11 Ανάλυση Συστάδων

Σύνοψη

Η Ανάλυση Συστάδων (ΑΣ) (Clustering) είναι μια από τις βασικότερες εργασίες Εξόρυξης Δεδομένων. Στόχος της ΑΣ είναι ο επιμερισμός ενός συνόλου παραδειγμάτων σε συστάδες. Οι συστάδες συγκροτούνται στη βάση της ομοιότητας των μελών τους. Το γεγονός ότι δεν υπάρχει εκ των προτέρων γνώση σχετικά με την ύπαρξη ομάδων χαρακτηρίζει την ΑΣ ως μη επιβλεπόμενη μάθηση. Ένας τρόπος καθορισμού του βαθμού ομοιότητας δύο παρατηρήσεων είναι με τη χρήση της απόστασης τους. Οι παρατηρήσεις θεωρούνται ως σημεία σε έναν πολυδιάστατο χώρο. Η απόσταση τους σε αυτόν τον χώρο αποτελεί το μέτρο της ομοιότητας τους. Εάν όλα τα γνωρίσματα είναι αριθμητικά, τότε για τον υπολογισμό της ανομοιότητας χρησιμοποιείται η Ευκλείδεια απόσταση ή κάποια παραλλαγή της, όπως η απόσταση Manhattan ή η απόσταση Minkowski. Συνήθως, σε μια βάση δεδομένων, εκτός από τα αριθμητικά πεδία, υπάρχουν και δυαδικά, ονομαστικά και διατακτικά πεδία. Για κάθε έναν από αυτούς τους τύπους των γνωρισμάτων έχουν καθοριστεί τρόποι υπολογισμού της απόστασης. Οι τρόποι αυτοί παρουσιάζονται αναλυτικά. Επίσης, ορίζεται ο υπολογισμός της απόστασης για περιπτώσεις, όπου οι παρατηρήσεις περιλαμβάνουν γνωρίσματα διαφορετικών τύπων, δηλαδή αριθμητικά, δυαδικά, ονομαστικά κλπ.
Υπάρχει μια μεγάλη ποικιλία μεθόδων ΑΣ. Οι μέθοδοι αυτές χωρίζονται σε Ιεραρχικές, Διαχωριστικές, μεθόδους βασισμένες στην πυκνότητα, μεθόδους πλέγματος και μεθόδους βασισμένες σε μοντέλα. Οι Ιεραρχικές μέθοδοι δημιουργούν μια ιεραρχία επιπέδων, κάθε ένα από τα οποία περιλαμβάνει ένα σύνολο συστάδων. Η επιλογή του κατάλληλου συνόλου συστάδων εναπόκειται στον χρήστη. Η ιεραρχία των επιπέδων και οι αντίστοιχες συστάδες αναπαριστώνται γραφικά με τη χρήση δενδρογραμμάτων. Οι Ιεραρχικές μέθοδοι υποδιαιρούνται σε συσσωρευτικές, οι οποίες δημιουργούν την ιεραρχία μέσα από μια διαδικασία διαδοχικών συγχωνεύσεων, και σε διαιρετικές, οι οποίες δημιουργούν την ιεραρχία μέσω διαδοχικών διασπάσεων. Για τη συγχώνευση ή διάσπαση συστάδων απαιτείται καθορισμός της απόστασης τους. Έχουν προταθεί διάφοροι τρόποι μέτρησης της απόστασης των συστάδων. Οι βασικότεροι από αυτούς είναι η μέθοδος της Απλής Σύνδεσης, η μέθοδος της Πλήρους Σύνδεσης, η Σύνδεση Μέσου Όρου, η μέθοδος Ward κλπ. Οι μέθοδοι αυτές παρουσιάζονται αναλυτικά. Στη Διαχωριστική ΑΣ, τα αντικείμενα επιμερίζονται σε k συστάδες. Τυπικά, ο αριθμός των συστάδων προκαθορίζεται από τον χρήστη. Στη συνέχεια εφαρμόζεται μια επαναληπτική διαδικασία, κατά την οποία τα αντικείμενα μετακινούνται από μια συστάδα σε μια άλλη. Η ποιότητα της κάθε λύσης ενδεχόμενων συστάδων μετράται με τη βοήθεια ενός κριτηρίου. Σε κάθε επανάληψη, και με τη μετακίνηση των σημείων, η τιμή του κριτηρίου μειώνεται. Ο πιο γνωστός αλγόριθμος Διαχωριστικής ΑΣ είναι ο k-Means. Στις βασισμένες στην πυκνότητα μεθόδους, ελέγχεται η πυκνότητα των αντικειμένων, και η συστάδα επεκτείνεται, όσο η γειτονιά των παρακείμενων σημείων διαθέτει την απαιτούμενη πυκνότητα. Οι μέθοδοι πλέγματος επιμερίζουν τον χώρο των δεδομένων σε διακριτά κελιά, τα οποία συγκροτούν ένα πλέγμα, και η αναζήτηση των συστάδων γίνεται στα κελιά του πλέγματος. Τέλος, στις βασισμένες στα μοντέλα μεθόδους, γίνεται χρήση μοντέλων, με στόχο τη βελτιστοποίηση της προσαρμογής ανάμεσα στα δεδομένα και τα μοντέλα. Μια πολύ διαδεδομένη μέθοδος αυτής της κατηγορίας είναι οι Αυτοοργανούμενοι Χάρτες. Οι Αυτοοργανούμενοι Χάρτες είναι ένας ειδικός τύπος Νευρωνικού Δικτύου με ένα επίπεδο. Οι νευρώνες είναι χωρικά διατεταγμένοι σε ένα πλέγμα, και περιέχουν ένα διάνυσμα ίδιων διαστάσεων με τα δεδομένα εισόδου. Με κατάλληλη, μη επιβλεπόμενη μάθηση, όμοια δεδομένα εισόδου αντιστοιχίζονται με γειτονικούς νευρώνες του πλέγματος. Οι Αυτοοργανούμενοι Χάρτες παρουσιάζονται αναλυτικά στο υποκεφάλαιο 11.6. Στο τέλος του παρόντος κεφαλαίου γίνεται μια σύντομη αναφορά στις εφαρμογές της ΑΣ στη σύγχρονη επιχείρηση

Προηγούμενη Γνώση

Το παρόν Κεφάλαιο εισάγει τον αναγνώστη στη θεματική ενότητα της Ανάλυσης Συστάδων, η οποία είναι αυτόνομη και για τον λόγο αυτό δεν απαιτούνται ιδιαίτερες προηγούμενες γνώσεις. Ωστόσο, για την καλύτερη κατανόηση των περιεχομένων θα συνιστούσαμε την προηγούμενη ανάγνωση του Κεφαλαίου 6, το οποίο αποτελεί εισαγωγή στην Εξόρυξη Δεδομένων. Η θεματική ενότητα της Ανάλυσης Συστάδων είναι πολύ εκτεταμένη και φυσικά είναι αδύνατο να καλυφθεί στα πλαίσια ενός κεφαλαίου. Υπάρχουν πολλά εξειδικευμένα βιβλία, τα οποία ασχολούνται αποκλειστικά με την Ανάλυση Συστάδων. Στα βιβλία αυτά, ο αναγνώστης μπορεί να βρει πρόσθετες μεθόδους ΑΣ, καθώς και παρουσίαση τεχνικών για την ομαδοποίηση δεδομένων ειδικού τύπου, όπως κειμένων, δικτύων, πολυμέσων κλπ. Ενδεικτικά αναφέρουμε το βιβλίο των Aggarwal and Reddy (2014).

11.1 Εισαγωγή

Η Ανάλυση Συστάδων (ΑΣ) (Clustering) είναι μια από τις βασικότερες εργασίες Εξόρυξης Δεδομένων. Σε γενικές γραμμές, η ΑΣ αφορά την ένταξη οντοτήτων σε ομοειδείς ομάδες. Η δραστηριότητα αυτή είναι εγγενής στους ανθρώπους, και εκτελείται αυθόρμητα από την παιδική τους ηλικία. Ένας άνθρωπος σε πρωτόγονες συνθήκες, αλλά με σχετική εμπειρία, κατανοεί αυθόρμητα ομάδες, όπως δένδρα, πουλιά κλπ. (Kruskal, 1977). Στην επιστημονική ΑΣ, οι ομάδες εξάγονται βάσει αλγορίθμων από τα δεδομένα.

