βελτίωσης, δεν πρέπει να δίδεται κάποια προκαθορισμένη εννοιολογική ιεραρχία.

Η προσέγγιση *STING* μπορεί να θεωρηθεί ως τεχνική ιεραρχικής συσταδοποίησης. Το πρώτο βήμα είναι η δημιουργία μιας ιεραρχικής αναπαράστασης (όπως ένα δενδρόγραμμα). Το δημιουργηθέν δένδρο διαδοχικά διαιρεί το χώρο σε τεταρτημόρια. Το κορυφαίο στοιχείο στην ιεραρχία αποτελείται από όλο το χώρο. Το κατώτερο επίπεδο έχει ένα φύλλο για καθένα από τα μικρότερα κελιά. Η αρχική πρόταση ήταν ένα κελί να έχει τέσσερα υποκελιά (πλέγματα) στο επόμενο κατώτερο επίπεδο. Η διαίρεση των κελιών είναι ίδια με αυτήν που εφαρμόζεται στα τετραδικά δένδρα. Γενικά, η προσέγγιση δουλεύει με οποιαδήποτε ιεραρχική διάσπαση του χώρου. Το Σχήμα 3.4 επεξηγεί τους κόμβους στα τρία πρώτα επίπεδα του δένδρου.



(α) Επίπεδο 1 (β) Επίπεδο 2 (γ) Επίπεδο 3

Σχήμα 3.4: Κόμβοι στην δομή *STING*
Πηγή: [Dun04]

Η διαδικασία δημιουργίας του δένδρου φαίνεται στον παρακάτω αλγόριθμο:

```
Input:
  D  // Data to be placed in the hierarchical structure
  k  //  Number of desired cells at the lowest level
Output
  T  //Tree
STING BUILD algorithm
  //  Create empty tree from top down.
  T=root node with data values initialized;  // Initially only root node
  i=1;
  repeat
    for each node in level i do
      create 4 children nodes with initial values;
    i=i+1
  until 4^i=k;
  //  Populate tree from bottom up.
  for each item in D do
    determine leaf node j associated with the position of D;
    update values of j based on attribute values in item;
  i: =log_4(k);
  repeat
    i: =i-1;
    for each node j in level i do
      update values of j based on attribute values
        in its 4 children;
  until i=1;
```

Κάθε κελί στο χώρο αντιστοιχεί σε έναν κόμβο του δένδρου και περιγράφεται τόσο από τα δεδομένα ανεξάρτητα των γνωρισμάτων (αριθμός), όσο και από τα δεδομένα εξαρτώμενα των γνωρισμάτων (μέση τιμή, τυπική απόκλιση, μέγιστο ελάχιστο, κατανομή). Καθώς τα δεδομένα φορτώνονται στη βάση, δημιουργείται η ιεραρχία. Η τοποθέτηση ενός στοιχείου σε ένα κελί καθορίζεται πλήρως από τη φυσική του θέση. Ο παραπάνω αλγόριθμος διαιρείται σε δύο μέρη. Το πρώτο μέρος δημιουργεί την ιεραρχία και το δεύτερο μέρος συμπληρώνει τις τιμές. Από τη στιγμή που ο αριθμός των κόμβων στο δένδρο είναι μικρότερος από τον αριθμό των στοιχείων στη βάση δεδομένων, η πολυπλοκότητα του *STING BUILD* είναι *O(n)*.

Ο παρακάτω αλγόριθμος αφορά στον τρόπο επεξεργασίας ερωτημάτων από τον *STING*.

ΑΛΓΟΡΙΘΜΟΣ 3

```
Input:
    T    // Tree
    q    // Query
Output:
    R    // Regions of relevant cells
STING algorithm:
    i=1
    repeat
        for each node in level i do
            determine if this cell is relevant to q and mark as such;
            i=i+1
    until all layers in the tree have been visited;
    identify neighboring cells of relevant cells to create
            regions of cells;
```

Ο αλγόριθμος υποθέτει ότι τίθεται μια ερώτηση *q* που μπορεί να απαντηθεί από την αποθηκευμένη στατιστική πληροφορία στο κατασκευασμένο δένδρο *T*. Μια τέτοια ερώτηση μπορεί να είναι η εύρεση του εύρους των τιμών ενός φαινομένου (ποιοτικό χαρακτηριστικό, πχ τιμές διαμερισμάτων) κοντά σε μια συγκεκριμένη περιοχή, έστω Κ (πχ στο κέντρο της πόλης). Θα πρέπει να καθορισθούν τα στατιστικά (*max* και *min*) των τιμών ενοικίασης των διαμερισμάτων για τα κατάλληλα κελιά. Το κελί στο οποίο βρίσκεται η περιοχή Κ θα καθορίζει τις πραγματικές τιμές για αυτά που βρίσκονται κοντά στο Κ. Επιπλέον, η ερώτηση θα μπορούσε να ανακτήσει τις πληροφορίες για τα κελιά που το περιβάλουν , ή ίσως για το επόμενο υψηλότερο επίπεδο στο δένδρο που περιέχει το κελί που βρίσκεται το Κ. Τα κοντινά κελιά θα μπορούσαν να προσδιορισθούν χρησιμοποιώντας κάποια συνάρτηση απόστασης. Το κρίσιμο σημείο εδώ είναι ότι πρέπει να προσδιορισθούν τα κατάλληλα κελιά και στη συνέχεια πρέπει να ανακτηθεί η πληροφορία από αυτά τα κελιά στο κατασκευασμένο δένδρο. Μια διάσχιση κατά πλάτος (*breadth - first*) χρησιμοποιείται για την εξέταση του δένδρου. Παρόλα αυτά, δεν εφαρμόζεται μια πλήρης διάσχιση του δένδρου. Εξετάζονται μόνο παιδιά σχετικών κόμβων. Εδώ η έννοια της σχέσης είναι περίπου ίδια με αυτή των ερωτήσεων ανάκτησης πληροφοριών, εκτός του ότι η σχέση καθορίζεται εκτιμώντας την αναλογία των αντικειμένων σε εκείνο το κελί που ικανοποιούν τις συνθήκες της ερώτησης. Η πολυπλοκότητα του αλγορίθμου *STING* είναι *O(k)*, όπου *k* είναι ο αριθμός των κελιών στο κατώτατο επίπεδο. Προφανώς αυτός είναι ο χώρος που λαμβάνεται από το ίδιο το δένδρο. Όταν χρησιμοποιείται για σκοπούς συσταδοποίησης, το *k* θα είναι ο μεγαλύτερος αριθμός από τις δημιουργηθείσες συστάδες.

Ο υπολογισμός της πιθανότητας ένα κελί να είναι σχετικό με μια ερώτηση βασίζεται στο ποσοστό των αντικειμένων στο κελί που ικανοποιούν τους περιορισμούς της ερώτησης. Χρησιμοποιώντας ένα προκαθορισμένο ασφαλές διάστημα, εάν το ποσοστό είναι αρκετά μεγάλο, τότε αυτό το κελί ορίζεται ως σχετικό. Η στατιστική πληροφορία που σχετίζεται με αυτά τα σχετικά κελιά χρησιμοποιείται για την απάντηση της ερώτησης. Εάν προσεγγιστική απάντηση

δεν είναι αρκετά καλή, τότε μπορεί τα συνδεόμενα σχετικά αντικείμενα στη βάση δεδομένων να μην χρειάζεται να εξεταστούν για να δώσουν μια πιο ακριβή απάντηση. [Dun04]

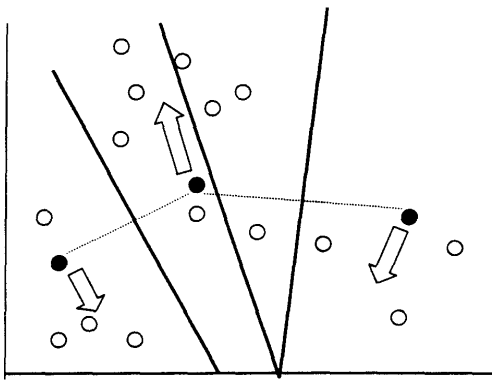
3.4.3 Αλγόριθμος *K – means*

Η μέθοδος *k – means* αποτελεί μία από τις πιο συχνά χρησιμοποιούμενες μεθόδους συσταδοποίησης [BL97]. Ανήκει στην κατηγορία της διαμεριστικής συσταδοποίησης, βασίζεται δηλαδή στην άμεση αποσύνθεση του συνόλου των δεδομένων σε ένα σύνολο ασυσχέτιστων συστάδων. Η αντικειμενική συνάρτηση την οποία προσπαθεί να ελαχιστοποιήσει ο αλγόριθμος είναι η μέση τετραγωνική απόσταση των δεδομένων από τα πλησιέστερα κέντρα των συστάδων.

$$E_K = \sum \left\| x_k -, m_{c(xk)} \right\|^2$$

όπου $c(x_k)$ είναι ο δείκτης του κέντρου το οποίο είναι πλησιέστερα στο x_k.

Ο βασικός αλγόριθμος για να ελαχιστοποιήσει την αντικειμενική συνάρτηση, αρχίζει θεωρώντας ένα σύνολο από *k* σημεία – δεδομένα ως τα κέντρα των *k* συστάδων (Σχήμα 3.5).



Σχήμα 3.5: Αρχικοποίηση *K - means*

Πηγή: [BX02]

Αν η σειρά των δεδομένων δεν έχει κάποια ιδιαίτερη σημασία, τότε λαμβάνονται τα πρώτα *k records*. Αλλιώς επιλέγονται σημεία αντιπροσωπευτικά για τις συστάδες, τα οποία απέχουν μεταξύ τους. Καθένα από τα κέντρα αντιπροσωπεύει μία συστάδα. Στο δεύτερο βήμα, κάθε σημείο αντιστοιχίζεται στην συστάδα της

οποίας το κέντρο βρίσκεται πιο κοντά. Στη συνέχεια υπολογίζονται τα νέα κέντρα των συστάδων με χρήση του μέσου όρου των σημείων τους. Για άλλη μια φορά αντιστοιχείται κάθε σημείο στην συστάδα της οποίας το κέντρο είναι πλησιέστερο. Η διαδικασία επαναλαμβάνεται συνεχώς έως ότου τα όρια των συστάδων παύουν να μεταβάλλονται, ή η συνάρτηση Ε δεν μεταβάλλεται σημαντικά. Ο αλγόριθμος *k – means* χρησιμοποιεί σταθερό και δοσμένο εξαρχής αριθμό συστάδων που θα δημιουργηθούν (όσα και τα κέντρα).

Παρακάτω περιγράφονται με την μορφή ψευδοκώδικα τα βασικά βήματα του αλγορίθμου *k – means*. Ο αλγόριθμος ξεκινά καθορίζοντας με τυχαίο τρόπο *c* κέντρα που θα αντιπροσωπεύουν οι *c* συστάδες. Στην συνέχεια προσδιορίζεται η απόσταση κάθε στοιχείου του συνόλου δεδομένων από το κέντρο κάθε συστάδας και κάθε στοιχείο τοποθετείται στην συστάδα από την οποία απέχει λιγότερο. Τα κέντρα των νέων συστάδων υπολογίζονται σαν ο μέσος όρος των στοιχείων που ανήκουν μέχρι στιγμής σε κάθε συστάδα. Η διαδικασία επαναλαμβάνεται μέχρις ότου οι συστάδες σταματήσουν να μεταβάλλονται. Αυτό σημαίνει ότι η απόκλιση μεταξύ των κέντρων των συστάδων που προέκυψαν τελευταία από αυτά της προηγούμενης επανάληψης είναι κοντά στο μηδέν (τα κέντρα ταυτίζονται).

Τα βήματα του αλγορίθμου σε μορφή ψευδοκώδικα είναι τα εξής:

- Εύρεση των αρχικών κέντρων, v_i, $I = 1, 2, ..., c$ για τις c συστάδες. Για κάθε επανάληψη $r = 1, ..., r_{max}$
- Υπολογισμός της απόστασης κάθε στοιχείου του συνόλου δεδομένων από το κέντρο κάθε συστάδας:

$$d_{k,i} = (x_k - v_i)^2,\ k = 1, 2, ..., \nu \text{ και } i = 1, 2, ..., c$$

- Κάθε στοιχείο x_k αντιστοιχίζεται στην συστάδα για την οποία ισχύει:

$$min_{k\,i}(d_{i,k}),\ \forall\ i, k$$

- Υπολογισμός των νέων κέντρων των συστάδων:

$$m_i^{r+1} = \frac{\sum_{i=1}^{n_i} x_k}{n_i}, \text{ όπου } n_i \text{ ο αριθμός των στοιχείων που ανήκουν στην } i \text{ συστάδα μέχρι στιγμής}$$

- *If* $\left\| m_i^{(r)} - m_i^{(r+1)} \right\| \prec \varepsilon$ *then*

  stop

  else

  $r = r+1$, *go to 2*

[BX02]

Ο αλγόριθμος *k – means* είναι σχετικά επιδεκτικός στην μεταβολή της κλίμακάς του και αποτελεσματικός στην επεξεργασία μεγάλων συνόλων δεδομένων, καθότι η υπολογιστική πολυπλοκότητά του είναι *O(nkt)*, όπου *n* είναι ο συνολικός αριθμός αντικειμένων, *k* ο αριθμός των συστάδων και *t* ο αριθμός των επαναλήψεων. Ισχύει $k \le n$ και $t \le n$. Η μέθοδος συνήθως ολοκληρώνεται σε ένα τοπικό μέγιστο.

Ο αλγόριθμος *k – means* είναι πολύ ευαίσθητος στον θόρυβο και τα απομακρυσμένα σημεία, δεδομένου ότι ένας μικρός αριθμός τέτοιων δεδομένων μπορούν να επηρεάσουν ουσιαστικά την μέση τιμή. [HKT01]

Προκειμένου ο αλγόριθμος *k – means* να γίνει επιδεκτικότερος στην μεταβολή της κλίμακάς του, κρίθηκε σκόπιμο να ταυτοποιηθούν τρία είδη περιοχών στα δεδομένα: περιοχές που είναι συμπιέσιμες, περιοχές που πρέπει να διατηρηθούν στην κύρια μνήμη και περιοχές προς απόρριψη, δηλαδή αγνοήσιμες. Ένα αντικείμενο είναι αγνοήσιμο εάν είναι εξακριβωμένα μέλος μιας συστάδας και συμπιέσιμο εάν δεν είναι αγνοήσιμο. Μία δομή δεδομένων γνωστή ως γνώρισμα συσταδοποίησης χρησιμοποιείται προκειμένου να συνοψίσει αντικείμενα τα οποία έχουν αγνοηθεί ή συμπιεσθεί. Εάν ένα αντικείμενο δεν είναι ούτε αγνοήσιμο ούτε συμπιέσιμο, θα πρέπει να διατηρηθεί στην κύρια μνήμη. Προκειμένου να επιτευχθεί η επιδεκτικότητα στην αλλαγή της κλίμακας, ο επαναληπτικός αλγόριθμος περιλαμβάνει μόνο εκείνα τα χαρακτηριστικά συσταδοποίησης των συμπιεζόμενων αντικειμένων, καθώς και τα αντικείμενα που πρέπει να διατηρηθούν στην κύρια μνήμη, μετατρέπει δηλαδή έναν αλγόριθμο κύριας μνήμης σε έναν αλγόριθμο δευτερεύουσας μνήμης. [HK03]

Κεφάλαιο 4
Πειραματική Αξιολόγηση

4.1 Εισαγωγή

Η υλοποίηση των τριών αλγορίθμων που αναπτύχθηκαν στις ενότητες 3.4.1 έως 3.4.3 γίνεται σε περιβάλλον *Visual Basic 6* και η πειραματική αξιολόγηση σε περιβάλλον *ARCGIS – ArcMap 8.3.* Προτού προχωρήσουμε στη ανάπτυξη αυτών των δύο θεμάτων, κρίνεται σκόπιμη μία περιγραφή των ως άνω πακέτων λογισμικού.