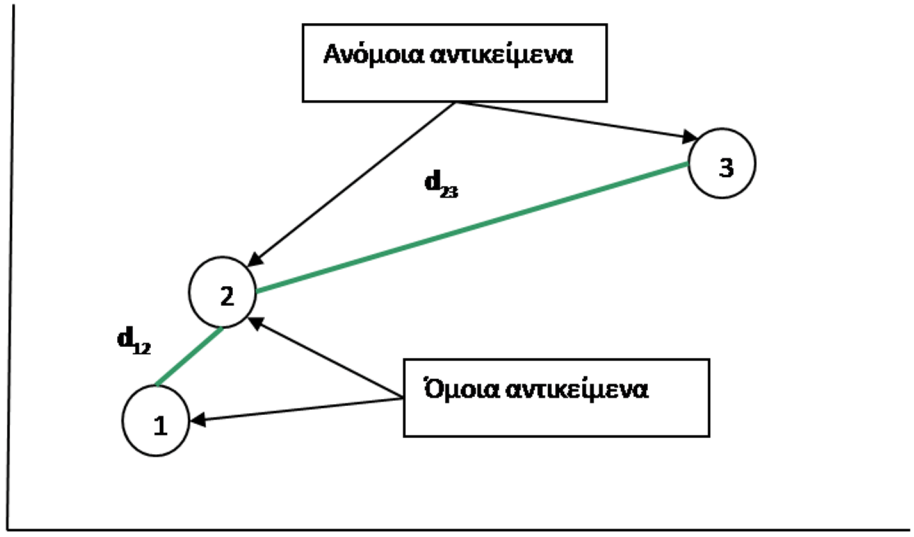
Στόχος της ΑΣ είναι ο επιμερισμός ενός συνόλου παραδειγμάτων σε υποσύνολα. Τα υποσύνολα καλούνται συστάδες. Για τον επιμερισμό, καθοριστικό ρόλο παίζει η **ομοιότητα**. Τα παραδείγματα μιας συστάδας «μοιάζουν» μεταξύ τους, ενώ «δεν μοιάζουν» με τα παραδείγματα των άλλων συστάδων. Ένα σχετικά συγγενές αντικείμενο είναι η Κατηγοριοποίηση, η οποία στοχεύει στην πρόβλεψη της κατηγορίας κάθε παρατήρησης. Όμως οι διαφορές ανάμεσα στην ΑΣ και την Κατηγοριοποίηση είναι πολλές. Στην Κατηγοριοποίηση, οι κατηγορίες είναι γνωστές εκ των προτέρων. Στα δεδομένα υπάρχει ένα γνώρισμα, το γνώρισμα της κλάσης, στο οποίο καταγράφεται η κατηγορία της εκάστοτε παρατήρησης. Οι αλγόριθμοι μοντελοποιούν τις σχέσεις ανάμεσα στο γνώρισμα της κλάσης και στα υπόλοιπα γνωρίσματα. Η Κατηγοριοποίηση είναι μια μορφή **εκπαίδευσης μέσω παραδειγμάτων** (learning by examples). Το γεγονός ότι υπάρχει εκ των προτέρων γνώση σχετικά με τις κατηγορίες, και ότι η γνώση αυτή καθοδηγεί τη διαδικασία εκπαίδευσης, χαρακτηρίζει την Κατηγοριοποίηση ως επιβλεπόμενη μάθηση (supervised learning). Στην ΑΣ δεν υπάρχει κάποιο γνώρισμα στο οποίο να καταγράφεται η κλάση των παραδειγμάτων, και οι συστάδες δεν είναι γνωστές εκ των προτέρων. Αντιθέτως, το ζητούμενο είναι να εντοπιστούν συστάδες και να ενταχθούν τα παραδείγματα στην κατάλληλη συστάδα. Οι συστάδες συγκροτούνται στη βάση της ομοιότητας των μελών τους. Για τον λόγο αυτό, η ΑΣ θεωρείται μια μορφή **εκπαίδευσης μέσω παρατήρησης** (learning by observation). Επίσης, το γεγονός ότι δεν υπάρχει εκ των προτέρων γνώση χαρακτηρίζει την ΑΣ ως **μη επιβλεπόμενη μάθηση** (unsupervised learning). Ο κύριος σκοπός της Κατηγοριοποίησης είναι η διατύπωση προβλέψεων (predictive), ενώ ο κύριος σκοπός της ΑΣ είναι περιγραφικός (descriptive). Σε διαδικαστικό επίπεδο, η ΑΣ αντιμετωπίζει όλα τα γνωρίσματα ισότιμα και τα χρησιμοποιεί για τον υπολογισμό της ομοιότητας των παρατηρήσεων. Αντιθέτως, η Κατηγοριοποίηση χρησιμοποιεί τα υπόλοιπα γνωρίσματα για να προβλέψει τις τιμές του γνωρίσματος της κλάσης.

Στα πλαίσια της Εξόρυξης Δεδομένων, η ΑΣ έχει πολλαπλή χρησιμότητα. Ως αυτόνομη αναλυτική εργασία, επιτρέπει στον αναλυτή να επιμερίσει τα δεδομένα σε ομάδες ομοειδών παρατηρήσεων. Ακολούθως, ο αναλυτής μπορεί να επικεντρωθεί στην εκάστοτε ομάδα, να αναγνωρίσει τα κοινά χαρακτηριστικά της, και να εξάγει γνώση χρήσιμη για τη λήψη αποφάσεων. Η πιο γνωστή εφαρμογή της ΑΣ στις επιχειρηματικές διαδικασίες είναι στη διαφήμιση, και ειδικότερα για την τμηματοποίηση της αγοράς. Ο όρος τμηματοποίηση της αγοράς περιγράφει τον επιμερισμό των καταναλωτών σε ομάδες με όμοια καταναλωτική συμπεριφορά. Η τμηματοποίηση της αγοράς είναι κεφαλαιώδους σημασίας για το μάρκετινγκ. Οι διαφημίσεις μαζικής απεύθυνσης έχουν υψηλό κόστος και μικρό ποσοστό ανταπόκρισης. Με τον εντοπισμό ομάδων όμοιων καταναλωτών μπορούν να σχεδιαστούν διαφημιστικές εκστρατείες προσαρμοσμένες στα ιδιαίτερα χαρακτηριστικά της κάθε ομάδας. Η στοχευμένη σε συγκεκριμένες ομάδες διαφήμιση κοστίζει λιγότερο και επιτυγχάνει σημαντικά υψηλότερα ποσοστά ανταπόκρισης των καταναλωτών. Πέρα από την αξία της ως αυτόνομο εργαλείο ανάλυσης, η ΑΣ μπορεί να συνδυαστεί με άλλες εργασίες Εξόρυξης Δεδομένων και να αποτελέσει στάδιο προεπεξεργασίας. Χάρη στην ικανότητα των αλγορίθμων της να ομαδοποιούν τις παρατηρήσεις σύμφωνα με την ομοιότητα τους, μπορεί να χρησιμοποιηθεί για τον εντοπισμό παρατηρήσεων με ακραίες τιμές (outliers) (Ng & Han, 1994; Shekhar & Chawla, 2003). Οι ακραίες παρατηρήσεις θα απομακρυνθούν από το σύνολο δεδομένων, ώστε να προκύψει ένα βελτιωμένο σύνολο εκπαίδευσης. Επίσης, οι συστάδες, οι οποίες θα προκύψουν, μπορούν να θεωρηθούν κατηγορίες. Σε ακόλουθο στάδιο, μπορεί να εκτελεστεί κατηγοριοποίηση για την ανάπτυξη μοντέλων ικανών να προβλέπουν την κατηγορία. Συνδυασμός μεθόδων ΑΣ και Κατηγοριοποίησης μπορεί να αποφέρει υβριδικούς κατηγοριοποιητές.

11.2 Ομοιότητα και απόσταση

Εφόσον στην ΑΣ οι παρατηρήσεις ομαδοποιούνται σύμφωνα με την ομοιότητα τους, είναι φανερό ότι ένα από τα βασικότερα ζητήματα είναι ο καθορισμός μέτρων ομοιότητας. Ένας τρόπος καθορισμού του βαθμού ομοιότητας δύο παρατηρήσεων είναι με τη χρήση της **απόστασης** τους. Ας θεωρήσουμε αρχικά μια απλή περίπτωση, όπου οι παρατηρήσεις αποτελούνται από δύο μόνο γνωρίσματα X και Y και ότι και τα δύο γνωρίσματα παίρνουν αριθμητικές τιμές. Κάθε παρατήρηση μπορεί να αναπαρασταθεί στον δισδιάστατο χώρο X,Y ως ένα σημείο. Δύο σημεία, τα οποία βρίσκονται κοντά στον δισδιάστατο χώρο, θεωρούνται όμοια, ενώ δύο σημεία,

τα οποία βρίσκονται μακριά στον δισδιάστατο χώρο, θεωρούνται ανόμοια. Στο Σχήμα 11.1 απεικονίζονται τρία σημεία στον δισδιάστατο χώρο. Τα σημεία 1 και 2 θεωρούνται όμοια, ενώ τα σημεία 2 και 3 θεωρούνται ανόμοια.



Σχήμα 11.1 *Ομοιότητα με χρήση απόστασης*

Εάν οι παρατηρήσεις έχουν *n* γνωρίσματα, τότε θεωρούνται σημεία στον χώρο των *n* διαστάσεων, και η ομοιότητα τους υπολογίζεται από την απόσταση τους σε αυτόν τον χώρο. Για τον υπολογισμό της απόστασης υπάρχει διαφοροποίηση ανάλογα με το εάν τα γνωρίσματα περιέχουν αριθμητικές, δυαδικές, ή ονομαστικές τιμές.

11.2.1 Απόσταση με αριθμητικά γνωρίσματα

Εάν όλα τα γνωρίσματα των παρατηρήσεων έχουν αριθμητικές τιμές, τότε ως μέτρο ομοιότητας δύο παρατηρήσεων x_a και x_b μπορεί να χρησιμοποιηθεί η Ευκλείδεια απόσταση. Θεωρούμε ότι οι παρατηρήσεις έχουν *n* γνωρίσματα. Η απόσταση μεταξύ των σημείων x_a και x_b συμβολίζεται ως $d(x_a,x_b)$. Η Ευκλείδεια απόσταση των σημείων x_a και x_b δίνεται από την Εξίσωση 11.1

$$d(x_a, x_b) = \sqrt{\sum_{j=1}^{n} (x_{aj} - x_{bj})^2}$$

(11.1)

όπου xaj είναι η τιμή της μεταβλητής *j* της παρατήρησης x_a.

Η Ευκλείδεια απόσταση είναι η πιο διαδεδομένη, ωστόσο δεν είναι η μοναδική. Μια παραλλαγή της, η οποία χρησιμοποιείται συχνά, είναι η απόσταση Manhattan. Η απόσταση Manhattan ορίζεται από την Εξίσωση 11.2

$$d(x_a, x_b) = \sum_{j=1}^{n} |x_{aj} - x_{bj}|$$

(11.2)

Γενίκευση της Ευκλείδειας απόστασης και της απόστασης Manhattan είναι η απόσταση Minkowski, η

οποία ορίζεται από την Εξίσωση 11.3.

$$d(x_a, x_b) = \sqrt[q]{\sum_{j=1}^{n} |x_{aj} - x_{bj}|^q}$$

(11.3)

Η Ευκλείδεια απόσταση ταυτίζεται με την Minkowski για *q=2*, ενώ η απόσταση Manhattan ταυτίζεται με την Minkowski για *q=1*.

Από τους παραπάνω ορισμούς προκύπτει ότι η απόσταση ενός σημείου από τον εαυτό του είναι ίση με μηδέν, και ότι η απόσταση είναι ένας θετικός αριθμός. Επίσης, η απόσταση μεταξύ δύο σημείων είναι συντομότερη από οποιαδήποτε άλλη διαδρομή, η οποία συνδέει τα δύο αυτά σημεία μέσω ενός τρίτου σημείου.

Η Ευκλείδεια απόσταση δεν επηρεάζεται από την προσθήκη νέων παρατηρήσεων και αποδίδει καλά όταν στα δεδομένα υπάρχουν συμπαγείς ή απομονωμένες συστάδες (Mao & Jain, 1996). Ένα μειονέκτημα της Ευκλείδειας απόστασης είναι ότι μεταβολή των μονάδων μέτρησης μιας μεταβλητής (πχ έκφραση μιας απόστασης από χιλιόμετρα σε μέτρα ή μετατροπή ενός χρηματικού ποσού από ευρώ σε γιεν) επηρεάζει σημαντικά την απόσταση, και μπορεί να οδηγήσει σε σημαντικά διαφορετικές συστάδες. Επίσης, οι μεταβλητές, οι οποίες παίρνουν μεγαλύτερες τιμές ή που παρουσιάζουν μεγάλες διαφορές τιμών μεταξύ των παρατηρήσεων, επηρεάζουν δυσανάλογα την απόσταση. Ένας τρόπος αντιμετώπισης αυτού του προβλήματος είναι η κανονικοποίηση των τιμών, με χρήση κάποιας τεχνικής όπως η Z-Score. Η τεχνική Z-Score και άλλες τεχνικές κανονικοποίησης παρουσιάζονται στο Κεφάλαιο 7.

Στον υπολογισμό της Ευκλείδειας απόστασης όλα τα γνωρίσματα θεωρούνται ισότιμα. Ωστόσο, ο αναλυτής πιθανόν να επιθυμεί να προσδώσει ιδιαίτερη βαρύτητα σε ορισμένα από αυτά. Αυτό μπορεί να επιτευχθεί με εκχώρηση συντελεστών βαρύτητας στα γνωρίσματα, έτσι ώστε η διαφορά των τιμών δυο αντικειμένων για το συγκεκριμένο γνώρισμα να πολλαπλασιάζεται με τον συντελεστή βαρύτητας του γνωρίσματος. Με τη χρήση συντελεστών βαρύτητας, η Ευκλείδεια απόσταση υπολογίζεται σύμφωνα με την Εξίσωση 11.4

$$d(x_a, x_b) = \sqrt{\sum_{j=1}^{n} w_j * (x_{aj} - x_{bj})^2}$$

(11.4)

όπου w_j ο συντελεστής βαρύτητας του γνωρίσματος j.

Η ύπαρξη γραμμικής συσχέτισης μεταξύ των γνωρισμάτων μπορεί να προκαλέσει στρεβλώσεις στον υπολογισμό της απόστασης. Το πρόβλημα αυτό μπορεί να μετριαστεί με τη χρήση της απόστασης Mahalanobis. Η απόσταση Mahalanobis λαμβάνει υπόψη της τις συσχετίσεις μεταξύ των γνωρισμάτων, και δεν επηρεάζεται από την κλίμακα μέτρησης των μεταβλητών. Ο υπολογισμός της γίνεται σύμφωνα με την Εξίσωση 11.5

$$d(x_a, x_b) = (x_a - x_b)^T S^{-1} (x_a - x_b)$$

(11.5)

όπου *S* ο πίνακας συνδιασποράς.

11.2.2 Απόσταση με δυαδικά γνωρίσματα

Η Ευκλείδεια απόσταση είναι ένα κατάλληλο μέτρο ομοιότητας, όταν οι παρατηρήσεις αποτελούνται από γνωρίσματα αριθμητικών τιμών. Σε μια βάση δεομένων όμως υπάρχουν και πεδία άλλων τύπων, όπως δυαδικά πεδία και ονομαστικά πεδία. Για παρατηρήσεις με γνωρίσματα άλλων τύπων έχουν προταθεί άλλα, πιο κατάλληλα μέτρα ομοιότητας.

Τα δυαδικά (binary) γνωρίσματα δέχονται δύο δυνατές τιμές, την τιμή 0 και την τιμή 1. Ένα δυαδικό γνώρισμα μπορεί να αναπαριστά μια πληροφορία, όπου οι δύο δυνατές τιμές κωδικοποιούν δύο καταστάσεις ίσης αξίας ή σημασίας. Παράδειγμα τέτοιου γνωρίσματος είναι το πεδίο «Φύλο», όπου το 1 συμβολίζει το «άρρεν»

και το 0 συμβολίζει το «θήλυ». Μια τέτοια δυαδική μεταβλητή ονομάζεται **συμμετρική**. Υπάρχουν όμως δυαδικές μεταβλητές, όπου οι δύο καταστάσεις τις οποίες συμβολίζουν οι τιμές 0 και 1 δεν είναι ισότιμες. Συνήθως τέτοιες δυαδικές μεταβλητές καταγράφουν την ύπαρξη ή την ανυπαρξία συμβάντος (πχ χρεοκοπία επιχείρησης ή μη χρεοκοπία). Κατά κανόνα η ύπαρξη συμβάντος είναι πιο σπάνια και κωδικοποιείται με την τιμή 1. Σε τέτοιου τύπου δυαδικές μεταβλητές, οι δύο καταστάσεις δεν είναι ίσης αξίας. Οι μεταβλητές αυτές καλούνται μη **συμμετρικές**.