4.2 *Visual Basic*

Η *Visual Basic* συνιστά μία γλώσσα ανάπτυξης εφαρμογών σε περιβάλλον *Windows.* Η εργασία σε ένα τέτοιο περιβάλλον εμπεριέχει τις έννοιες «παράθυρο» (*window*), συμβάν (*event*) και μήνυμα (*message*).

4.2.1 Παραθυρικό Περιβάλλον

Το λειτουργικό σύστημα των *Microsoft Windows* διαχειρίζεται τα διάφορα είδη παραθύρων αναθέτοντας σε κάθε ένα έναν μοναδικό αριθμό (*id number*). Το σύστημα καταγράφει συνεχώς κάθε ένα από αυτά τα παράθυρα για συμβάντα.

Κάθε φορά που λαμβάνει χώρα ένα συμβάν, στέλνεται ένα μήνυμα στο λειτουργικό σύστημα. Το σύστημα επεξεργάζεται το μήνυμα και το αποστέλλει στα υπόλοιπα παράθυρα, τα οποία εκτελούν τις απαιτούμενες ενέργειες αναφορικά με το συγκεκριμένο μήνυμα.

Η αντιμετώπιση όλων των πιθανών συνδυασμών παραθύρων είναι χαοτική. Χάρις στη *Visual Basic*, πολλά από αυτά τα μηνύματα διαχειρίζονται αυτόματα ενώ κάποια άλλα λογίζονται ως διαδικασίες συμβάντων. Έτσι, ο χρήστης είναι σε θέση να δημιουργήσει ισχυρές εφαρμογές χωρίς να απαιτείται η ενασχόλησή του με ασήμαντες λεπτομέρειες.

4.2.2 Μοντέλο Συμβάντων

Στις παραδοσιακές ή διαδικαστικές εφαρμογές η ίδια η εφαρμογή καθορίζει ποια τμήματα του κώδικα θα εκτελεσθούν και με ποια σειρά. Η εκτέλεση ξεκινά με την πρώτη σειρά κώδικα και ακολουθεί ένα προκαθορισμένο μονοπάτι κατά μήκος της εφαρμογής το οποίο καλεί τις διαδικασίες που απαιτούνται κάθε φορά.

Σε μία εφαρμογή συμβάντων ο κώδικας δεν ακολουθεί ένα προκαθορισμένο μονοπάτι· εκτελεί διαφορετικά τμήματα κώδικα, ανάλογα με τα συμβάντα τα οποία πυροδοτούνται από τις ενέργειες του χρήστη, τα μηνύματα του συστήματος, άλλες εφαρμογές ή ακόμη και από την εφαρμογή καθεαυτή. Η ακολουθία αυτών των συμβάντων καθορίζει την ακολουθία με την οποία εκτελείται ο κώδικας, ούτως ειπείν το μονοπάτι κατά μήκος του κώδικα της εφαρμογής διαφοροποιείται κάθε φορά που «τρέχει» το πρόγραμμα.

Δεδομένου ότι η ακολουθία των συμβάντων δεν μπορεί να προβλεφθεί, ο κώδικας πρέπει να κάνει υποθέσεις όταν εκτελείται. Όταν γίνονται υποθέσεις (π.χ. ένα πεδίο εισόδου πρέπει να περιέχει μία τιμή πριν την εκτέλεση της διαδικασίας η οποία θα επεξεργασθεί την μεταβλητή), η εφαρμογή οφείλει να είναι δομημένη με τέτοιο τρόπο, ώστε να διασφαλίζεται ότι η υπόθεση είναι έγκυρη. Επίσης, ο κώδικας μπορεί να πυροδοτεί συμβάντα κατά την διάρκεια της εκτέλεσης. Για παράδειγμα, μεταβάλλοντας από την σκοπιά του προγραμματιστή το κείμενο σε

ένα πλαίσιο κειμένου, λαμβάνει χώρα το συμβάν αλλαγής αυτού του πλαισίου. Εάν είχε γίνει η υπόθεση ότι αυτό το συμβάν θα πυροδοτούνταν μέσω της αλληλεπίδρασης με τον χρήστη, θα υπήρχαν ανεπιθύμητα αποτελέσματα. Γίνεται σαφής λοιπόν η σημασία κατανόησης του μοντέλου συμβάντων κατά τον σχεδιασμό μιας εφαρμογής.

4.2.3 Διαδραστική Ανάπτυξη Κώδικα

Η ανάπτυξη εφαρμογών με τον παραδοσιακό τρόπο συνίσταται σε τρία διακριτά βήματα: γραφή, αποδελτίωση και δοκιμή κώδικα. Σε αντίθεση με τις παραδοσιακές γλώσσες, η *Visual Basic* χρησιμοποιεί μία διαδραστική προσέγγιση για την ανάπτυξη, *blurring* και διάκριση ανάμεσα στα τρία αυτά βήματα.

Η *Visual Basic* μεταφράζει τον κώδικα κατά την εισαγωγή του, εντοπίζοντας και επισημαίνοντας τα περισσότερα συντακτικά και ορθογραφικά λάθη *on the fly*. Επιπρόσθετα, αποδελτιώνει τον κώδικα κατά την εισαγωγή του. Υπάρχει μόνο μία μικρή καθυστέρηση πριν την δοκιμή της εφαρμογής. Εάν εντοπισθεί σφάλμα, αυτό επισημαίνεται. Το σφάλμα αυτό μπορεί να διορθωθεί και η αποδελτίωση να συνεχισθεί χωρίς να απαιτείται επανέναρξη της όλης διαδικασίας. Τέλος, λόγω της διαδραστικής φύσης της *Visual Basic*, η εφαρμογή μπορεί να εκτελεσθεί συχνά κατά την ανάπτυξή της. Έτσι μπορεί κανείς να ελέγχει παράλληλα τις επενέργειες του κώδικα. [RP99]

4.3 *ARCGIS – ArcMap8.3*

Το Σύστημα Γεωγραφικών Πληροφοριών *ArcGIS* διαχειρίζεται τις θέσεις των γεωγραφικών δεδομένων στο *ARC* περιβάλλον, ενώ τα περιγραφικά χαρακτηριστικά και τις μεταξύ τους σχέσεις στο *INFO* περιβάλλον. Ειδικότερα, σε ό,τι αφορά το λογισμικό *ArcGIS*, αυτό στηρίζεται στα παρακάτω χαρακτηριστικά:

- Εισαγωγή της χωρικής πληροφορίας και δημιουργία του ψηφιακού χάρτη: Αφού γίνει η εισαγωγή της γραφικής και της μη γραφικής (περιγραφικής) πληροφορίας, στη φάση της δημιουργίας του ψηφιακού χάρτη, το *ArcGIS* συγκεντρώνει τα γραφικά χαρακτηριστικά - που είναι σημεία και γραμμές, είτε αυτές αντιπροσωπεύουν γραμμικά χαρακτηριστικά, είτε όρια πολυγώνων, είτε πληροφορία σε μορφή ψηφιδωτού - τα δομεί, τα συσχετίζει με πίνακες, τα μετατρέπει ή όχι σε μορφή διανυσματική ή ψηφιδωτού και τα αποθηκεύει με τις αντίστοιχες μορφές.

- Διόρθωση και ενημέρωση του ψηφιακού χάρτη, ώστε ο χάρτης να ανταποκρίνεται στην πραγματικότητα
- Αποθήκευση του ψηφιακού χάρτη
- Διαχωρισμός του ψηφιακού χάρτη σε επίπεδα ομοιογενούς πληροφορίας και δημιουργία επικαλυπτόμενων ψηφιακών χαρτών διαφορετικών περιεχομένων
- Αναζήτηση χαρακτηριστικών (γραφικών και μη γραφικών, τοπολογικά δομημένων και μη δομημένων)
- Παρουσίαση πρωτογενών ή δευτερογενών χαρτών ή συνδυασμό αυτών στην οθόνη γραφικών
- Δημιουργία φύλλων χάρτη με τη χρήση *plotters* ή *printers*
- Επικοινωνία του πακέτου με άλλα πακέτα

Τα υποσυστήματα του *ArcGIS Desktop* είναι ο *ArcCatalog* , ο *ArcMap* και ο *ArcToolobox*. Πιο συγκεκριμένα, ο *ArcCatalog* παρέχει όλα τα εργαλεία για τη διαχείριση και οργάνωση της γεωγραφικής και περιγραφικής πληροφορίας, τόσο

τις γνωστές τεχνικές των *Coverage* και *Shapefiles* αλλά και με το χωρικό μοντέλο *Geodatabase*. Παρέχει ακόμη εργαλεία για την επισκόπηση των δεδομένων, καθώς και τη δυνατότητα για οργάνωση των μεταδεδομένων (*metadata*).

Ο *ArcMap* παρέχει όλα τα εργαλεία για να επιτελεστούν χαρτοσυνθέσεις, αναλύσεις, εισαγωγή δεδομένων, διόρθωση δεδομένων, δημιουργία αναφορών και γραφημάτων κλπ. Το περιβάλλον εργασίας στον *ArcMap* χωρίζεται σε τρεις ενότητες: το χώρο εμφάνισης των δεδομένων –καμβά, το χώρο του υπομνήματος και τον περιβάλλοντα χώρο με τις εργαλειοθήκες. Το περιβάλλον εργασίας μπορεί να μεταβληθεί και να τροποποιηθεί ανάλογα με τις ανάγκες και τις απαιτήσεις του χρήστη με την εισαγωγή ή αφαίρεση εργαλειοθηκών, ή ακόμη και τη δημιουργία νέων. Οι απαιτήσεις συστήματος για το εν λόγω λογισμικό είναι διαθέσιμες στη διεύθυνση www.esri.com.

Ο *ArcToolobox* τέλος δίνει τη δυνατότητα για τη διενέργεια γεωγραφικών επεξεργασιών μέσω διαχείρισης, ανάλυσης και μετατροπών δεδομένων, καθώς και με τη βοήθεια εργαλείων και οδηγών.[MDS03]

4.4 Δεδομένα

Προκειμένου για την αξιολόγηση των αποτελεσμάτων των υλοποιηθέντων αλγορίθμων χρησιμοποιήθηκαν συνθετικά δεδομένα. Αυτό συνέβη λόγω ακαταλληλότητας των πραγματικών δεδομένων που χορηγήθηκαν από την ΔΕΗ, η οποία οφειλόταν σε χρονικές και χωρικές ασυνέχειες, καθώς και σε ανισομερή κατανομή τους στην περιοχή ενδιαφέροντος.

Τα δεδομένα τα οποία θα εισαχθούν προκειμένου για την αξιολόγηση των αποτελεσμάτων των υλοποιηθέντων αλγορίθμων, αφορούν:

- Αφενός μεν για τον αλγόριθμο *STING*, σε θερμοκρασίες για δέκα χιλιάδες (10.000) θέσεις σε όλη την επιφάνεια της Πελοποννήσου. Τα δεδομένα αυτά παρήχθησαν με την μέθοδο *Kriging* ως εξής: έγινε επιφανειακή ολοκλήρωση σε πραγματικές σημειακές μετρήσεις θερμοκρασιών από είκοσι (20) σταθμούς με αποτέλεσμα την παραγωγή ενός ψηφιδωτού (*raster*). Σε αυτό επιλέγησαν δέκα χιλιάδες (10.000) τυχαία σημεία και για κάθε ένα από αυτά ελήφθη η τιμή του *pixel* που αντιστοιχεί στην τιμή της θερμοκρασίας.
- Αφετέρου, για τον αλγόριθμο *DBSCAN* και τον αλγόριθμο *K - means* τα δεδομένα αυτά αφορούν σε ατυχήματα κατά μήκος ενός τμήματος οδικού δικτύου. Πρόκειται επίσης για συνθετικά δεδομένα, τα οποία παρήχθησαν εν γένει με τη λήψη έξι χιλιάδων (6.000) τυχαία κατανεμημένων σημείων κατά μήκος δικτύου γραμμών. Τα δεδομένα αυτά κατετάγησαν σε κατηγορίες από 1 έως 4, όπου:
  - 1 = Πολύ σοβαρό
  - 2 = Σοβαρό
  - 3 = Μέτρια σοβαρό
  - 4 = Ελαφρύ

4.5 Υλοποίηση Αλγορίθμων

Η υλοποίηση των αλγορίθμων *STING* και *DBSCAN* έγινε σε περιβάλλον *Visual Basic 6*, ενώ ο *K – means* προέκυψε από αναζήτηση στην βιβλιογραφία [Πατ2005], επίσης σε *Visual Basic 6*. Ο κώδικας στο σύνολό του παρατίθεται στο ΠΑΡΑΡΤΗΜΑ Π2.

4.6.1. Αλγόριθμος *STING*

Ο αλγόριθμος *STING* αναπτύχθηκε στις ενότητες 2.3.6.1 και 3.4.2. Εν συντομία, ο στόχος του αλγορίθμου είναι η δημιουργία ενός δένδρου που να περιλαμβάνει σε κάθε κόμβο του βασικά στατιστικά στοιχεία (μέσος όρος, ελάχιστο και μέγιστο) για τα δεδομένα που αντιστοιχούν στις συντεταγμένες του (του κόμβου). Τα δεδομένα που χρησιμοποιεί ο αλγόριθμος, είναι αυτά που διαβάζει από το αρχείο εισόδου. Το αρχείο αυτό, περιλαμβάνει τις συντεταγμένες των στοιχείων (X, Y) και μία τιμή η οποία αφορά θερμοκρασίες.

Το δένδρο στη συγκεκριμένη υλοποίηση αντιστοιχίζεται σε έναν πίνακα του οποίου τα στοιχεία είναι οι κόμβοι του δένδρου. Για να μπορεί να αναπαριστά δένδρο, οι κόμβοι αυτοί θα πρέπει να γνωρίζουν τα παιδιά τους, ενώ για τη διευκόλυνση του αλγορίθμου, θα γνωρίζουν και τον πατέρα τους. Επίσης, κάθε κόμβος θα πρέπει να γνωρίζει τις συντεταγμένες σύμφωνα με τις οποίες ορίζεται η περιοχή του. Σύμφωνα με την πρακτική του αλγορίθμου, η περιοχή του είναι ορθογώνιου σχήματος και για να προσδιορισθεί και ακολούθως διαιρεθεί, χρησιμοποιούνται οι συντεταγμένες του άνω αριστερά (*Min X*, *Max Y* από τα δεδομένα) και του κάτω δεξιά σημείου (*Max X*, *Min Y* από τα δεδομένα), ορίζεται δηλαδή κάθε φορά η διαγώνιος από τις δοσμένες συντεταγμένες.

Τα στατιστικά στοιχεία αποτελούν το βασικό στόχο του αλγόριθμου, ενώ σε κάθε κόμβο – φύλλο του δένδρου, διατηρούνται και τα δεδομένα που αντιστοιχούν στις συντεταγμένες του ώστε να γίνεται ταχύτερα ο υπολογισμός των στατιστικών αποτελεσμάτων.Ο αλγόριθμος όπως υλοποιήθηκε, υποστηρίζει μέχρι και οκτώ (8) επίπεδα.

4.6.2. Αλγόριθμος *DBSCAN*

Ο αλγόριθμος *DBSCANs* αναπτύχθηκε στην ενότητα 3.4.1. Εν συντομία, στον αλγόριθμο αρχικά εισάγονται η ελάχιστη ακτίνα και ο ελάχιστος αριθμός σημείων. Ο αλγόριθμος εξελίσσει περιοχές με ικανοποιητικά υψηλή πυκνότητα σε συστάδες και ανακαλύπτει συστάδες ακανόνιστου σχήματος .

Τα δεδομένα που χρησιμοποιεί ο αλγόριθμος, είναι αυτά που διαβάζει από το αρχείο εισόδου. Το αρχείο αυτό, περιλαμβάνει τις συντεταγμένες των στοιχείων (X, Y) και μία τιμή η οποία αφορά στην σοβαρότητα των ατυχημάτων.

Το αποτέλεσμα είναι να ανατίθεται κάθε μία εγγραφή σε συστάδα ή θόρυβο.

4.6.3. Αλγόριθμος *K – means*

Ο αλγόριθμος *K – means* αναπτύχθηκε στην ενότητα 3.4.3. Εν συντομία, στον αλγόριθμο αρχικά εισάγονται ο επιθυμητός αριθμός των συστάδων *k* και μία βάση δεδομένων με ένα συγκεκριμένο πλήθος στοιχείων. Ο αλγόριθμος ακανόνιστα και τυχαία επιλέγει *k* κέντρα ως την αρχική λύση και αναθέτει κάθε αντικείμενο στο πιο κοντινό του κέντρο μορφώνοντας ένα σύνολο συστάδων. Ακολούθως γίνεται υπολογισμός του μέσου για κάθε συστάδα, έως ότου ικανοποιηθεί ένα κριτήριο σύγκλισης.

Τα δεδομένα που χρησιμοποιεί ο αλγόριθμος, είναι αυτά που διαβάζει από το αρχείο εισόδου. Το αρχείο αυτό, περιλαμβάνει τις συντεταγμένες των στοιχείων (Χ, Υ) και μία τιμή η οποία αφορά στην σοβαρότητα των ατυχημάτων.

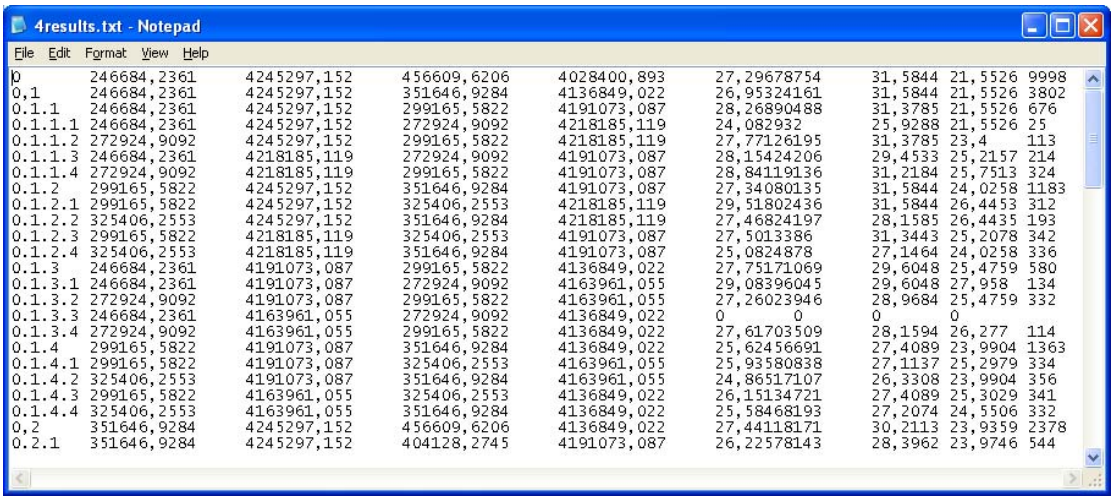
Τα στατιστικά στοιχεία που προκύπτουν είναι οι μέσοι για κάθε μία από τις τελικές συστάδες.

4.6 Οπτικοποίηση Αλγορίθμων

Προκειμένου για την αξιολόγηση των υπό μελέτη αλγορίθμων, είναι θεμιτή η οπτικοποίησή τους, η οποία πραγματοποιείται σε περιβάλλον *ARCGIS – ArcMap*.

4.6.1. Αλγόριθμος *STING*

Τα αποτελέσματα της συσταδοποίησης έστω για αριθμό επιπέδων ίσο προς τέσσερα (4) που εκτελεί ο αλγόριθμος *STING* παρατίθενται στο ΠΑΡΑΡΤΗΜΑ Π3. Πρόκειται για ένα *.txt* αρχείο, το οποίο αποσπασματικά φαίνεται στο Σχήμα 4.1:

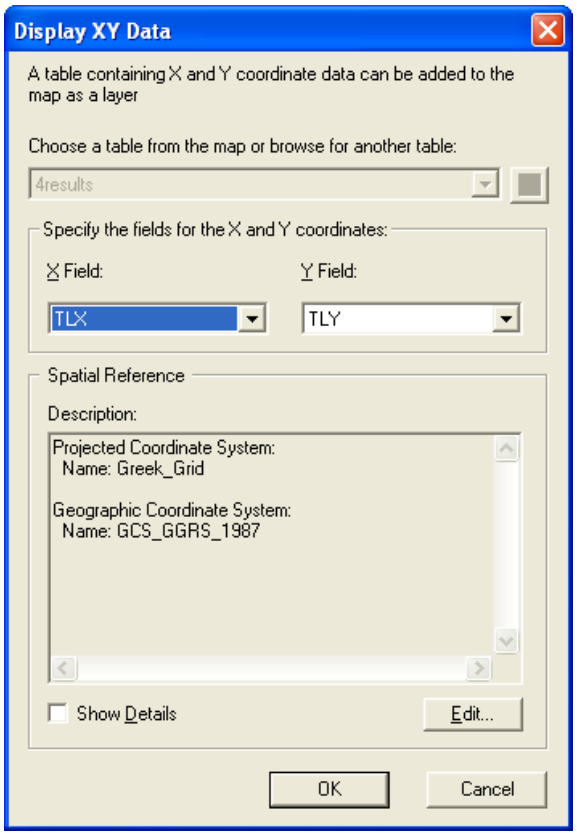


Σχήμα 4.1: Αποτελέσματα Συσταδοποίησης Αλγορίθμου *STING* για αριθμό επιπέδων ίσο προς 4.

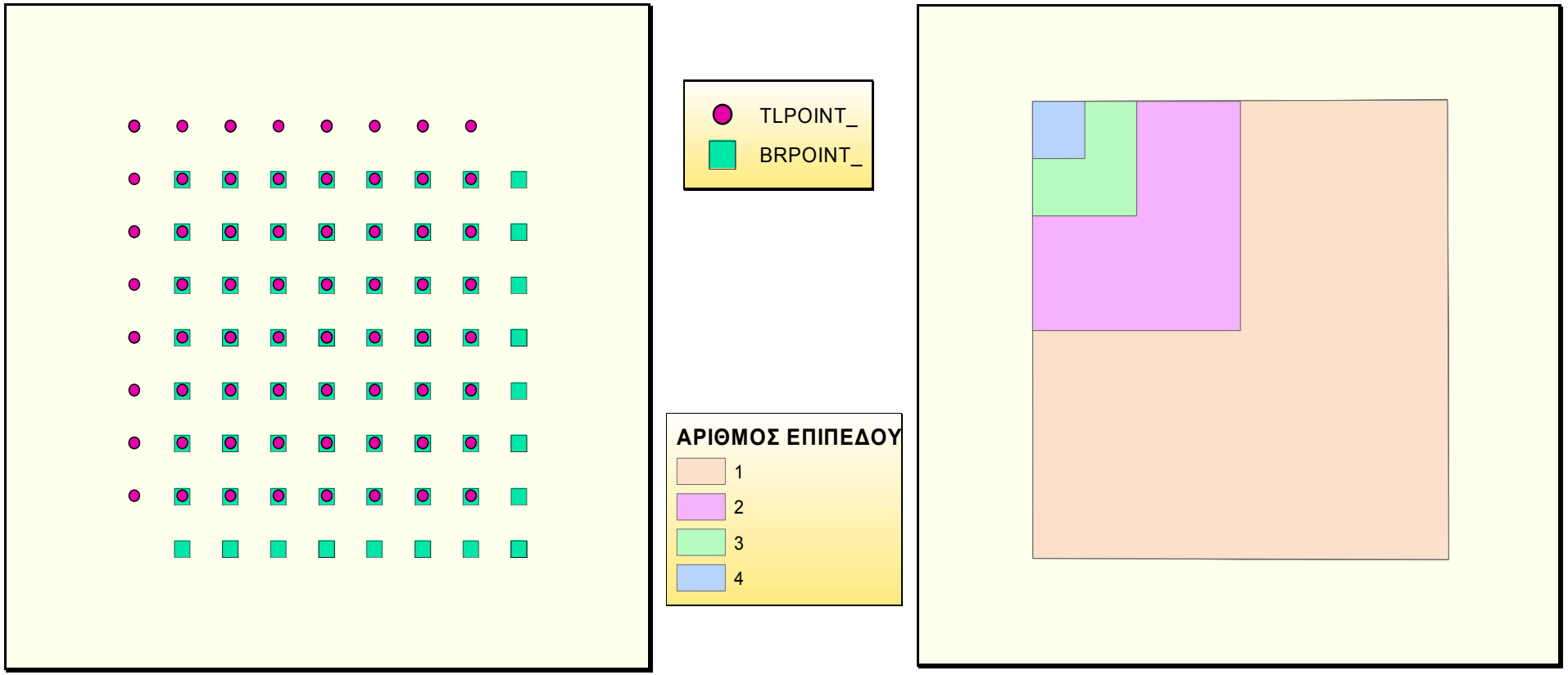
Στο αρχείο αυτό διακρίνονται εννέα στήλες, οι οποίες αντιστοιχούν στον κόμβο, το X του άνω αριστερά σημείου, το Y του άνω αριστερά σημείου, το X του κάτω δεξιά σημείου, το Y του κάτω δεξιά σημείου, τον μέσο όρο των θερμοκρασιών για το εκάστοτε κελί που ορίζεται από τις συντεταγμένες του πάνω αριστερά και του κάτω δεξιά σημείου, την μέγιστη τιμή αυτών των θερμοκρασιών, την ελάχιστη τιμή τους, καθώς και τον αριθμό των στοιχείων που περιέχονται σε κάθε κελί.

Το αρχείο εισάγεται στη συνέχεια σε περιβάλλον *Microsoft Excel* και μετατρέπεται σε *.dbf*, το *4results.dbf*, προκειμένου να οπτικοποιηθεί και να αξιολογηθεί σε περιβάλλον *ArcMap*.

Με *Add Data* γίνεται η προσθήκη του *4results.dbf* σε περιβάλλον *ArcMap*. Προκειμένου να απεικονισθεί η ορθογωνική διαίρεση του χώρου σε τεταρτημόρια ακολουθείται η εξής διαδικασία: με δεξί κλικ στον εισαχθέντα πίνακα > *Display XY Data*, εμφανίζεται το παρακάτω παράθυρο διαλόγου, το οποίο συμπληρώνεται με τις συντεταγμένες (*Top Left X – TLX*, *Top Left Y - TLY*) για το πάνω αριστερά σημείο:

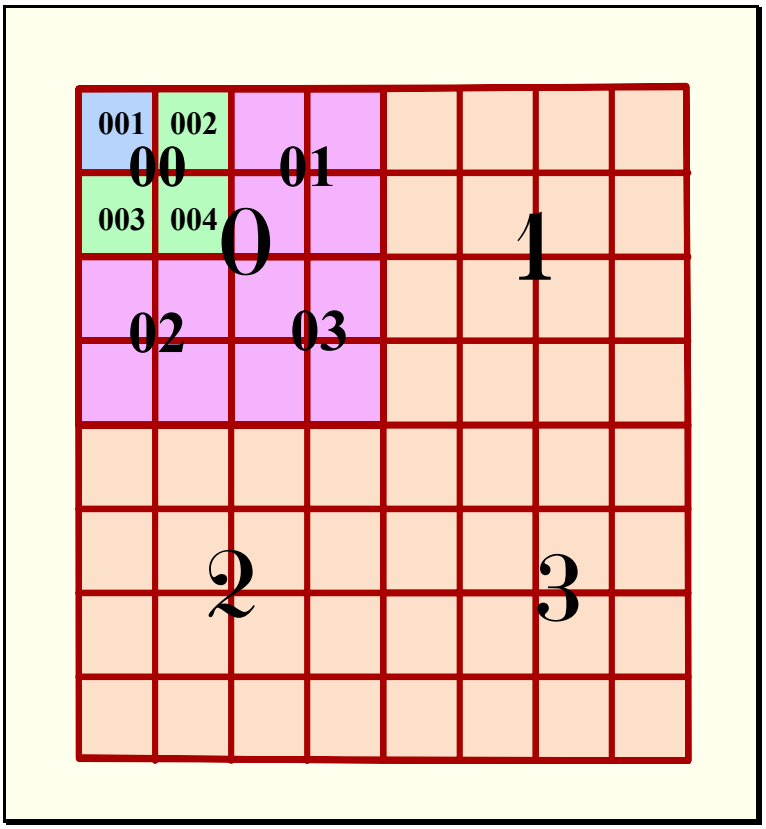


Η διαδικασία επαναλαμβάνεται για το κάτω δεξιά σημείο, εισάγοντας τις αντίστοιχες συντεταγμένες (*Bottom Right X – BRX*, *Bottom Right Y - BRY*). Στη συνέχεια, με δεξί κλικ στα προκύπτοντα επίπεδα και επιλέγοντας *Data > Export Data*, δημιουργούνται δύο αντίστοιχα *shapefiles*, τα *TLPOINT* και *BRPOINT*, με τα στοιχεία του αρχικού πίνακα. Οπτικά, το αποτέλεσμα φαίνεται στο Σχήμα 4.2:



Σχήμα 4.2: Διαδικασία διαίρεσης του χώρου σε τεταρτημόρια βάσει του αλγορίθμου *STING*.