Θεωρούμε ένα σύνολο δεδομένων που περιέχει μόνο δυαδικές μεταβλητές. Τα x_a και x_b είναι δύο αντικείμενα αυτού του συνόλου. Τα x_a και x_b μπορούν να έχουν σε μια μεταβλητή ίδιες τιμές (1 ή 0) ή διαφορετικές τιμές (1 και 0 ή 0 και 1). Η συμφωνία τιμών των δύο παρατηρήσεων στις διάφορες μεταβλητές τους έχει ως ακολούθως:

- *k* είναι το πλήθος των μεταβλητών όπου και το x_a και το x_b έχουν την τιμή 1,
- *l* είναι το πλήθος των μεταβλητών όπου το x_a έχει την τιμή 1, ενώ το x_b έχει την τιμή 0,
- *m* είναι το πλήθος των μεταβλητών όπου το x_a έχει την τιμή 0, ενώ το x_b έχει την τιμή 1,
- *n* είναι το πλήθος των μεταβλητών όπου και το x_a και το x_b έχουν την τιμή 0.

Τα παραπάνω συνοψίζονται στον πίνακα συνάφειας που παρουσιάζεται στο Σχήμα 11.2

		Αντικείμενο x_b		
		1	0	Άθροισμα
Αντικείμενο x_a	1	k	l	k+l
	0	m	n	m+n
	Άθροισμα	k+m	l+n	k+l+m+n

Σχήμα 11.2 *Πίνακας συνάφειας με δυαδικές μεταβλητές*

Εάν οι δυαδικές μεταβλητές είναι συμμετρικές, τότε η απόσταση των αντικειμένων x_a και x_b δίνεται από τον συντελεστή **simple matching** (simple matching coefficient), ο οποίος ορίζεται με την Εξίσωση 11.6.

$$d(x_a x_b) = \frac{l + m}{k + l + m + n} \tag{11.6}$$

Εάν οι δυαδικές μεταβλητές δεν είναι συμμετρικές, η σύμπτωση τιμών ίσων με 0 είναι μικρότερης σημασίας. Για τον λόγο αυτό, έχουν προταθεί πιο κατάλληλα μέτρα. Το πιο γνωστό μέτρο είναι ο συντελεστής **Jaccard**, ο οποίος ορίζεται με την Εξίσωση 11.7.

$$d(x_a x_b) = \frac{l + m}{k + l + m} \tag{11.7}$$

11.2.3 Απόσταση με ονομαστικά γνωρίσματα

Ονομαστικά (nominal) καλούνται τα γνωρίσματα τα οποία δέχονται ονομαστικές τιμές, δηλαδή λέξεις. Ένα ονομαστικό γνώρισμα μπορεί να λάβει ένα πεπερασμένο πλήθος τιμών. Οι τιμές αυτές δεν υποδηλώνουν κάποια μορφή εσωτερικής ιεράρχησης, όπως συμβαίνει στα διατακτικά ονομαστικά γνωρίσματα. Παράδειγμα ονομαστικού γνωρίσματος είναι η κατηγορία προϊόντος (τρόφιμα, είδη ένδυσης, είδη καλλωπισμού κλπ.).

Θεωρούμε δύο αντικείμενα x_a και x_b με *n* ονομαστικά γνωρίσματα. Τα δύο αντικείμενα έχουν σε *m* γνωρίσματα τις ίδιες τιμές. Η απόσταση των δύο αντικειμένων μπορεί να υπολογιστεί με τρόπο αντίστοιχο με τον συντελεστή simple matching (Εξίσωση 11.8)

$$d(x_a x_b) = \frac{n-m}{n}$$

(11.8)

Ένας εναλλακτικός τρόπος υπολογισμού της απόστασης είναι με την εισαγωγή ψευδομεταβλητών (dummy variables). Για μια ονομαστική μεταβλητή, η οποία μπορεί να δεχτεί *k* διαφορετικές τιμές, δημιουργούμε *k* ψευδομεταβλητές, μία για κάθε δυνατή τιμή. Για παράδειγμα, αν η κατηγορία προϊόντος μπορεί να δεχτεί μόνο τις τρείς τιμές που αναφέρθηκαν προηγουμένως, τότε δημιουργούμε μια μεταβλητή «Τρόφιμα», μια μεταβλητή «Είδη Ένδυσης» και μια μεταβλητή «Είδη καλλωπισμού». Εάν ένα αντικείμενο έχει μια συγκεκριμένη τιμή στην ονομαστική μεταβλητή, τότε η αντίστοιχη ψευδομεταβλητή παίρνει την τιμή 1, ενώ οι υπόλοιπες ψευδομεταβλητές την τιμή 0. Για παράδειγμα, αν ένα αντικείμενο έχει στην ονομαστική μεταβλητή «Κατηγορία Προϊόντος» την τιμή «Τρόφιμα», τότε η ψευδομεταβλητή «Τρόφιμα» θα πάρει την τιμή 1, ενώ οι ψευδομεταβλητές «Είδη Ένδυσης» και «Είδη καλλωπισμού» θα πάρουν την τιμή 0. Με τον τρόπο αυτό, μια ονομαστική μεταβλητή μετατρέπεται σε πολλές δυαδικές μεταβλητές. Μετά τη μετατροπή, ο υπολογισμός της απόστασης μπορεί να γίνει με τον τρόπο που περιγράφτηκε για τα γνωρίσματα δυαδικών τιμών.

11.2.4 Απόσταση με διατακτικά γνωρίσματα

Τα διατακτικά (ordinal) γνωρίσματα δέχονται τιμές, οι οποίες υποδηλώνουν μια θέση σε μια διάταξη ή σειρά. Ένα παράδειγμα διατακτικών τιμών είναι οι δείκτες αξιολόγησης της πιστοληπτικής ικανότητας, τους οποίους εκδίδουν οι οίκοι αξιολόγησης. Η πιστοληπτική ικανότητα ενός φορέα βαθμολογείται με μια τιμή της μορφής AAA, AA, A, BBB, BB, B, CCC, CC, C, D(edault) ή κάποια παραλλαγή της. Οι τιμές είναι λεκτικές, δηλώνουν όμως μια θέση σε μια ποιοτική διαβάθμιση.

Το γεγονός ότι οι διατακτικές τιμές υποδηλώνουν μια σειρά μας επιτρέπει να τις χειριστούμε σαν αριθμητικές τιμές. Αν ένα διατακτικό γνώρισμα δέχεται n τιμές, τότε η τιμή που δηλώνει τη χαμηλότερη θέση στη σειρά μπορεί να αντικατασταθεί με τον αριθμό 1, η επόμενη τιμή με τον αριθμό 2, μέχρι την τελευταία, η οποία θα αντικατασταθεί με τον αριθμό *n*. Η προσέγγιση αυτή έχει το μειονέκτημα ότι μεταβλητές με πολλές διατακτικές τιμές μπορούν να δημιουργήσουν μεγάλες διαφορές μεταξύ αντικειμένων, και οι διαφορές αυτές να επηρεάσουν δυσανάλογα την απόσταση. Για τον λόγο αυτό, οι αριθμητικές τιμές κανονικοποιούνται και ανάγονται στο διάστημα [0,0..1,0]. Ο μετασχηματισμός των τιμών γίνεται σύμφωνα με την Εξίσωση 11.9

$$m_{new} = \frac{m-1}{n-1}$$

(11.9)

όπου m_{new} είναι η νέα τιμή, *m* είναι η τιμή πριν την κανονικοποίηση και *n* είναι το πλήθος δυνατών τιμών της διατακτικής μεταβλητής. Στον πίνακα 11.1 παρουσιάζεται ο μετασχηματισμός των τιμών για τους δείκτες πιστοληπτικής ικανότητας

Διατακτικές τιμές	Αριθμητικές τιμές	Κανονικοποιημένες τιμές
AAA	10	1,00
AA	9	0,89
A	8	0,78
BBB	7	0,67
BB	6	0,56
B	5	0,44
CCC	4	0,33
CC	3	0,22
C	2	0,11
D	1	0,00

Πίνακας 11.1 *Μετασχηματισμός διατακτικών τιμών δείκτη πιστοληπτικής ικανότητας*

Μετά τον μετασχηματισμό των διατακτικών τιμών και την αντιστοίχιση τους σε αριθμητικές τιμές της περιοχής [0,0..1,0], μπορεί να γίνει υπολογισμός της απόστασης δύο αντικειμένων με τη χρήση της Ευκλείδειας απόστασης ή κάποιας παραλλαγής της.

11.2.5 Απόσταση με μεικτών τύπων γνωρίσματα

Όλοι οι υπολογισμοί αποστάσεων, οι οποίοι αναφέρθηκαν μέχρι αυτό το σημείο, θεωρούν ότι όλα τα γνωρίσματα είναι του ιδίου τύπου. Σε πραγματικές βάσεις δεδομένων όμως τα πεδία είναι διαφόρων τύπων, δηλαδή αριθμητικά, δυαδικά, ονομαστικά κλπ. Για τον υπολογισμό της απόστασης αντικειμένων με διάφορους τύπους γνωρισμάτων μπορεί να γίνει συνδυασμός των προηγούμενων τεχνικών.