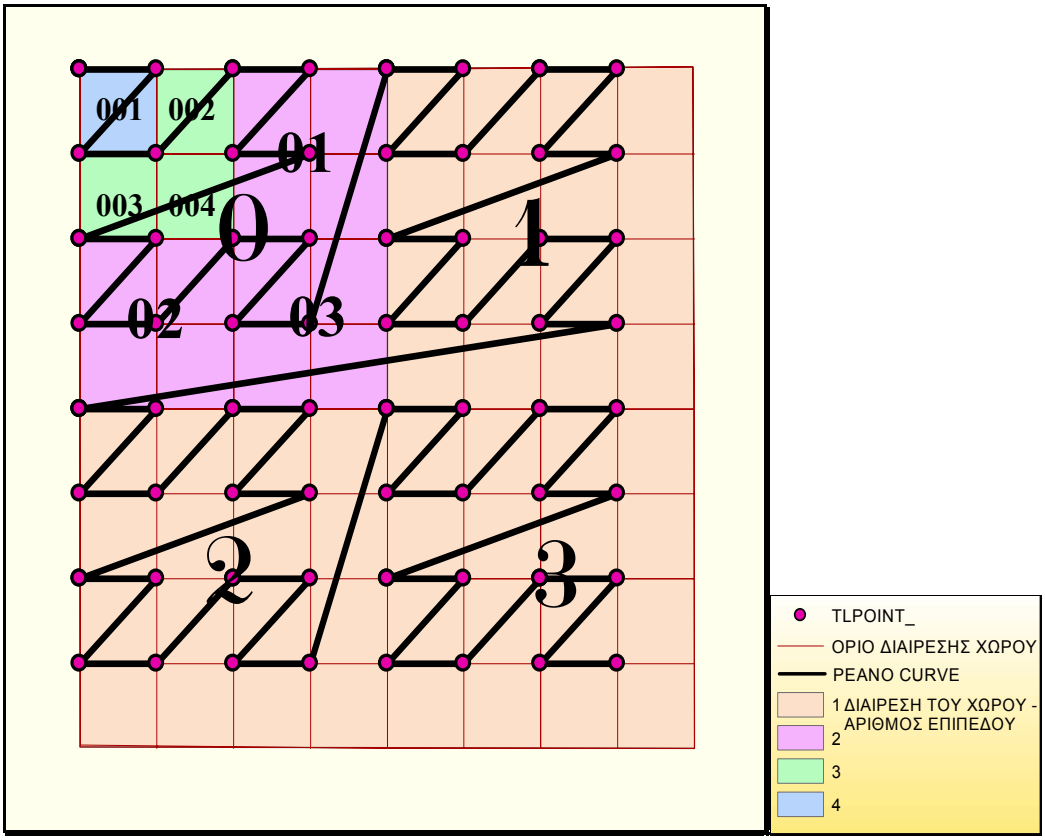
Με διαδικασία *editing* (επιλογή *Editor > Start Editing >…*) πραγματοποιείται η διαίρεση του χώρου, η οποία απεικονίζεται στο Σχήμα 4.3 συνολικά:



Σχήμα 4.3: Συνολική ιεραρχική διαίρεση του χώρου βάσει του αλγορίθμου *STING*.

Κάτι το οποίο παρουσιάζει ενδιαφέρον, είναι η μορφή της πολυγραμμής που προκύπτει από τις συντεταγμένες των «πάνω αριστερών σημείων» του θεματικού επιπέδου *TLPOINT* μέσω της διαδικασίας *XTools > Feature Conversions >Make One Polyline From Points*. Πρόκειται για την γραμμή Peano σχήματος Z (*Z – shaped Peano curve*). Κάτι τέτοιο ήταν αναμενόμενο, δεδομένης της αντιστοιχίας της δομής τετραδικού δένδρου και της εν λόγω γραμμής εν γένει, καθώς και του πίνακα Morton (*Morton matrix indexing scheme*). (Κα62004)

Τα παραπάνω φαίνονται στο Σχήμα 4.4:

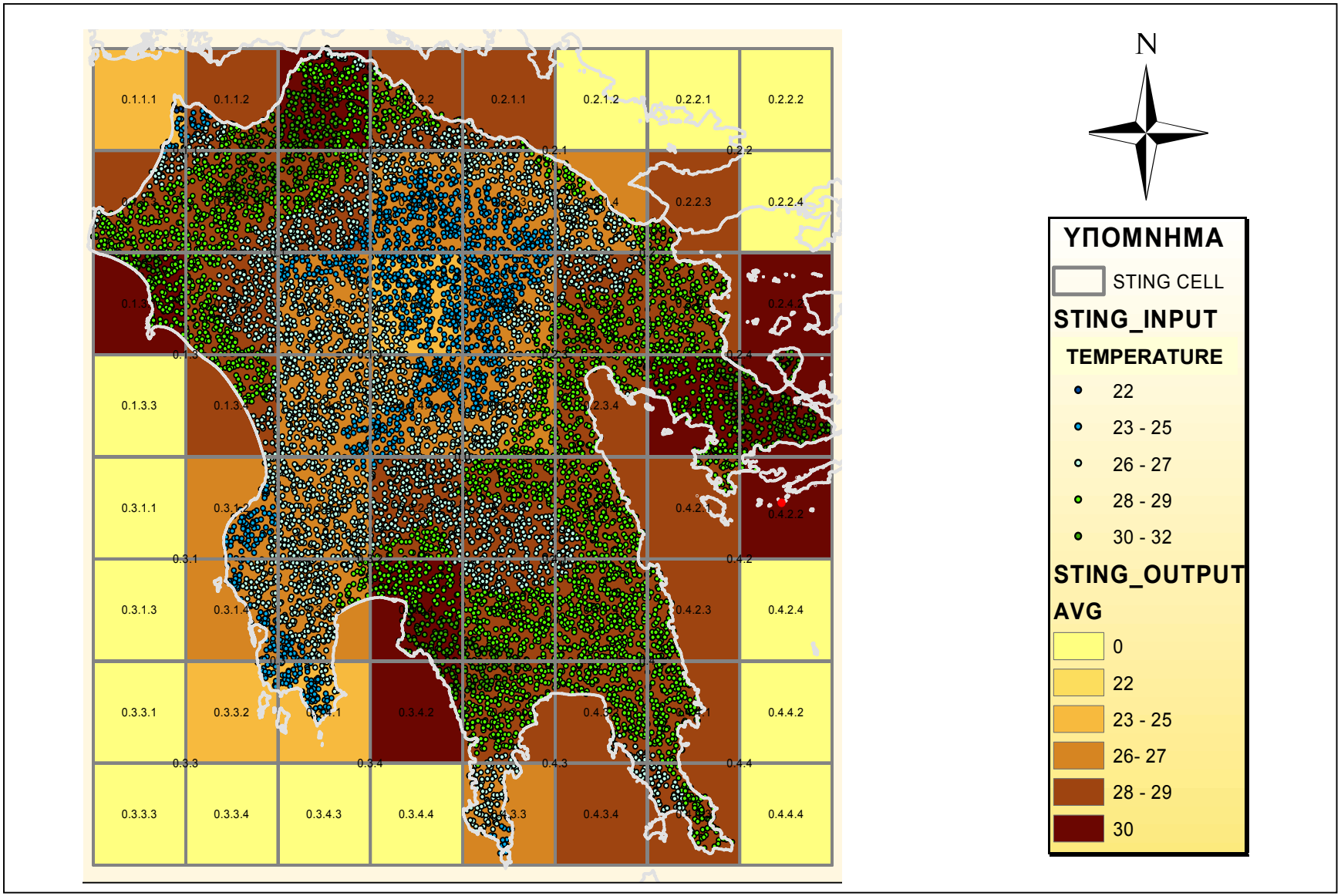


Σχήμα 4.4: Αντιστοιχία ιεραρχικής διαίρεσης χώρου βάσει του αλγορίθμου *STING*, *Peano Curve* και *Morton Matrix*.

Για κάθε ένα από τα κελιά που σχηματίζονται, ο αλγόριθμος *STING* υπολογίζει τα εξής στατιστικά στοιχεία:

- Μέσο όρο θερμοκρασίας
- Μέγιστη θερμοκρασία
- Ελάχιστη θερμοκρασία

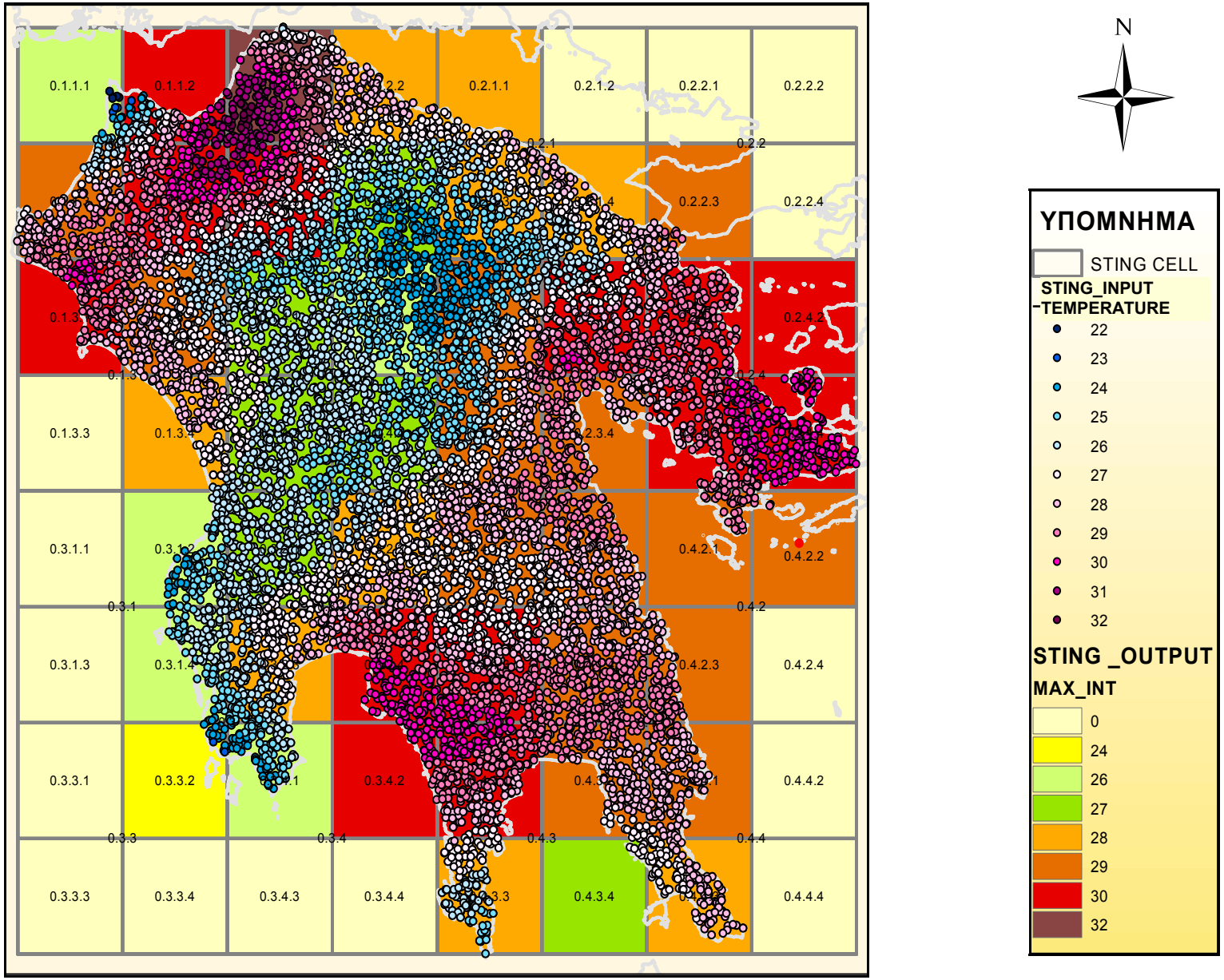
Ο υπολογισμός αυτός γίνεται από το κατώτατο στο ανώτατο επίπεδο. Η οπτικοποίηση για τον μέσο όρο της θερμοκρασίας (*AVG*) φαίνεται στο Σχήμα 4.5:



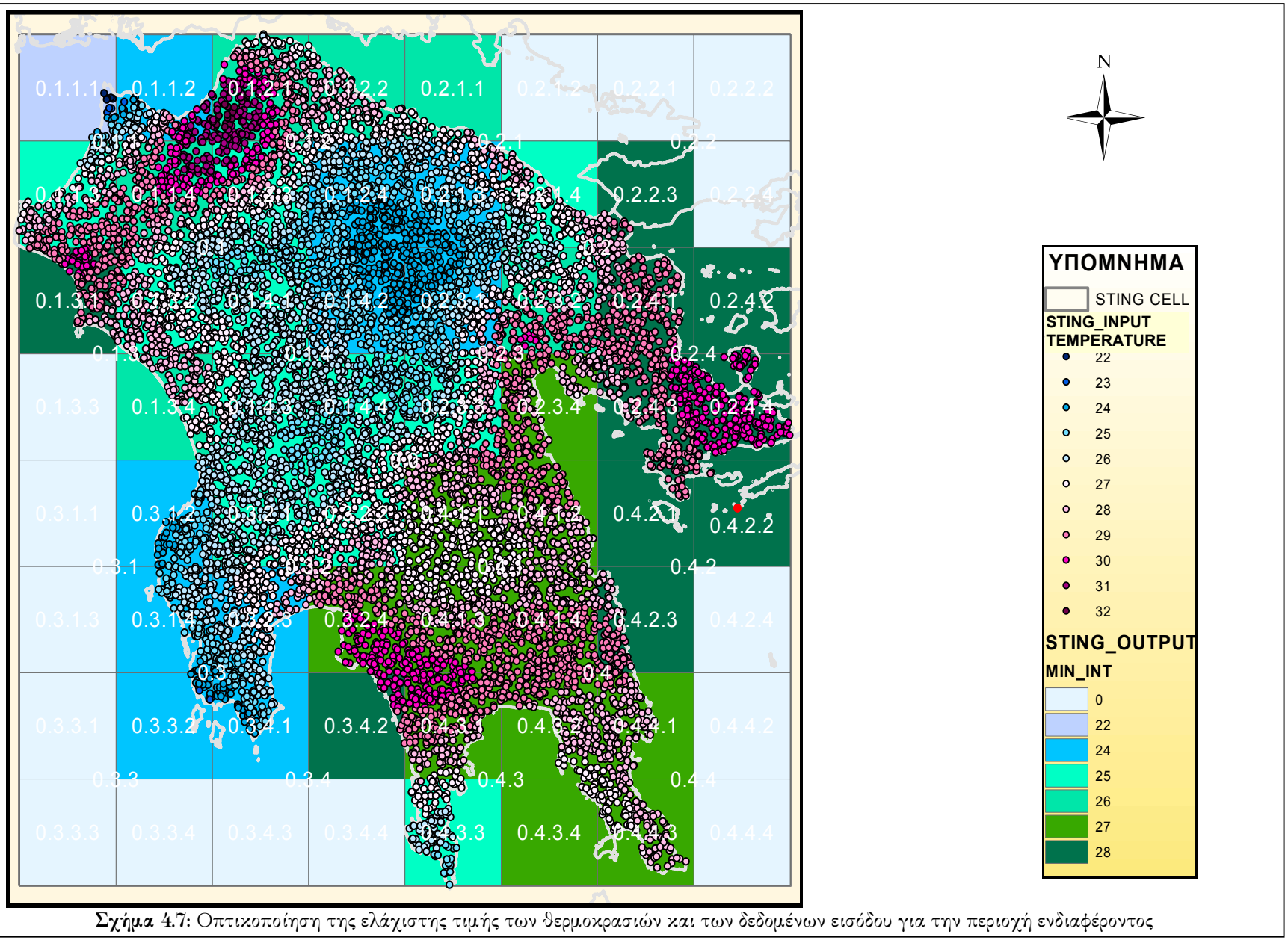
Σχήμα 4.5: Οπτικοποίηση του μέσου όρου των θερμοκρασιών και των δεδομένων εισόδου για την περιοχή ενδιαφέροντος

Η οπτικοποίηση για τα στατιστικά μεγέθη μέγιστης και ελάχιστης θερμοκρασίας φαίνεται στα Σχήματα 4.5 και 4.6 αντίστοιχα.

Σημειώνεται ότι η κωδικοποίηση των κελιών διαφοροποιείται από αυτήν στα Σχήματα 4.3 και 4.4, στα οποία έχει ακολουθηθεί κωδικοποίηση με βάση την βιβλιογραφία. Πάντως, ο αριθμός των επιπέδων είναι προφανές ότι είναι ο ίδιος στα Σχήματα 4.3, 4.4, 4.5, 4.6 και 4.7.



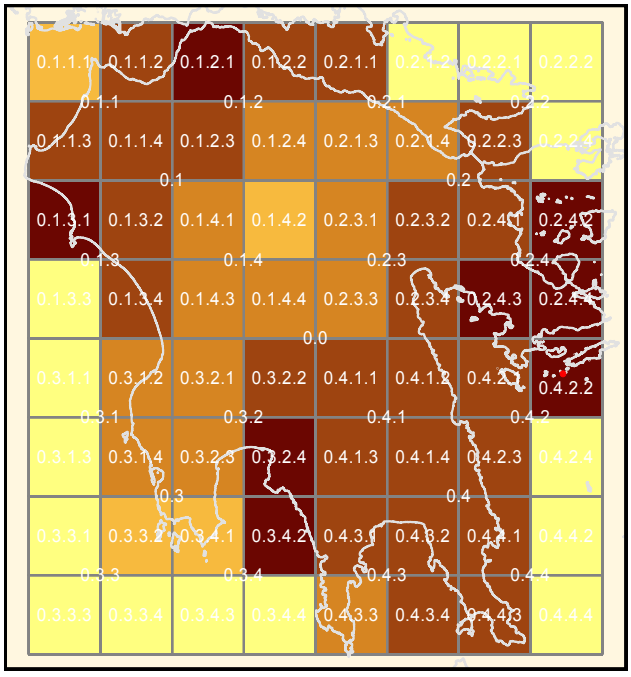
Σχήμα 4.6: Οπτικοποίηση της μέγιστης τιμής των θερμοκρασιών και των δεδομένων εισόδου για την περιοχή ενδιαφέροντος



Σχήμα 4.7: Οπτικοποίηση της ελάχιστης τιμής των θερμοκρασιών και των δεδομένων εισόδου για την περιοχή ενδιαφέροντος

Από πλευράς ποιοτικής αξιολόγησης των αποτελεσμάτων, παρατηρείται σύμπτωση των αποτελεσμάτων του αλγορίθμου, σε σχέση με τα αρχικά δεδομένα. Πιο συγκεκριμένα, είναι:

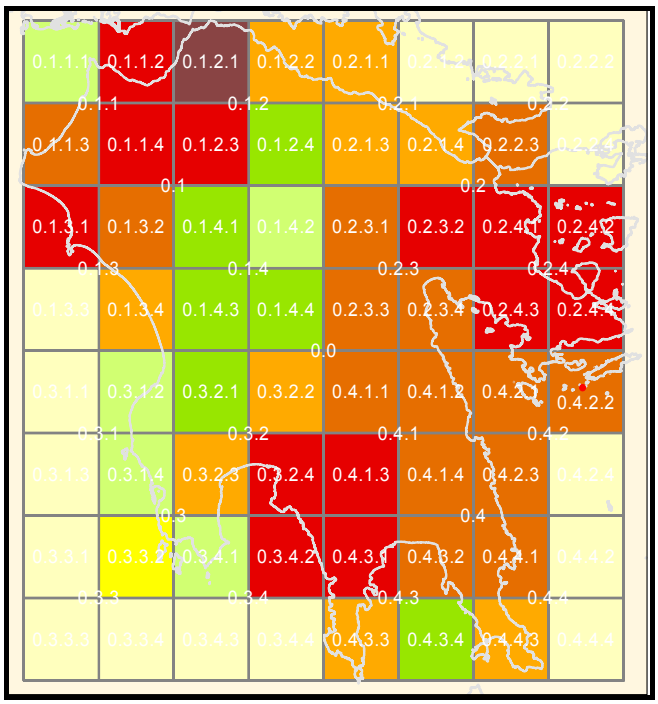
- Μέση θερμοκρασία (τυχαία παρατήρηση στο κελί «0.2.4.1»): Το εύρος μέσης τιμής θερμοκρασίας είναι 28-29 ° C (Σχήμα 4.8) και συνάδει με το εύρος τιμής «28-29 ° C» που προκύπτει από το υπόμνημα για τα δεδομένα εισόδου (Σχήμα 4.5).



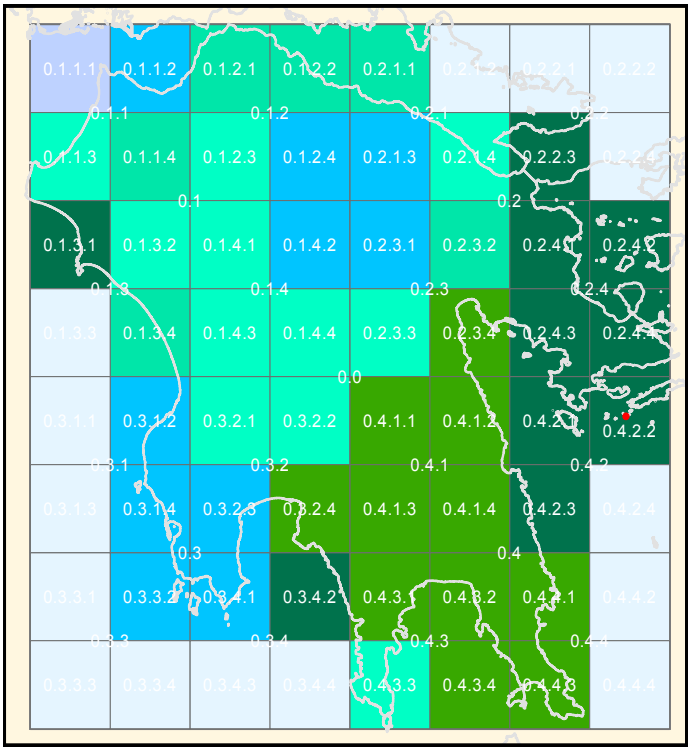
Σχήμα 4.8: Οπτικοποίηση μέσου όρου θερμοκρασιών (κελί 0.2.4.1)

- Μέγιστη τιμή θερμοκρασίας: Είναι προφανές (Σχήμα 4.6 και Σχήμα 4.9) ότι οι υψηλότερες μέγιστες θερμοκρασίες παρατηρούνται σε περιοχές που τα δεδομένα εισόδου παρουσιάζουν επίσης υψηλές τιμές θερμοκρασίας.

- Ελάχιστη τιμή θερμοκρασίας: Αντίστοιχα, οι χαμηλότερες ελάχιστες θερμοκρασίες παρατηρούνται σε περιοχές στις οποίες τα δεδομένα εισόδου παρουσιάζουν αντίστοιχα χαμηλές τιμές θερμοκρασίας (Σχήμα 4.7 και 4.10).



Σχήμα 4.9: Οπτικοποίηση μεγίστων θερμοκρασιών ανά κελί



Σχήμα 4.10: Οπτικοποίηση ελαχίστων θερμοκρασιών ανά κελί

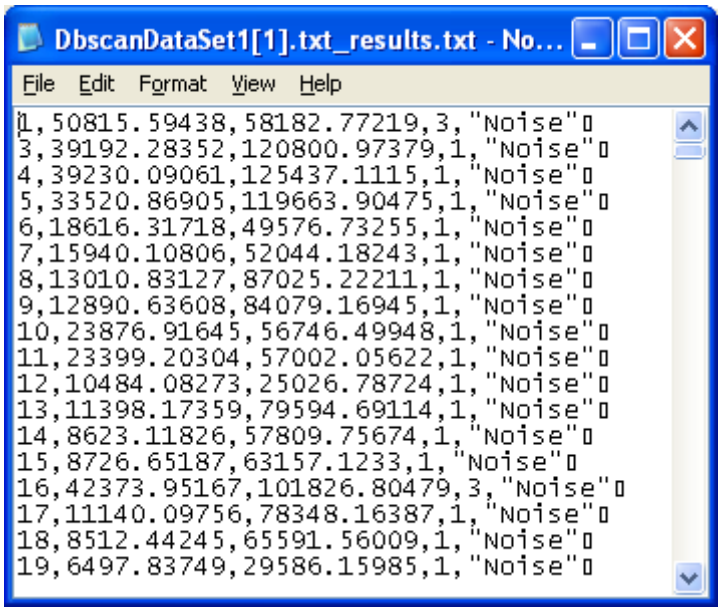
Όπως προκύπτει, τα αποτελέσματα της συσταδοποίησης με τον αλγόριθμο *STING* είναι σχετικά αξιόπιστα συγκρινόμενα με τα αρχικά δεδομένα, σε περιοχές που η αριθμητική τιμή του εύρους παρουσιάζει κάποια συνέχεια, ώστε να είναι εφικτή η σύγκριση μόνο με παρατήρηση επί χάρτου.

Λόγω του ότι τα αρχικά δεδομένα αφορούσαν σε σημειακές μετρήσεις, η εξασφάλιση του οπτικού αποτελέσματος της συνέχειας έγινε με χρήση μεγάλου μέγεθος συμβόλου για την σημειακή τιμή της μέτρησης. Συνεπώς, η παρέκκλιση της περιοχής μελέτης από τα όρια που προκύπτουν με την εφαρμογή του συγκεκριμένου αλγορίθμου– όπου αυτή διαπιστλωνεται - οφείλεται σε αυτό και μόνο το λόγο. Για μεγαλύτερο αριθμό επιπέδων, ο αλγόριθμος διαιρεί το χώρο σε περισσότερα κελιά και το στατιστικό αποτέλεσμα είναι πιο κοντά στην πραγματική τιμή του φαινομένου. Τα κελιά στα οποία ο αλγόριθμος αντιστοιχίζει τιμή μηδέν (0), είναι κελιά για τα οποία δεν υπάρχουν δεδομένα

4.6.2. Αλγόριθμος *DBSCAN*

Τα αποτελέσματα της συσταδοποίησης που εκτελεί ο αλγόριθμος *DBSCAN* παρατίθενται αποσπασματικά, για τις πρώτες οκτακόσιες εγγραφές, στο ΠΑΡΑΡΤΗΜΑ Π3. Αυτό οφείλεται στο μεγάλο χρονικό διάστημα που απαιτεί η εκτέλεση του αλγορίθμου. Αναφέρεται ότι για αυτά τα αποτελέσματα απαιτήθηκε χρόνος μεγαλύτερος της μιας ώρας, ενώ για τις επόμενες δύο ώρες δεν είχε ολοκληρωθεί η εκτέλεση του αλγορίθμου.

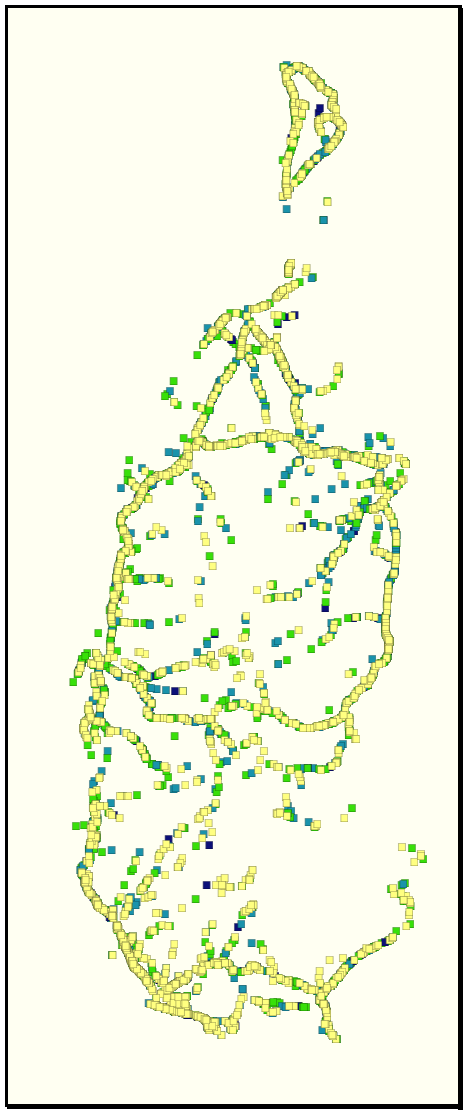
Τα προκύπτοντα αποτελέσματα σε μορφή *.txt* φαίνονται στο Σχήμα 4.8:



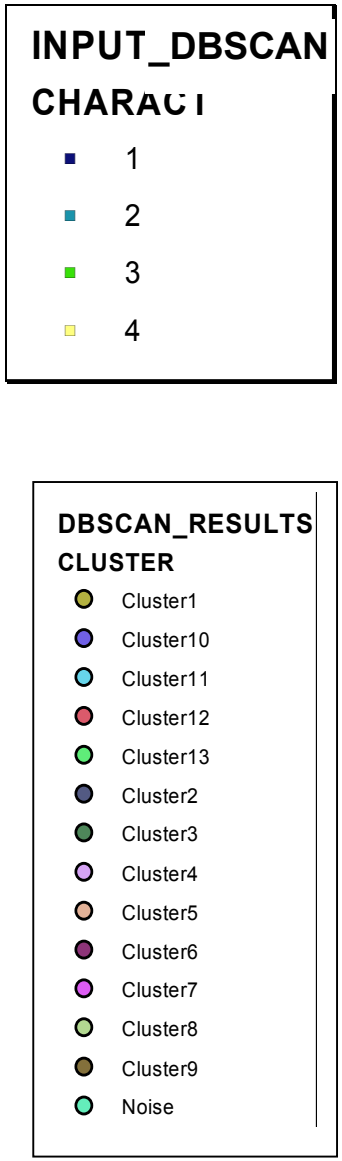
Σχήμα 4.11: Αποτελέσματα συσταδοποίησης αλγορίθμου *DBSCAN* για ελάχιστη ακτίνα ίση προς 1200 και ελάχιστο αριθμό σημείων ίσο προς 10

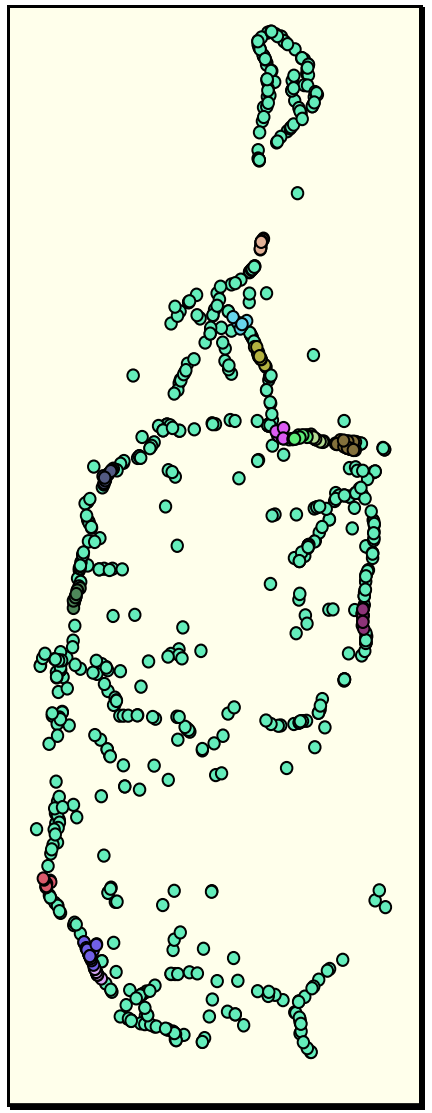
Στο αρχείο αυτό διακρίνονται τέσσερις στήλες, οι οποίες αντιστοιχούν στο X, το Y, την κατηγορία των ατυχημάτων και την συστάδα / θόρυβο.