Η απόσταση δύο αντικειμένων x_a και x_b με n γνωρίσματα διαφόρων τύπων μπορεί να υπολογιστεί σύμφωνα με την Εξίσωση 11.10.

$$d(x_a, x_b) = \frac{\sum_{j=1}^{n} \delta_{abj} \Delta_{abj}}{\sum_{j=1}^{n} \delta_{abj}}$$

(11.10)

Το δ_{abj} παίρνει τιμές ως ακολούθως:

- Τιμή = 0 εάν η τιμή του x_a (x_{aj}) ή του x_b (x_{bj}) στη μεταβλητή j λείπει.
- Τιμή = 0 εάν η μεταβλητή j είναι μη συμμετρική και η τιμή των x_a και x_b στη μεταβλητή j είναι ίση με 0 ($x_{aj} = x_{bj} = 0$).
- Τιμή = 1 σε οποιαδήποτε άλλη περίπτωση.

Ο υπολογισμός της τιμής του Δ_{abj} εξαρτάται από τον τύπο της μεταβλητής j:

- Εάν η μεταβλητή j είναι δυαδική ή ονομαστική το Δ_{abj} παίρνει την τιμή 0 εάν $x_{aj} = x_{bj.}$ Διαφορετικά παίρνει την τιμή 1.
- Εάν η μεταβλητή j είναι αριθμητική, τότε το Δ_{abj} υπολογίζεται σύμφωνα με την Εξίσωση 11.11, όπου maxj είναι η μέγιστη τιμή του γνωρίσματος j και min_j είναι η ελάχιστη τιμή του γνωρίσματος j.

$$\Delta_{abj} = \frac{|x_{aj} - x_{bj}|}{max_j - min_j}$$

(11.11)

- Εάν η μεταβλητή j είναι διατακτική, τότε οι τιμές της μετασχηματίζονται και ανάγονται στην περιοχή [0,0..1,0] και ακολούθως το Δ_{abj} υπολογίζεται με τρόπο αντίστοιχο των αριθμητικών μεταβλητών.

11.3 Κατηγορίες Μεθόδων ΑΣ

Η επιστημονική βιβλιογραφία περιλαμβάνει έναν μεγάλο αριθμό διαφορετικών μεθόδων Ανάλυσης Συστάδων. Οι μέθοδοι αυτές παρουσιάζουν σημαντικές διαφορές στις επαγωγικές αρχές τους και στον τρόπο σχηματισμού των συστάδων. Ένας από τους λόγους ύπαρξης αυτής της ποικιλίας μεθόδων είναι το γεγονός ότι δεν υπάρχει ένας αυστηρός ορισμός της έννοιας της συστάδας (Estivill-Castro & Yang, 2000). Οι Han, Kamber and Pei (2011) ορίζουν πέντε κατηγορίες μεθόδων ΑΣ:

- **Ιεραρχικές μέθοδοι**. Οι ιεραρχικές μέθοδοι (hierarchical methods) δημιουργούν μια ιεραρχία από συστάδες. Στο κατώτατο επίπεδο της ιεραρχίας βρίσκονται τα μεμονωμένα αντικείμενα. Στο ανώτατο επίπεδο βρίσκεται μια υπερσυστάδα, η οποία περιλαμβάνει όλα τα αντικείμενα. Κάθε ενδιάμεσο επίπεδο ορίζει ένα σύνολο συστάδων. Η ιεραρχία προκύπτει από μια διαδικασία διαδοχικών διασπάσεων ή συγχωνεύσεων συστάδων. Η επιλογή του κατάλληλου συνόλου συστάδων εναπόκειται στον χρήστη. Αναλυτική παρουσίαση των ιεραρχικών μεθόδων γίνεται στο υποκεφάλαιο 11.4.
- **Διαχωριστικές μέθοδοι**. Οι διαχωριστικές μέθοδοι (partitioning methods) επιμερίζουν τα αντικείμε-

να σε *k* συστάδες. Τυπικά το πλήθος των συστάδων προκαθορίζεται από τον χρήστη. Στις μεθόδους αυτής της κατηγορίας εφαρμόζεται μια επαναληπτική διαδικασία, κατά την οποία τα αντικείμενα μετακινούνται από μια συστάδα σε μια άλλη. Η ποιότητα της κάθε λύσης ενδεχόμενων συστάδων μετράται με τη βοήθεια ενός κριτηρίου. Σε κάθε επανάληψη και με τη μετακίνηση των σημείων, η τιμή του κριτηρίου μειώνεται. Ο πιο γνωστός αλγόριθμος διαχωριστικής ΑΣ είναι ο k-Means. Μέθοδοι και ζητήματα διαχωριστικής ΑΣ παρουσιάζονται στο υποκεφάλαιο 11.5.

- **Μέθοδοι βασισμένες στην πυκνότητα**. Στις βασισμένες στην πυκνότητα μεθόδους (density based methods) ελέγχεται η πυκνότητα των αντικειμένων στον χώρο και δημιουργούνται συστάδες, οι οποίες καλύπτουν τις πυκνές περιοχές. Για κάθε παρατήρηση που ανήκει σε μια συστάδα, η γειτονιά της, η οποία είναι καθορισμένης διαμέτρου, πρέπει να περιλαμβάνει έναν ελάχιστο αριθμό παρατηρήσεων. Η συστάδα συνεχίζει να επεκτείνεται όσο η γειτονιά των παρακείμενων σημείων διαθέτει την απαιτούμενη πυκνότητα. Οι μέθοδοι αυτές μπορούν να δημιουργήσουν συστάδες με μη κυρτά και περίπλοκα σχήματα. Επίσης, είναι ιδιαίτερα ικανές να απομονώνουν τις εξαιρέσεις.
- **Μέθοδοι πλέγματος**. Οι μέθοδοι πλέγματος (grid based methods) επιμερίζουν τον χώρο των δεδομένων σε διακριτά κελιά, τα οποία συγκροτούν ένα πλέγμα. Τα αντικείμενα πλέον αντιπροσωπεύονται από τα κελιά στα οποία ανήκουν. Η αναζήτηση των συστάδων γίνεται στα κελιά του πλέγματος και όχι στα αντικείμενα. Στις μεθόδους πλέγματος ο χρόνος επεξεργασίας εξαρτάται από το πλήθος των κελιών και όχι από το πλήθος των αντικειμένων. Επειδή κατά κανόνα ο αριθμός των κελιών είναι πολύ μικρότερος από τον αριθμό των αντικειμένων, οι μέθοδοι αυτές είναι σημαντικά ταχύτερες. Ένα σημαντικό ζήτημα είναι ο καθορισμός κελιών κατάλληλου μεγέθους.
- **Μέθοδοι βασισμένες σε μοντέλα**. Στις βασισμένες σε μοντέλα μεθόδους (model based methods), όπως υπονοεί το όνομα τους, γίνεται χρήση μοντέλων. Στόχος τους είναι να βελτιστοποιηθεί η προσαρμογή ανάμεσα στα δεδομένα και στα μοντέλα. Το μοντέλο εκπαιδεύεται με μη επιβλεπόμενη μάθηση σχετικά με τη συμμετοχή των παρατηρήσεων σε συστάδες. Μια πολύ διαδεδομένη μέθοδος αυτής της κατηγορίας είναι ένα ειδικός τύπος νευρωνικών δικτύων, που ονομάζονται Αυτοοργανούμενοι Χάρτες (Self Organizing Maps). Οι Αυτοοργανούμενοι Χάρτες παρουσιάζονται στο υποκεφάλαιο 11.7.

11.4 Ιεραρχική Ανάλυση Συστάδων

Η Ιεραρχική ΑΣ συνίσταται σε μια διαδικασία διαδοχικών συγχωνεύσεων ή διασπάσεων συστάδων. Οι σχετικές τεχνικές αντιστοίχως χωρίζονται σε συσσωρευτικές και διαιρετικές.

Οι **συσσωρευτικές** (agglomerative) μέθοδοι αρχικά θεωρούν κάθε ξεχωριστό αντικείμενο ως μια συστάδα. Τα πιο όμοια αντικείμενα επιλέγονται και συγχωνεύονται, δημιουργώντας μια νέα συστάδα. Από τις συστάδες που προκύπτουν, επιλέγονται οι πιο όμοιες και συγχωνεύονται. Η διαδικασία επαναλαμβάνεται μέχρι να ενταχθούν όλα τα αντικείμενα σε μια ενιαία συστάδα. Οι συσσωρευτικές μέθοδοι έχουν ως αφετηριακό σημείο το κατώτερο επίπεδο της ιεραρχίας των διαδοχικών συγχωνεύσεων, και σταδιακά ανέρχονται τα επίπεδα. Υιοθετούν δηλαδή μια προσέγγιση «από κάτω προς τα επάνω» (bottom up).