Το αρχείο κατά τα γνωστά εισάγεται σε περιβάλλον *ArcMap* (βλ. ενότητα 4.6.1). Με την ίδια διαδικασία εισάγεται σε περιβάλλον *ArcMap* και το αρχείο με τα δεδομένα εισόδου. Το οπτικό αποτέλεσμα φαίνεται στο Σχήμα 4.9:



(α)





(β)

Σχήμα 4.12: Δεδομένα εισόδου και πληροφορία εξόδου στον αλγόριθμο *K - means*

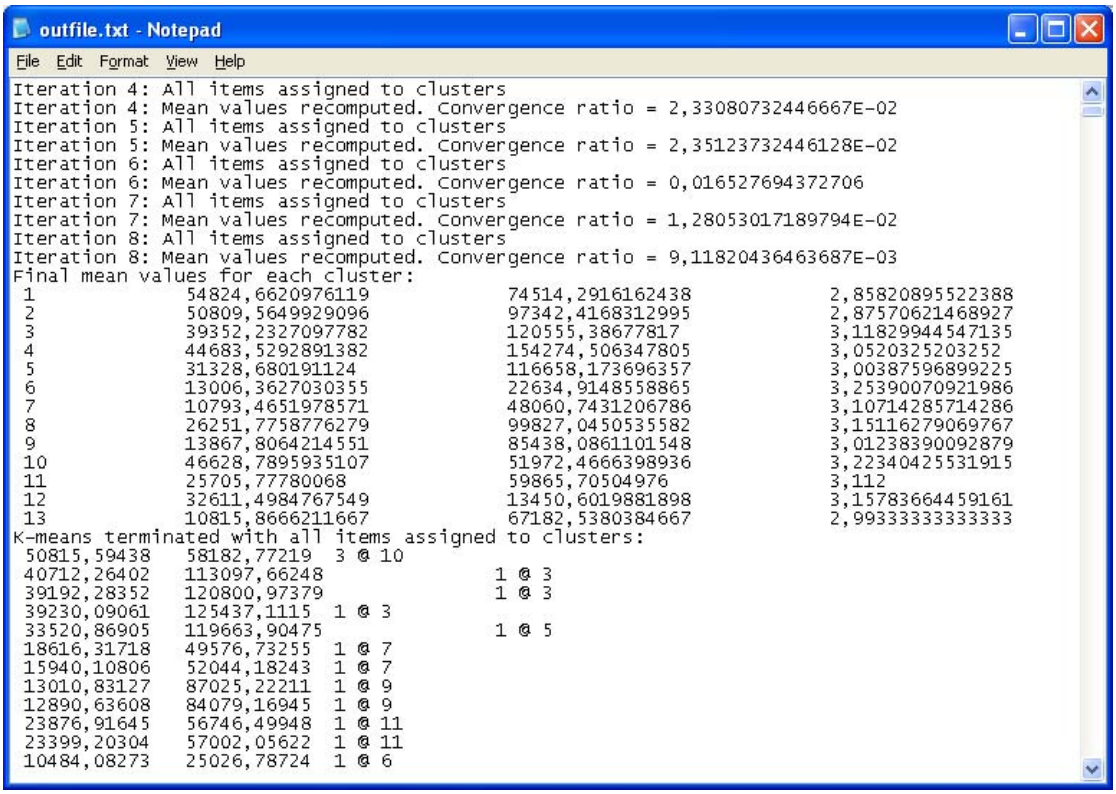
Προκειμένου να προκύψει η (τμηματική) συσταδοποίηση του Σχήματος 4.9, απαιτήθηκαν αρκετές δοκιμές με τις παραμέτρους *Eps* και *MinPts*., καθότι όλες οι εγγραφές στα αποτελέσματα λογίζονταν θόρυβος. Αυτό ακριβώς το γεγονός καθώς και το ότι η εκτέλεση του αλγορίθμου δεν ολοκληρώθηκε, δεν προκύπτει σωστή αξιολόγηση του *DBSCAN*. Είναι προφανές λοιπόν ότι το αποτέλεσμα του αλγορίθμου επηρεάζεται σε μεγάλο βαθμό από τις τιμές των παραμέτρων αυτών. Εδώ θεωρήθηκε *Eps* = 1200 και *MinPts* = 10.

Με δεδομένο ότι η εκτέλεση του αλγορίθμου δεν ολοκληρώθηκε, ο αλγόριθμος *DBSCAN* καθίσταται ασύμφορος για μεγάλες βάσεις δεδομένων.

Τέλος, από το τμηματικό αυτό αποτέλεσμα εξόδου του αλγορίθμου, δεν μπορούν να εξαχθούν ασφαλή συμπεράσματα, δεδομένου ότι μπορεί να υπάρχουν μεγάλες διακυμάνσεις στην πυκνότητα των σημείων μέσα στο τυχαίο δείγμα [BX02].

4.6.2. Αλγόριθμος *K – means*

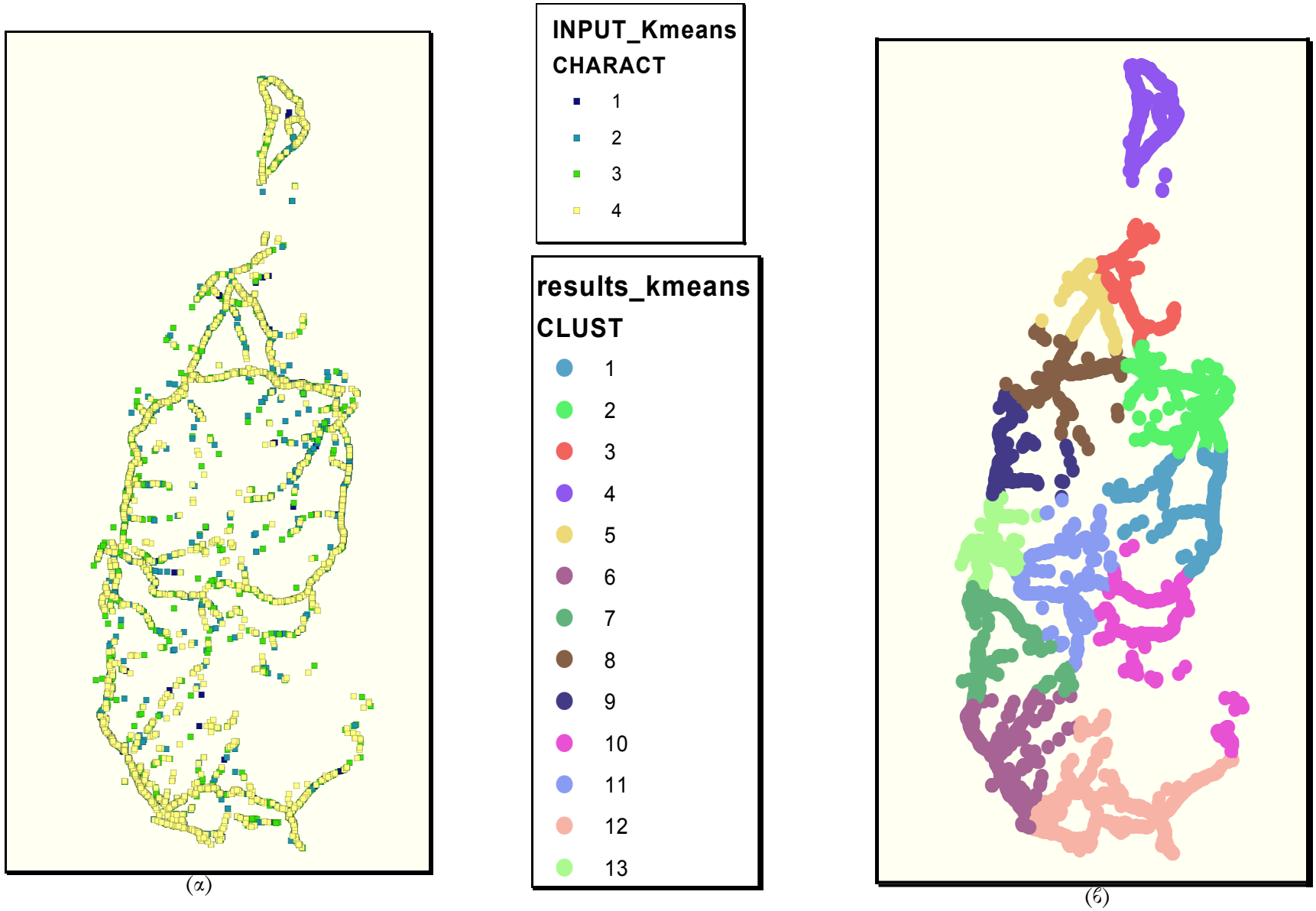
Τα αποτελέσματα της συσταδοποίησης που εκτελεί ο αλγόριθμος *K - means* παρατίθενται αποσπασματικά, λόγω μεγάλου πλήθους δεδομένων, για τις πρώτες χίλιες εγγραφές, στο ΠΑΡΑΡΤΗΜΑ Π3. Πρόκειται για ένα *.txt* αρχείο, το οποίο αποσπασματικά φαίνεται στο Σχήμα 4.10:



Σχήμα 4.13: Αποτελέσματα συσταδοποίησης αλγορίθμου *K - means* για αριθμό συστάδων ίσο προς 13.

Στο αρχείο αυτό διακρίνονται τέσσερις στήλες, οι οποίες αντιστοιχούν στο *X*, το *Y*, την κατηγορία των ατυχημάτων και την συστάδα.

Ομοίως το αρχείο εισάγεται σε περιβάλλον *ArcMap*. Με την ίδια διαδικασία εισάγεται σε περιβάλλον *ArcMap* και το αρχείο με τα δεδομένα εισόδου. Το οπτικό αποτέλεσμα φαίνεται στο Σχήμα 4.11. Το μεγάλο πλήθος των δεδομένων δεν επιτρέπει την οπτικοποίηση της συσχέτισης συστάδας και χαρακτηρισμού ατυχήματος μέσω γραφήματος.



Σχήμα 4.14: Δεδομένα εισόδου και πληροφορία εξόδου στον αλγόριθμο *K - means*

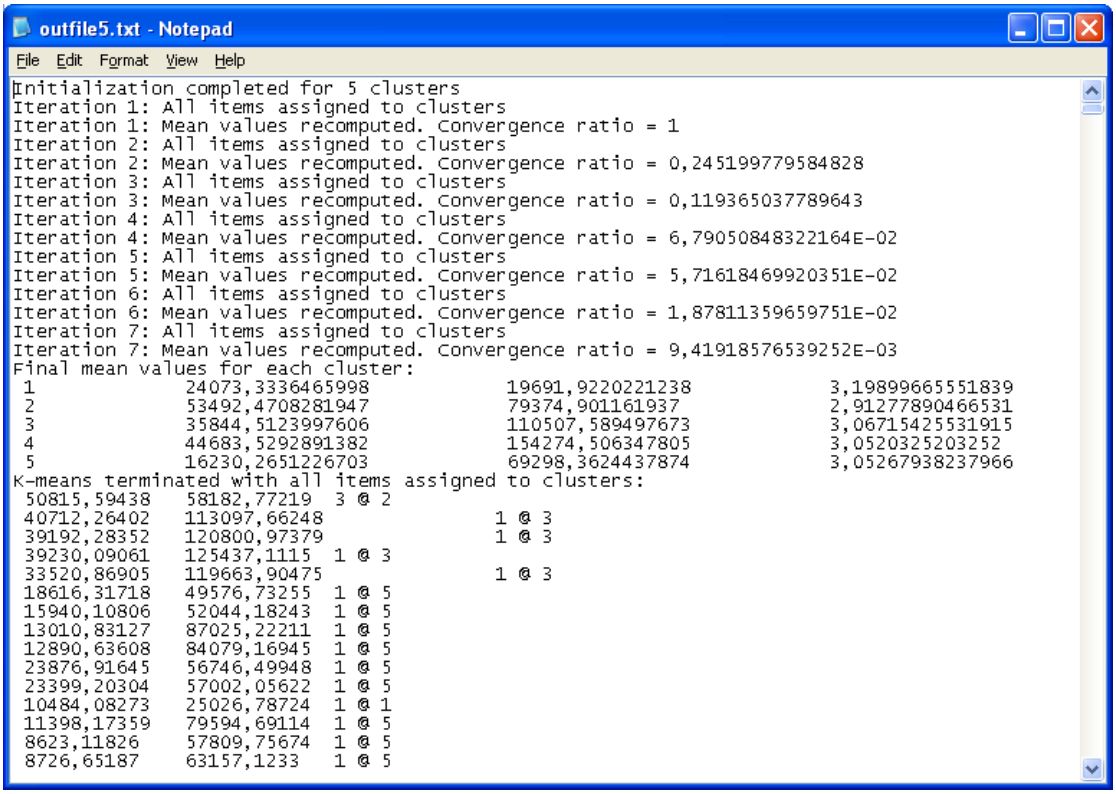
Για κάθε συστάδα, η μέση τιμή της σοβαρότητας (χαρακτηρισμού) του συμβάντος σύμφωνα με τις τιμές που επιστρέφει ο *K – means*, έχει ως εξής:

- Συστάδα 1: 2,858
- Συστάδα 2: 2,876
- Συστάδα 3: 3,118
- Συστάδα 4: 3,052
- Συστάδα 5: 3,003
- Συστάδα 6: 3,254
- Συστάδα 7: 3,107
- Συστάδα 8: 3,151
- Συστάδα 9: 3,012
- Συστάδα 10: 3,223
- Συστάδα 11: 3,112
- Συστάδα 12: 3,158
- Συστάδα 13: 2,993

Όπως προκύπτει, τα αποτελέσματα της συσταδοποίησης με τον αλγόριθμο *K – means* δίνουν μια γενική εικόνα της σοβαρότητας του συμβάντος. Έτσι, τα πιο ελαφρά συμβάντα παρατηρούνται στην Συστάδα 10, ενώ τα πιο σοβαρά στην Συστάδα 1, βάσει της κλίμακας που δόθηκε στην Ενότητα 4.1.

Εκτέλεση του αλγορίθμου με αριθμό συστάδων ίσο προς επτά (7), δίνει αποτελέσματα (αποσπασματικά εδώ) που φαίνονται στο Σχήμα 4.12. Ειδικότερα, είναι:

- Συστάδα 1: 3,199
- Συστάδα 2: 2,913
- Συστάδα 3: 3,067
- Συστάδα 4: 3,052
- Συστάδα 5: 3,053



Σχήμα 4.15: Αποτελέσματα συσταδοποίησης αλγορίθμου *K - means* για αριθμό συστάδων ίσο προς πέντε (5).

Τα αποτελέσματα του αλγόριθμου *K – means* εξαρτώνται από τον αρχικό αριθμό των συστάδων και το κριτήριο σύγκλισης.

Κεφάλαιο 5
Συμπεράσματα

Η εξέλιξη της Εξόρυξης Γνώσης είναι αποτέλεσμα πολύχρονης επιρροής και μελέτης πληθώρας επιστημονικών κλάδων, όπως είναι οι βάσεις δεδομένων, η ανάκτηση πληροφοριών, η στατιστική, οι αλγόριθμοι και η μηχανική μάθηση, τα συστήματα υποστήριξης αποφάσεων, κ.ά. Η χρησιμότητα της εξόρυξης γνώσης γίνεται αντιληπτή, δεδομένης της αύξησης του όγκου της πληροφορίας και των διαθέσιμων συστημάτων βάσεων δεδομένων, τα οποία καθιστούν επιτακτική την ανάγκη για την εύρεση και χρήση τεχνικών και εργαλείων τα οποία υποστηρίζουν την αυτόματη μετατροπή των υπό επεξεργασία δεδομένων σε χρήσιμη πληροφορία και γνώση.

Προκειμένου να διασφαλισθεί η λειτουργικότητα των συστημάτων διαχείρισης βάσεων δεδομένων, αυτά πρέπει να επεκταθούν προκειμένου να συμπεριλαμβάνουν δεδομένα με χωρική αναφορά. Η επεξεργασία αυτών των δεδομένων, ώστε να προκύψει (χωρική) πληροφορία χρήσιμη και διαχειρίσιμη, πραγματοποιείται με την Εξόρυξη Χωρικής Γνώσης. Ορισμένες από τις εφαρμογές Εξόρυξης Χωρικής Γνώσης εντάσσονται στα πεδία των γεωγραφικών συστημάτων πληροφοριών, της

γεωλογίας, της περιβαλλοντικής επιστήμης, της διαχείρισης πόρων, της γεωργίας, της ιατρικής και της ρομποτικής. Η διαδικασία εξόρυξη (χωρικής) γνώσης πραγματοποιείται με τη βοήθεια εργασιών (συσταδοποίηση, κατηγοριοποίηση, κανόνες χωρικών συσχετίσεων, κ.ά) και αλγορίθμων. Το εξαγόμενο της διαδικασίας είναι ένα σύνολο υποθέσεων (πρότυπα), τα οποία μπορούν να επιβεβαιωθούν με ακρίβεια κάνοντας χρήση στατιστικών εργαλείων και να οπτικοποιηθούν με τη χρήση γεωγραφικών συστημάτων πληροφοριών.
Πιο συγκεκριμένα, η συσταδοποίηση συνιστά ένα πολλά υποσχόμενο ερευνητικό πεδίο της οποίας οι δυνητικές εφαρμογές θέτουν τις δικές τους απαιτήσεις. Μερικές τυπικές εφαρμογές της συσταδοποίησης παρατηρούνται στα πεδία μάρκετινγκ και οικονομίας, ιατρικής, ανθρωπολογίας, βιολογίας (*taxonomy*), επεξεργασία εικόνας, ανάκτηση κειμένων. Πρόσφατες χρήσεις της συσταδοποίησης περιλαμβάνουν την εξέταση των δεδομένων των αρχείων λειτουργίας του *Web* για τον εντοπισμό προτύπων σχετικά με τον τρόπο χρήσης του δικτύου. Ακόμη, σημαντικότατη εφαρμογή της συσταδοποίησης συνιστά η ανάλυση χωρικών δεδομένων. Η συσταδοποίηση βοηθά στην αυτοματοποίηση της διαδικασίας ανάλυσης και κατανόησής τους. Χρησιμοποιείται προκειμένου να ταυτοποιήσει και να εξάγει ενδιαφέροντα χαρακτηριστικά και πρότυπα που ενδέχεται να υπάρχουν σε μεγάλες βάσεις χωρικών δεδομένων.

Σημαντικοί αλγόριθμοι συσταδοποίησης είναι ο *STING*, ο *DBSCAN* και ο *K – means*. Πιο συγκεκριμένα, ο αλγόριθμος *STING* απαντά πολλά αιτήματα για εξόρυξη γνώσης από δεδομένα, εξετάζοντας τα στατιστικά που προκύπτουν για τα δημιουργούμενα κελιά. Ακόμη, ενδέχεται να μην χρειάζεται η σάρωση ολόκληρης της βάσης δεδομένων, γεγονός το οποίο είναι πολύ αποδοτικό όταν γίνονται πολλαπλές αιτήσεις για εξόρυξη γνώσης από τα δεδομένα. Ο αλγόριθμος *DBSCAN* επηρεάζεται από τιμές παραμέτρων (*Eps, MinPts*)οι οποίες είναι δύσκολο να προσδιορισθούν, δεν χρησιμοποιεί κάποια μορφή προσυσταδοποίησης, αλλά εφαρμόζεται απευθείας στο σύνολο των δεδομένων, με αποτέλεσμα να καθίσταται ασύμφορος για μεγάλες βάσεις δεδομένων, λόγω του κόστους *I / O*, ενώ η χρήση δείγματος προκειμένου να περιορισθεί το μέγεθος της εισόδου, δεν είναι εφικτή. Ο λόγος είναι ότι ακόμη και αν το δείγμα είναι μεγάλο, μπορεί να

υπάρχουν μεγάλες διακυμάνσεις στην πυκνότητα των σημείων μέσα σε κάθε συστάδα στο τυχαίο δείγμα. Ο αλγόριθμος *K – means* είναι σχετικά επιδεκτικός στην μεταβολή της κλίμακάς και αποτελεσματικός στην επεξεργασία μεγάλων συνόλων δεδομένων, ενώ είναι πολύ ευαίσθητος στον θόρυβο και τα απομακρυσμένα σημεία, δεδομένου ότι ένας μικρός αριθμός τέτοιων δεδομένων μπορούν να επηρεάσουν ουσιαστικά την μέση τιμή. Η χρήση τεχνικών οπτικοποίησης επιτρέπει να εξαχθούν και να γίνουν αντιληπτά πιο πολύπλοκα αποτελέσματα από αυτά των μαθηματικών και περιγραφικών τρόπων παρουσίασης των αποτελεσμάτων. Έτσι, η οπτικοποίηση των αποτελεσμάτων των παραπάνω τριών αλγορίθμων αφενός επιβεβαιώνει σε μεγάλο βαθμό την βιβλιογραφία και, αφετέρου, δίνει μια ποιοτική αξιολόγηση της πληροφορίας εξόδου. Πιο συγκεκριμένα, τα αποτελέσματα της συσταδοποίησης με τον αλγόριθμο *STING* είναι σχετικά αξιόπιστα συγκρινόμενα με τα αρχικά δεδομένα, σε περιοχές που αυτά παρουσιάζουν συνέχεια, ώστε να είναι εφικτή η σύγκριση μόνο με παρατήρηση επί χάρτου.

Η εκτέλεση του αλγορίθμου *DBSCAN* απαιτεί μεγάλο χρονικό διάστημα.. Προκειμένου να προκύψει (τμηματική) συσταδοποίηση απαιτήθηκαν αρκετές δοκιμές με τις παραμέτρους *Eps* και *MinPts*., καθότι όλες οι εγγραφές στα αποτελέσματα λογίζονταν θόρυβος. Αυτό ακριβώς το γεγονός καθώς και το ότι η εκτέλεση του αλγορίθμου δεν ολοκληρώθηκε, δεν προκύπτει σωστή αξιολόγηση του *DBSCAN*. Είναι προφανές λοιπόν ότι το αποτέλεσμα του αλγορίθμου επηρεάζεται σε μεγάλο βαθμό από τις τιμές των παραμέτρων αυτών. Εδώ θεωρήθηκε *Eps* = 1200 και *MinPts* = 10. Έτσι, ο αλγόριθμος *DBSCAN* καθίσταται ασύμφορος για μεγάλες βάσεις δεδομένων, κάτι το οποίο προκύπτει και από την βιβλιογραφία..

Τα αποτελέσματα της συσταδοποίησης με τον αλγόριθμο *K – means* δίνουν μια γενική εικόνα της σοβαρότητας του συμβάντος και εξαρτώνται από τον αρχικό αριθμό των συστάδων και το κριτήριο σύγκλισης.

Όπως προκύπτει και από την οπτικοποίηση των αποτελεσάτων των υλοποιηθέντων αλγορίθμων, η ερμηνεία της σημασιολογίας κάθε συστάδας ενδέχεται να είναι δύσκολη. Καθοριστική κρίνεται η συμβολή ενός ειδικού του πεδίου προκειμένου να ανατεθούν «ετικέτες» (προσδιορισμοί) στις συστάδες. Ένα άλλο ζήτημα που ανακύπτει είναι τι δεδομένα θα πρέπει να χρησιμοποιηθούν για τη συσταδοποίηση. Σε αντίθεση με τη μάθηση κατά τη διάρκεια της διαδικασίας κατηγοριοποίησης, όπου υπάρχει εκ των προτέρων κάποια γνώση σχετικά με το ποια πρέπει να είναι τα γνωρίσματα της κατηγοριοποίησης, στη συσταδοποίηση δεν υπάρχει επιβλεπόμενη μάθηση για να βοηθήσει τη διαδικασία.

Αναφορικά με το παρόν και το μέλλον της Εξόρυξης (Χωρικής) Γνώσης, θα έλεγε κανείς ότι όλες οι λύσεις των προβλημάτων πρέπει να είναι ικανές να εφαρμόζονται στις βάσεις δεδομένων του πραγματικού κόσμου. Όσον αφορά στην αποτελεσματικότητα, υπάρχει ενδιαφέρον για τους αλγορίθμους και τις δομές δεδομένων που χρησιμοποιούνται. Είναι σημαντικό πώς συμπεριφέρονται οι προτεινόμενοι αλγόριθμοι καθώς τροποποιείται η βάση δεδομένων. Πολλοί αλγόριθμοι εξόρυξης γνώσης μπορούν να δουλέψουν καλά σε μία στατική βάση δεδομένων, αλλά είναι ιδιαίτερα αναποτελεσματικοί όταν γίνονται αλλαγές στη βάση δεδομένων. Το ενδιαφέρον εστιάζεται στο πώς αποδίδουν οι αλγόριθμοι σε πολύ μεγάλες βάσεις δεδομένων παρά για το πώς λειτουργούν σε απλοϊκά προβλήματα. Σε θέματα κοινωνικής φύσεως, η ενσωμάτωση των τεχνικών εξόρυξης γνώσης από στις καθημερινές δραστηριότητες αποτελεί πια συνηθισμένη δραστηριότητα. Τρανταχτό παράδειγμα συνιστούν οι διαφημίσεις, ενώ οι επιχειρήσεις έχουν γίνει πιο αποτελεσματικές στο να μειώσουν τα έξοδα τους με χρήση της διαδικασίας εξόρυξης γνώσης. Όμως, υπάρχει ανησυχία ότι αυτές οι πληροφορίες παρέχονται με κόστος την καταπάτηση της ιδιωτικής ζωής. Οι εφαρμογές εξόρυξης γνώσης μπορούν να εξάγουν πολλές δημογραφικές πληροφορίες που αφορούν πελάτες, οι οποίες ήταν πριν άγνωστες ή κρυμμένες στα δεδομένα. Η μη εξουσιοδοτημένη χρήση αυτών των δεδομένων θα μπορούσε να οδηγήσει στην αποκάλυψη πληροφοριών που θεωρούνται εμπιστευτικές. Από τεχνική σκοπιά, πολλές εταιρείες στον χώρο της βιομηχανίας βάσεων δεδομένων έχουν προϊόντα

ειδικά σχεδιασμένα προκειμένου γι την διαχείριση χωρικών δεδομένων (*Spatial Data Engine - SDE*). Η λειτουργικότητα που παρέχεται από αυτά τα συστήματα περιλαμβάνει ένα σύνολο χωρικών τύπων δεδομένων και χωρικών τελεστών. Ένα πρότυπο σύνολο χωρικών τύπων δεδομένων και λειτουργιών έχει αναπτυχθεί από *Open GIS Consortium – OGIS*. Οι εν λόγω τύποι και λειτουργίες μπορούν να αποτελέσουν τμήμα μια αντικειμενο – σχεσιακής γλώσσας ερωτημάτων, όπως η *SQL3*. Η βελτίωση της απόδοσης που παρέχεται από τα συστήματα αυτά περιλαμβάνει ένα πολυδιάστατο χωρικό ευρετήριο και αλγόριθμους για χωρικές λειτουργίες και ερωτήματα.

Σήμερα, η εξόρυξη γνώσης από δεδομένα είναι κάτι παραπάνω από ένα σύνολο από εργαλεία τα οποία μπορούν να χρησιμοποιηθούν για να ανακαλύψουν κρυμμένες πληροφορίες από τις βάσεις δεδομένων. Παρά την ύπαρξη πολλών εργαλείων που βοηθούν σε αυτή τη διαδικασία, δεν υπάρχει ένα μοντέλο ή μία προσέγγιση που να τα περιλαμβάνει όλα. Σύντομα στα επόμενα χρόνια, θα υπάρξουν όχι μόνο περισσότεροι αλγόριθμοι με καλύτερες διεπαφές, αλλά θα γίνουν και βήματα για την ανάπτυξη ενός μοντέλου εξόρυξης γνώσης από δεδομένα που θα τα περιέχει όλα. Εάν και δε θα μοιάζει με το σχεσιακό μοντέλο, πιθανότατα θα περιέχει παρόμοια στοιχεία: αλγόριθμους, μοντέλο δεδομένων και μέτρα αξιολόγησης. Τα σημερινά εργαλεία της εξόρυξης γνώσης από δεδομένα, απαιτούν υψηλή ανθρώπινη αλληλεπίδραση όχι μόνο για να οριστεί η απαίτηση αλλά επίσης και για να ερμηνευτούν τα αποτελέσματα. Καθώς τα εργαλεία γίνονται καλύτερα και πιο ολοκληρωμένα, αυτή η εκτεταμένη ανθρώπινη αλληλεπίδραση πιθανότατα θα μειωθεί. Οι εφαρμογές της εξόρυξης γνώσης από δεδομένα είναι διαφορετικών ειδών, με αποτέλεσμα να είναι επιθυμητή η δημιουργία ενός ολοκληρωμένου μοντέλου εξόρυξης γνώσης. Σημαντική ανάπτυξη θα ήταν η δημιουργία μίας εξειδικευμένης «γλώσσας ερωτήσεων» η οποία θα περιλαμβάνει τις παραδοσιακές *SQL* συναρτήσεις.