Οι **διαιρετικές** (divisive) μέθοδοι αρχικά θεωρούν όλα τα αντικείμενα ως μέλη μιας ενιαίας συστάδας. Η αρχική αυτή συστάδα διαιρείται σε δύο υποομάδες. Η διάσπαση γίνεται με τέτοιον τρόπο, ώστε οι υποομάδες οι οποίες θα προκύψουν θα έχουν τη μεγαλύτερη ανομοιότητα. Η διαδικασία των διαδοχικών διασπάσεων επαναλαμβάνεται μέχρι κάθε αντικείμενο να αποτελεί μια ξεχωριστή υποομάδα. Οι διαιρετικές μέθοδοι έχουν αφετηριακό σημείο το ανώτατο επίπεδο της ιεραρχίας και ακολουθούν μια προσέγγιση «από επάνω προς τα κάτω» (top down).

Για την επιλογή των συστάδων δημιουργείται ένας πίνακας ανομοιότητας. Εάν τα δεδομένα περιέχουν *N* σημεία, τότε ο πίνακας είναι διαστάσεων *NXN*. Κάθε εγγραφή του πίνακα είναι ένα μέτρο ανομοιότητας ή απόστασης μεταξύ δύο σημείων. Ο πίνακας ανομοιότητας έχει την ακόλουθη μορφή:

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \dots & \dots & \dots & 0 & \\ d(N,1) & \dots & \dots & d(N,N-1) & 0 \end{bmatrix} \quad (11.12)$$

όπου $d(x_1,x_2)$ είναι η απόσταση μεταξύ των σημείων x_1 και x_2. Εφόσον η απόσταση κάθε σημείου από τον εαυτό του είναι μηδενική ($d(x_i,x_i)=0$) οι εγγραφές της διαγωνίου από επάνω και αριστερά προς κάτω και δεξιά έχουν μηδενικές τιμές. Επειδή η απόσταση μεταξύ δύο σημείων είναι συμμετρική ($d(x_i,x_j)= d(x_j,x_i)$), η διαγώνιος χωρίζει τον πίνακα σε δύο κατοπτρικά μέρη, οπότε διατηρούνται μόνο οι εγγραφές οι οποίες βρίσκονται κάτω από τη διαγώνιο.

Στην Ιεραρχική ΑΣ δημιουργείται μια ιεραρχία, η οποία περιλαμβάνει ένα σύνολο από δυνατές συστάδες. Κάθε επίπεδο της ιεραρχίας περιγράφει ένα συγκεκριμένο τρόπο διαμοιρασμού των αντικειμένων σε συστάδες. Αποτελεί αρμοδιότητα του χρήστη να αποφασίσει πιο είναι το κατάλληλο επίπεδο, το οποίο περιγράφει έναν φυσικό τρόπο διαμοιρασμού των αντικειμένων, δηλαδή ποιες είναι οι συστάδες, οι οποίες είναι επαρκώς όμοιες μεταξύ τους. Εάν στα δεδομένα μας υπάρχουν *N* σημεία, τότε και στις δύο κατηγορίες μεθόδων υπάρχουν *N-1* επίπεδα.

Τα βασικά **πλεονεκτήματα** των Ιεραρχικών Μεθόδων είναι τα ακόλουθα:

- Οι ιεραρχικές μέθοδοι παρουσιάζουν καλή προσαρμοστικότητα. Μπορούν να εντοπίσουν καλά διαχωρισμένες, επιμήκεις και ομόκεντρες συστάδες.
- Δημιουργούν πολλαπλά επίπεδα φωλιασμένων συστάδων και επιτρέπουν στον χρήστη να επιλέξει το επίπεδο που αυτός επιθυμεί.

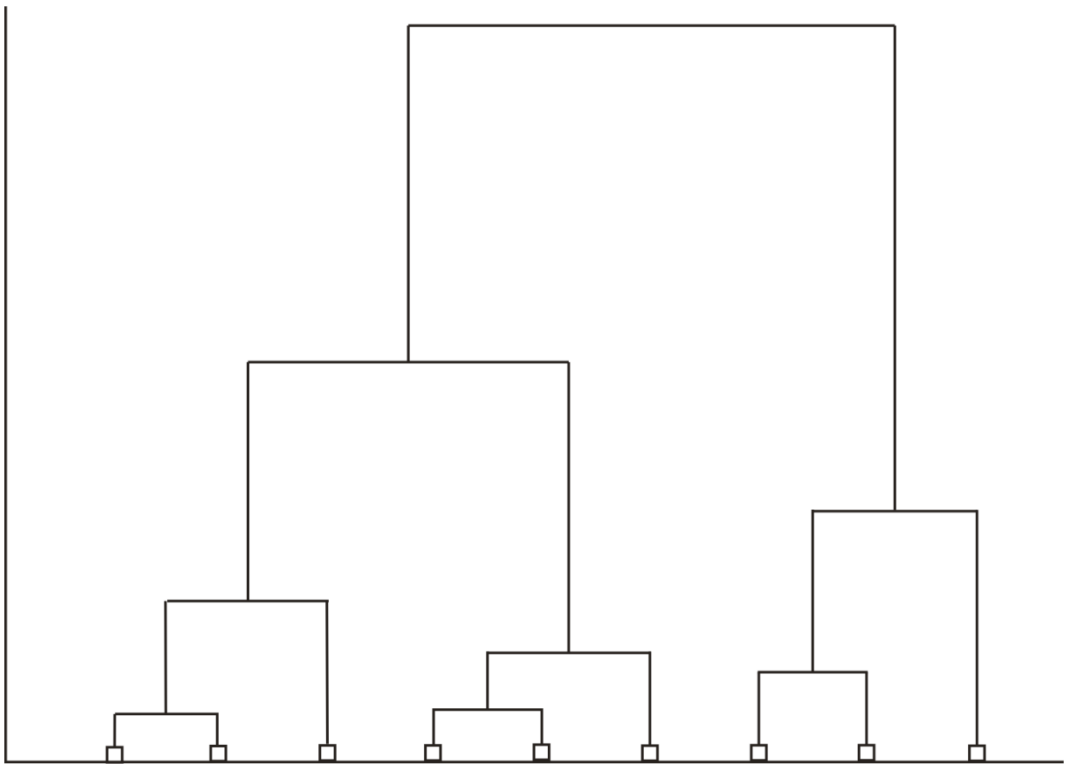
Μειονεκτήματα των Ιεραρχικών μεθόδων είναι τα εξής:

- Κάθε ενέργεια, η οποία πραγματοποιείται σε ένα στάδιο, δεν είναι αντιστρέψιμη. Από τη στιγμή που δύο αντικείμενα ενταχθούν στην ίδια ομάδα, θα παραμείνουν στην ίδια ομάδα, και δεν υπάρχει δυνατότητα να διαχωριστούν αργότερα και να ενταχθούν σε διαφορετικές ομάδες.
- Οι ιεραρχικές μέθοδοι χρειάζεται να ελέγξουν πολλές αποστάσεις, και για τον λόγο αυτό καθυστερούν όταν χρειάζεται να επεξεργαστούν μεγάλο αριθμό αντικειμένων. Το υπολογιστικό κόστος είναι τουλάχιστον $O(N^2)$ όπου *N* το πλήθος των αντικειμένων.

11.4.1 Δενδρογράμματα

Τα Δενδρογράμματα είναι ένας γραφικός τρόπος αναπαράστασης της διαδικασίας των διαδοχικών συγχωνεύσεων ή διασπάσεων. Το Δενδρόγραμμα έχει τη μορφή ανεστραμμένου δένδρου. Στα φύλλα του δένδρου, δηλαδή στο κατώτερο επίπεδο, βρίσκονται τα μεμονωμένα αντικείμενα. Κάθε κόμβος του δένδρου αντιπροσωπεύει μια συστάδα. Επίσης, κάθε κόμβος αποτελεί αφετηρία δύο κλάδων. Στη συσσωρευτική ομαδοποίηση, ένας κόμβος με τους κλάδους και τα τέκνα του συμβολίζει τη συγχώνευση των συστάδων-τέκνων, και τη δημιουργία της συστάδας-γονέα. Στη διαιρετική ομαδοποίηση, ένας κόμβος με τους κλάδους και τα τέκνα του συμβολίζει τη διάσπαση του κόμβου-γονέα, και τη δημιουργία των συστάδων-τέκνων.

Σε όλες τις συσωρευτικές μεθόδους και ορισμένες διαιρετικές μεθόδους, ο βαθμός ανομοιότητας αυξάνεται μονότονα με το επίπεδο. Ο σχεδιασμός του δένδρου γίνεται με τέτοιον τρόπο, ώστε η διαφορά ύψους των επιπέδων να αποτυπώνει την αύξηση της ανομοιότητας. Ο χρήστης μπορεί να χρησιμοποιήσει το Δενδρόγραμμα για να επιλέξει ένα επίπεδο και να αποφασίσει ένα συγκεκριμένο τρόπο διαμοιρασμού των αντικειμένων σε συστάδες.. Ωστόσο, ο χρήστης πρέπει να γνωρίζει ότι διαφορετικές μέθοδοι ιεραρχικής ομαδοποίησης ή και μικρές αλλαγές στα δεδομένα μπορούν να δημιουργήσουν σημαντικά διαφορετικά Δενδρογράμματα. Στο Σχήμα 11.3 παρουσιάζεται ένα Δενδρόγραμμα.



Σχήμα 11.3 *Δενδρόγραμμα Ιεραρχικής Ομαδοποίησης*

11.4.2 Ιεραρχική Συσσωρευτική Ανάλυση Συστάδων

*Ο*ι συσσωρευτικές μέθοδοι εκτελούν διαδοχικές συγχωνεύσεις συστάδων. Σε κάθε επανάληψη οι δύο πλησιέστερες συστάδες συνενώνονται. Ο γενικός αλγόριθμος της Ιεραρχικής Συσσωρευτικής ΑΣ έχει ως ακολούθως:

- Αρχικά, κάθε ένα από τα *N* σημεία θεωρείται ως μια ξεχωριστή συστάδα. Στον πίνακα αποστάσεων καταγράφονται οι αποστάσεις μεταξύ των σημείων.
- Εντοπίζεται στον πίνακα αποστάσεων η μικρότερη τιμή. Η τιμή αυτή είναι η απόσταση των δύο πιο όμοιων συστάδων *U* και *V (d(U,V))*.
- Οι συστάδες *U* και *V* συνενώνονται σε μια ενιαία συστάδα *UV*. Στον πίνακα αποστάσεων, διαγράφονται οι γραμμές και οι στήλες που αντιστοιχούν στις συστάδες *U* και *V*, και προστίθεται μια γραμμή και μια στήλη για τη νέα συστάδα *UV*. Επαναϋπολογίζονται οι αποστάσεις μεταξύ των συστάδων.
- Επαναλαμβάνονται τα βήματα 2 και 3, *N-1* φορές. Σε κάθε επανάληψη καταγράφονται οι συστάδες που συγχωνεύονται καθώς και οι αποστάσεις τους.

Για τον υπολογισμό της εγγύτητας των συστάδων είναι απαραίτητο ένα μέτρο. Έχουν προταθεί διάφοροι τρόποι μέτρησης της απόστασης μεταξύ των συστάδων. Εναλλακτικές μέθοδοι συσωρευτικής ΑΣ διαφοροποιούνται μεταξύ τους, ανάλογα με το μέτρο απόστασης το οποίο εφαρμόζουν. Οι κυριότεροι τρόποι μέτρησης της απόστασης είναι οι ακόλουθοι:

11.4.3 Απλή Σύνδεση

Η μέθοδος της Απλής Σύνδεσης (Simple Linkage) ονομάζεται και μέθοδος του κοντινότερου γείτονα. Σύμφωνα με αυτήν τη μέθοδο, η απόσταση ανάμεσα σε δύο συστάδες είναι η μικρότερη απόσταση από οποιοδήποτε μέλος της πρώτης συστάδας προς οποιοδήποτε μέλος της δεύτερης συστάδας (Sneath & Sokal, 1973). Με απλούστερα λόγια, η απόσταση των συστάδων είναι η απόσταση μεταξύ των δύο πλησιέστερων σημείων τους. Με μαθηματικό τρόπο η απόσταση αυτή ορίζεται από την Εξίσωση 11.13

$$d(C_1, C_2) = \min_{x_a \in C_1, x_b \in C_2} d(x_a, x_b)$$

(11.13)

όπου C_1, C_2 είναι οι δύο συστάδες, x_a, x_b είναι σημεία των συστάδων και $d(x_a,x_b)$ είναι η απόσταση μεταξύ των σημείων x_a και x_b.

Ένα σύνηθες πρόβλημα της απλής σύνδεσης είναι ότι συνενώνει συστάδες, οι οποίες έχουν δύο κοντινά σημεία και πολλά άλλα σημεία που βρίσκονται σε μεγάλες αποστάσεις. Ένα άλλο πρόβλημα είναι ότι μπορεί να προκληθεί η δημιουργία μιας επιμήκους συστάδας, και να προστίθενται διαρκώς νέα σημεία στην «ουρά» της συστάδας. Επίσης, εάν μεταξύ δύο πραγματικών συστάδων υπάρχουν μεμονωμένα σημεία που δημιουργούν μια «γέφυρα», τότε οι συστάδες αυτές θα ενωθούν. Το αποτέλεσμα αυτής της διαδικασίας είναι ότι τα σημεία που βρίσκονται στα δύο άκρα της συστάδας θα απέχουν πολύ μεταξύ τους. Το πρόβλημα αυτό είναι γνωστό ως φαινόμενο της αλυσίδας (chaining phenomenon). Πλεονέκτημα της απλής σύνδεσης είναι ότι μπορεί να εντοπίσει μη ελλειψοειδείς συστάδες.

11.4.4 Πλήρης Σύνδεση

Η μέθοδος της Πλήρους Σύνδεσης (Complete Linkage) ονομάζεται και μέθοδος του μακρινότερου γείτονα. Στη μέθοδο αυτή η λογική υπολογισμού της απόστασης των συστάδων είναι αντίστροφη από αυτήν της Απλής Σύνδεσης. Πιο συγκεκριμένα, η απόσταση μεταξύ δύο συστάδων C_1 και C_2 είναι η μεγαλύτερη απόσταση από οποιοδήποτε μέλος της C_1 προς οποιοδήποτε μέλος της C_2 (King, 1967). Με απλούστερα λόγια, η απόσταση μεταξύ δύο συστάδων είναι η απόσταση ανάμεσα στα δύο πιο απομακρυσμένα σημεία τους. Ο μαθηματικός ορισμός της πλήρους σύνδεσης δίνεται από την Εξίσωση 11.14

$$d(C_1, C_2) = \max_{x_a \in C_1, x_b \in C_2} d(x_a, x_b)$$

(11.14)

Με τη μέθοδο της πλήρους σύνδεσης αποφεύγονται προβλήματα που παρουσιάζονται με την απλή σύνδεση, όπως η δημιουργία επιμηκών συστάδων. Αντιθέτως, η πλήρης σύνδεση τείνει να δημιουργήσει συμπαγείς και σφαιρικές συστάδες με συγκρίσιμη διάμετρο. Αυτό συμβαίνει, γιατί από όλες τις υποψήφιες για συνένωση συστάδες, επιλέγει εκείνες τις δύο, οι οποίες θα δημιουργήσουν τη νέα συστάδα με τη μικρότερη διάμετρο. Το κριτήριο της μεθόδου δεν είναι τοπικό. Ολόκληρη η δομή της εκάστοτε συστάδας θα επηρεάσει την απόφαση για τη συνένωση. Η μέθοδος της πλήρους σύνδεσης ενδείκνυται, όταν γνωρίζουμε ότι αντικείμενα της ίδιας συστάδας είναι δυνατόν να βρίσκονται σε μεγάλες αποστάσεις μεταξύ τους. Ένα μειονέκτημα της μεθόδου είναι η ευαισθησία της στην ύπαρξη αντικειμένων με ακραίες τιμές. Εάν υπάρχει ένα αντικείμενο με ακραίες τιμές σε μια συστάδα, τότε δύσκολα αυτή η συστάδα θα συγχωνευθεί με κάποιαν άλλη.

11.4.5 Σύνδεση Μέσου Όρου

Η μέθοδος της Σύνδεσης Μέσου Όρου (Average Link). Σύμφωνα με αυτήν την προσέγγιση, η απόσταση δύο συστάδων είναι ίση με τη μέση απόσταση όλων των ζευγών αντικειμένων, όπου το πρώτο αντικείμενο ανήκει στην πρώτη συστάδα και το δεύτερο αντικείμενο ανήκει στη δεύτερη συστάδα (Murtagh, 1984). Πρόκειται δηλαδή για τη μέση απόσταση μεταξύ των αντικειμένων των συστάδων. Ο μαθηματικός ορισμός της απόστασης Μέσου Όρου δίνεται από την Εξίσωση 11.15

$$d(C_1, C_2) = \frac{\sum_{x_a \in C_1} \sum_{x_b \in C_2} d(x_a, x_b)}{N_{C_1} N_{C_2}}$$

(11.15)

όπου C_1 και C_2 είναι οι δύο συστάδες, $d(x_a,x_b)$ είναι η απόσταση μεταξύ των αντικειμένων x_a και x_b, και N_{C1}, N_{C2} είναι το πλήθος των συστάδων C_1 και C_2 αντίστοιχα.