Ήδη έχει προταθεί μια γλώσσα ερωτήσεων εξόρυξης γνώσης που ονομάζεται *DMQL* (*Data Mining Query Language*) και η οποία βασίζεται στην *SQL*.

Αντίθετα με την *SQL*, όπου υποτίθεται ότι υπάρχει προσπέλαση μόνο σε σχεσιακές βάσεις δεδομένων, η *DMQL* επιτρέπει την προσπέλαση σε πληροφορίες όπως η ιεραρχία εννοιών. Μία άλλη διαφορά είναι ότι τα δεδομένα που ανακτώνται δεν χρειάζεται να αποτελούν ένα υποσύνολο ή μία συνάθροιση των δεδομένων των σχέσεων. Έτσι, μία *DMQL* δήλωση πρέπει να υποδείξει το είδος της γνώσης που πρόκειται να εξορυχτεί. Μία άλλη διαφορά είναι ότι μία *DMQL* δήλωση μπορεί να υποδηλώνει την απαραίτητη σημασία ή το κατώφλι που πρέπει να ικανοποιεί η πληροφορία που εξορύσσεται.

Έχει δημιουργηθεί ο όρος Σύστημα Διαχείρισης Ανακάλυψης Γνώσης και Δεδομένων - ΣΔΑΓΔ (*Knowledge and Data Discovery Management System – KDDMS*) για να περιγράψει τη μελλοντική γενιά των συστημάτων εξόρυξης γνώσης από δεδομένα τα οποία δεν θα περιλαμβάνουν μόνο εργαλεία εξόρυξης γνώσης αλλά επίσης και τεχνικές που θα χειρίζονται τα σχετικά δεδομένα, θα εξασφαλίζουν τη συνέπεια τους και θα προσφέρουν συνδρομικότητα και ανάκαμψη. Ένα τέτοιο σύστημα θα παρέχει προσπέλαση μέσω ειδικών ερωτήσεων εξόρυξης γνώσης από δεδομένα, οι οποίες θα έχουν βελτιστοποιηθεί για να γίνει η προσπέλαση αποτελεσματική.

Πρόσφατα παρουσιάστηκε ένα καινούργιο μοντέλο επεξεργασίας της διαδικασίας *KDD*, το επονομαζόμενο *CRISP-DM (Cross-Industry Stadard Process for Data Mining)*, με πολλές διαφορετικές εφαρμογές. Το μοντέλο απευθύνεται σε όλα τα βήματα της *KDD*, συμπεριλαμβανομένης και της συντήρησης των αποτελεσμάτων του βήματος της εξόρυξης γνώσης από δεδομένα. Ο κύκλος ζωής του *CRISP-DM* περιλαμβάνει τα επόμενα βήματα: κατανόηση του είδους της επιχείρησης, κατανόηση των δεδομένων, προετοιμασία των δεδομένων, μοντελοποίηση, ανάπτυξη. Τα βήματα που περιλαμβάνονται στο μοντέλο *CRISP-DM* μπορούν συνοπτικά να ονομαστούν σαν «τα 5Α»: *assess*, *access*, *analyse*, *act*, *automate* - προσδιορίζω, προσπελαύνω, αναλύω, ενεργώ, αυτοματοποιώ.

Πρόσφατα έχει παρατηρηθεί ένα αυξανόμενο ενδιαφέρον στις τεχνικές εξόρυξης γνώσης από δεδομένα που χρησιμοποιούνται σε εφαρμογές όπως είναι η ανίχνευση

απάτης, η αναγνώριση υπόπτων για εγκλήματα και η πρόβλεψη των πιθανών τρομοκρατών. Αυτά μπορούν να θεωρηθούν σαν τύποι προβλημάτων κατηγοριοποίησης. Η προσέγγιση που συχνά χρησιμοποιείται εδώ είναι η δημιουργία ενός «προφίλ», με μια τυπική συμπεριφορά και τα κατάλληλα χαρακτηριστικά. Πράγματι, πολλές τεχνικές κατηγοριοποίησης λειτουργούν αναγνωρίζοντας τις τιμές των γνωρισμάτων που εμφανίζονται συχνά για την υπό εξέταση κατηγορία - κλάση. Στη συνέχεια, κατηγοριοποιούνται οι καταγραφές με βάση αυτές τις τιμές των γνωρισμάτων. Ας μην ξεχνάμε ότι αυτές οι προσεγγίσεις της κατηγοριοποίησης *δεν είναι τέλειες*. Μπορεί να γίνουν λάθη. Το ότι κάποιος αγοράζει με πιστωτική κάρτα μια σειρά από προϊόντα που συνήθως αγοράζονται όταν η πιστωτική κάρτα είναι κλεμμένη, δεν σημαίνει ότι η κάρτα του είναι κλεμμένη ή ότι ο συγκεκριμένος καταναλωτής είναι εγκληματίας. Οι χρήστες των τεχνικών εξόρυξης γνώσης πρέπει να είναι ευαισθητοποιημένοι σε αυτά τα θέματα και δεν θα πρέπει να παραβιάζουν κατευθύνσεις ή οδηγίες σχετικές με θέματα προστασίας προσωπικών δεδομένων.

Ορολογία

Ακολουθιακή Ανάλυση	*Sequential Analysis*
Ακραία Σημεία	*Outliers*
Αλγόριθμος «εκ των προτέρων»	*Apriori Algorithm*
Αλγόριθμος *CLARANS* μη Χωρικής Τάξης	*Non Spatial Dominant – NSD(CLARANS)*
Αλγόριθμος Απλού Συνδέσμου	*Single Link Algorithm*
Αλγόριθμος Διαμερισμού γύρω από Μέσους	*PAM - Partitioning Around Medoids*
Αλγόριθμος Ενέργειας Δεσμού	*Bond Energy Algorithm – BEA*
Αλγόριθμος Κατανομής Δεδομένων	*DDA - Data Distribution Algorithm*
Αλγόριθμος Κατανομής Μετρητών	*CDA - Count Distribution Algorithm*
Αλγόριθμος Κλιμακούμενης Παραλληλοποιήσιμης Επαγωγής Δένδρων Αποφάσεων	*SPRINT – Scalable Parallelizable Induction of Decision Trees*
Αλγόριθμος Μέσου Συνδέσμου	*Average Link Algorithm*
Αλγόριθμος Πλήρους Συνδέσμου	*Complete Link Algorithm*
Αλγόριθμος Πλησιέστερου Γείτονα	*Nearest Neighbor Algorithm*
Αλγόριθμος Συναθροιστικής Εγγύτητας	*CRH:C: Περικλείων Κύκλος (Encompassing Circle), R: Ισοθετικό Ορθογώνιο (Isothetic Rectangle), H: Πολύγωνο (Convex Hull)*
Αλγόριθμος Συσταδοποίησης Τετραγωνικού Σφάλματος	*Squared Error Clustering Algorithm*
Άμεση Αναλυτική Επεξεργασία	*Online Analytical Processing – OLAP*

Αναγνώριση Προτύπου	*Pattern Recognition*
Ανακάλυψη Ακολουθιών	*Sequence Discovery*
Ανακάλυψη Γνώσης Σε Βάσεις Δεδομένων	*Knowledge Discovery In Databases*
Ανάλυση Συγγένειας	*Affinity Analysis*
Ανάλυση Συνδέσμων	*Link Analysis*
Ανάλυση Χρονοσειρών	*Time Series Analysis*
Ανάλυση Χρονοσειρών	*Time Series Analysis*
Ανίχνευση Τάσης	*Data Detection*
Ανταγωνιστική Μάθηση	*Competitive Learning*
Απότοκο	*Consequent*
Αριθμητικός Μέσος	*Median*
Ασαφής Συσταδοποίηση	*Fuzzy Clustering*
Αυτοοργάνωση	*Self – Organization*
Γενίκευση	*Generalization*
Γενίκευση	*Generalization*
Γενίκευση Μη Χωρικής Τάξης	*Non spatial Data Dominant Generalization*
Γενίκευση Χωρικής Τάξης	*Spatial Data Dominant Generalization*
Γραμμή Πεάνω Σχήματος «Z»	*Z – shaped Peano Curve*
Γραφικές Διεπαφές Χρήστη	*Graphic User Interface*
Γράφος Γειτνίασης	*Neighborhood Graph*
Δειγματοληψία	*Sampling*
Δένδρα Απόφασης	*Decision Trees*
Δένδρα Κατηγοριοποίησης και Παλινδρόμησης	*Classification and Regression Trees*
Δένδρο Ελάχιστης Ζεύξης	*MST - Minimum Spanning Tree*
Δένδρο Κατακερματισμού	*Hash Tree*
Διάδοση	*Propagation*
Διαιρετικοί	*Divisive*
Διαμέριση	*Partitioning*
Διαμεριστική Συσταδοποίηση	*Partitional Clustering*
Διαμεριστικοί	*Divisive*
Διάσχιση Κατά Πλάτος	*Breadth - First*
Δομημένη Γλώσσα Ερωτημάτων	*SQL*
Εγγενείς	*Intrinsic*
Ελάχιστη Ακτίνα	*Eps*
Ελάχιστο Περιβάλλον Ορθογώνιο	*Minimum Bounding Rectangle*

Βιβλιογραφία

[BL97] Michael J. A. Berry, Gordon Linoff. *Data Mining Techniques for Marketing, Sales and Customer Support*. Jon Willey & Sons, Inc, 1996.

[BX02] Μιχάλης Βαζιργιάννης και Μαρία Χαλκίδη. *Εξόρυξη Γνώσης από Μεγάλες Βάεις Δεδομένων: Σημειώσεις Μαθήματος*. ΟΠΑ, Τμήμα Πληροφορικής 2002.

[Dun04] Margaret H. Dunham. *Data Mining Introductory and Advanced Topics*. New Jersey: Pearson Eduction, 2004.

[DXGH00] Margaret H. Dunham, Yongqiao Xiao, Le Gruenwald and Zalin Hossain. A survey of association rules. Technical report, Southern Methodist University, Department of Computer Science, Technical Report TR00-CSE-8, 2000.

[EFKS98] Martin Ester, Alexander Frommelt, Hans – Peter Kriegel and J"org Sander. Algorithms for characterization and trend detection in spatial databases. *Proceedings of the Fourth International Conference on Knowledge Discovery and data Mining,* pages 44-50, 1998.

[EKS97] Martin Ester, Alexander Frommelt, Hans – Peter Kriegel and J" org Sander. Spatial data Mining: A database approach. *Proceedings of the Fifth International Symposium on Large Spatial Databases (SSD)*, pages 47-66,1997.

[EKX95] Martin Ester, Hans – Peter Kriegel and Xiaowei Xu. Knowledge discovery in large spatial databases: Focusing techniques for efficient class identification. *Proceedings of the Fourth*

International Symposium on Large Spatial Databases (SSD), pages 67-82, 1995.

[EFKS00] Martin Ester, Alexander Frommelt, Hans – Peter Kriegel and J" org Sander. Spatial data Mining: Database primitives, algorithms and efficient dbms support. *Data Mining and Knowledge Discovery*, 4(2/3): 193-216, 2000.

[HCC92] Jiawei Han, Yandong Cai and Nick Cercone. Knowledge discovery in databases: An attribute – oriented approach. *Proceedings of the international Very Large Databases Conference*, pages 547-559, 1992.

[HK03] Jiawei Han and Micheline Kamber. *Data Mining – Concepts and Techniques*. New Jersey: Pearson Eduction, 2003.

[HKT01] Jiawei Han, Micheline Kamber and Anthony K. H. Tung. *Spatial Clustering Methods in Data Mining: A Survey*. Philadelphia: Taylor & Francis 2001.

[Καβ04] Μαρίνος Κάβουρας. *Αρχές Γεωπληροφορικής: Σημειώσεις Μαθήματος.* ΕΜΠ, ΣΑΤΜ, ΔΠΜΣ «ΓΕΩΠΛΗΡΟΦΟΡΙΚΗ» 2004.

[KHS98] Krzysztof Koperski, Jiawei Han and Nebosja Stefanovic. An efficient two-step method for classification of spatial data. *Proceedings of the International Symposium on Spatial data Handling*, pages 45-54, 1998.

[KN96] E. Knorr and R. Ng. Finding aggregate proximity relationships and commonalities in spatial data mining. *IEEE Transactions on Knowledge and data Engineering,* 8(6): 884-897, December 1996.

[LH093] W. Lu. J. Han and B.C. Ooi. Discovery of general Knowledge in large spatial databases. *Proceedings of Far East Workshopon Geographic Information Systems,* pages 275-289, 1993.

[MDS03] Marathon Data Systems. *Εισαγωγή στο ArcGIS - ArcView*. Σημειώσεις Σεμιναρίου, MDS 2003.

[Πατ05] Πατρούμπας Κώστας. *Συγγραφή Κώδικα Αλγορίθμου K - means*, ΣΗΜΜΥ, ΔΠΜΣ «ΓΕΩΠΛΗΡΟΦΟΡΙΚΗ» 2005 - 2006.

[RB99] Bob Reselman and Richard Peasley. *Practical Visual Basic 6*. Αθήνα: Γκιούρδας Εκδοτική 2000.

[Στε03] Εμμανουήλ Στεφανάκης. *Βάσεις Γεωγραφικών Δεδομένων και Συστήματα Γεωγραφικών Πληροφοριών*. Αθήνα: Εκδόσεις Παπασωτηρίου 2003

[SC03] Shashi Shekhar and Sanjay Chawla. *Spatial Databases – A Tour*. New Jersey: Pearson Eduction, 2003.

[WYM97] W. Wang, Yang, R. Muntz. STING: A Statistical Information Grid Approach for Very Large Databases. *VLDB '97*.

[XEKS98] X. Xu, M. Ester, Hans – Peter Kriegel and J. Sander. A distribution based clustering algorithm for mining in large spatial databases. *Proceedings of the IEEE International Conference on Data Engineering*, pages 324-331, 1998.

Δικτυακοί τόποι

www. cis.temple.edu ~ ingargio/cis587/readings/id3-c45.html

www.dblab.aueb.gr/index.php/corporate/conbnt/download/232/868/file/HBV_SSDBMOI.pdf

www.ted.unipi.gr/Uploads/Files/Material/courses/15_1105659031.pdf

http://db.cs.sfu.ca/GeoMiner/survey/html/survey.html