Η απόσταση μέσου όρου αποτελεί ενδιάμεση λύση ανάμεσα στην ευαισθησία στα αντικείμενα με ακραίες τιμές της μεθόδου πλήρους σύνδεσης, και στην τάση δημιουργίας επιμηκών συστάδων της απλής σύνδεσης. Χάρη στον υπολογισμό της μέσης απόστασης μεταξύ των ζευγών, δεν δημιουργείται το φαινόμενο της αλυσίδας. Επίσης, εξομαλύνεται η επιρροή των αντικειμένων με ακραίες τιμές. Από άποψη υπολογιστικού κόστους

η μέθοδος είναι ακριβή, καθώς υπολογίζει τις αποστάσεις όλων των δυνατών ζευγών. Ένα άλλο μειονέκτημα είναι ότι μπορεί να διασπάσει υπαρκτές επιμήκεις συστάδες.

11.4.6 Απόσταση Μέσων Σημείων (centroids)

Σύμφωνα με την προσέγγιση αυτή, η απόσταση μεταξύ δύο συστάδων είναι η απόσταση ανάμεσα στα μέσα σημεία των δύο συστάδων. Ο μαθηματικός ορισμός της απόστασης μέσων σημείων δίνεται από την Εξίσωση 11.16

$$d(C_1, C_2) = d(m_1, m_2)$$

(11.16)

όπου m_1, m_2 είναι τα μέσα σημεία των συστάδων C_1 και C_2.

Η απόσταση μέσων σημείων έχει το πλεονέκτημα ότι δεν επηρεάζεται σημαντικά από την ύπαρξη αντικειμένων με ακραίες τιμές.

11.4.7 Μέθοδος Ward

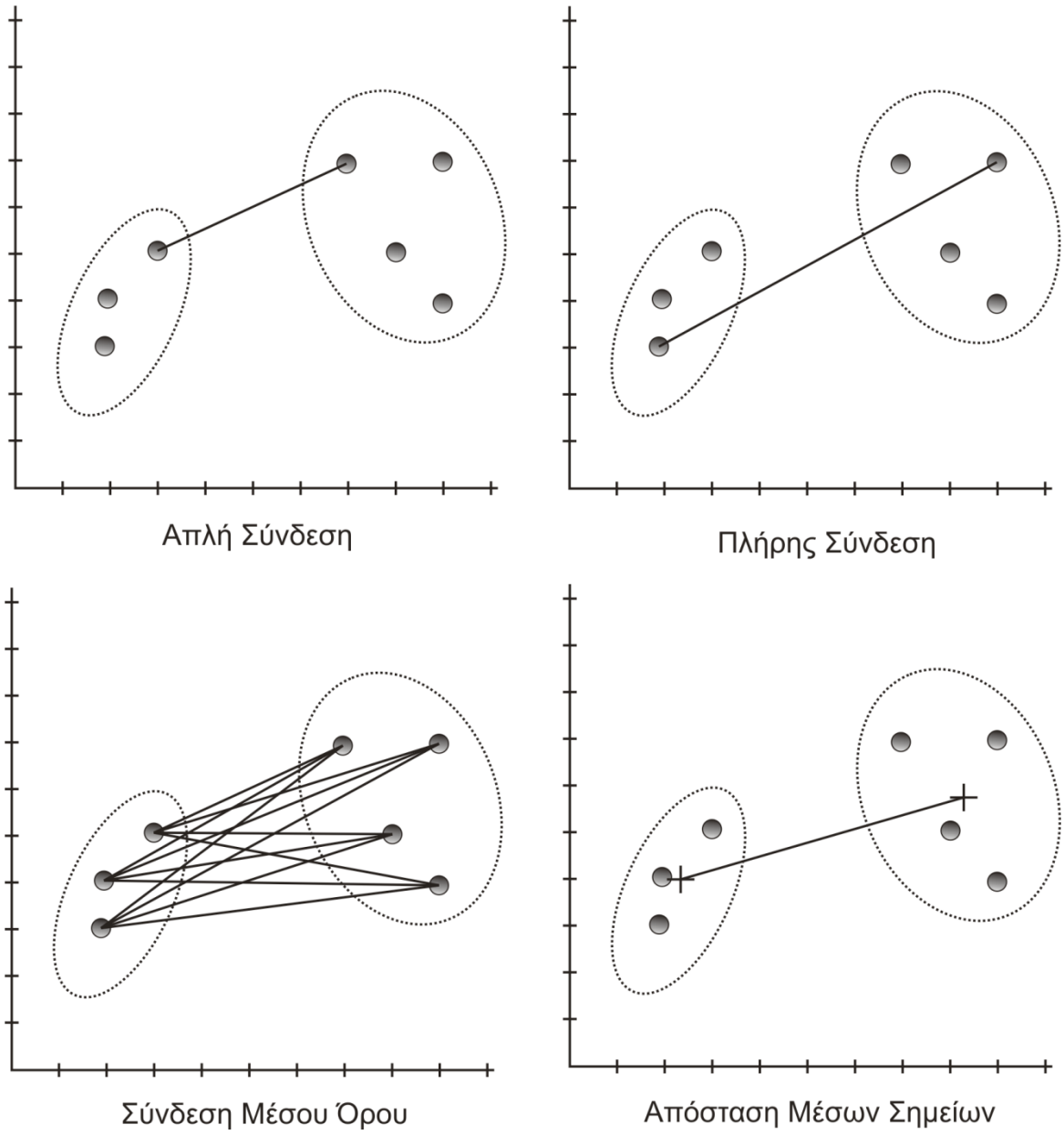
Η μέθοδος του Ward (1963) διαφέρει σημαντικά από τις προηγούμενες μεθόδους, καθώς δεν υπολογίζει κάποια «απόσταση» μεταξύ των συστάδων. Κριτήριο για τη δημιουργία συστάδων είναι η μεγιστοποίηση της ομοιογένειας στο εσωτερικό των συστάδων. Το μέτρο που εφαρμόζεται είναι το άθροισμα του τετραγωνικού σφάλματος, και επιδίωξη της μεθόδου είναι η ελαχιστοποίηση του. Το ίδιο κριτήριο χρησιμοποιείται και από τον αλγόριθμο k-Means, οπότε η μέθοδος Ward μπορεί να θεωρηθεί το ιεραρχικό ανάλογο του k-Means.

Το τετραγωνικό σφάλμα δίνεται από τη Σχέση 11.17

$$E = \sum_{x \in C_i} (x - m_i)^2$$

(11.17)

όπου C_i είναι μια κλάση και mi είναι το μέσο σημείο της. Η μέθοδος, για να συνενώσει δύο συστάδες από συνολικό πλήθος *k* συστάδων, ελέγχει τα δυνατά *k(k-1)/2* ζεύγη συστάδων τα οποία μπορούν να δημιουργηθούν, και επιλέγει το ζεύγος, το οποίο όταν ενωθεί θα μας δώσει τη συστάδα με το ελάχιστο τετραγωνικό σφάλμα. Η μέθοδος του Ward έχει την τάση να παράγει ισοπληθείς ομάδες.



Σχήμα 11.4 *Απόσταση μεταξύ συστάδων*

11.5 Διαχωριστική Ανάλυση Συστάδων

Οι διαχωριστικές μέθοδοι θεωρούν ένα πλήθος *N* σημείων και ένα πλήθος *k* συστάδων, και διαμερίζουν τα σημεία στις συστάδες. Τυπικά, το πλήθος των συστάδων *k* προκαθορίζεται από τον χρήστη. Ξεκινώντας από έναν αρχικό διαχωρισμό, με μια επαναληπτική διαδικασία, τα σημεία μετακινούνται από μια συστάδα σε μιαν άλλη. Ο σχηματισμός των συστάδων γίνεται με τρόπο τέτοιο, ώστε να βελτιστοποιείται ένα κριτήριο διαχωρισμού. Στόχος είναι να δημιουργηθούν συστάδες, οι οποίες να περιέχουν όμοια αντικείμενα, ενώ τα αντικείμενα διαφορετικών συστάδων να είναι ανόμοια.

Οι διαχωριστικές μέθοδοι παρουσιάζουν ευαισθησία στις αρχικές τους συνθήκες. Ένα σημαντικό πρόβλημα είναι το πλήθος των συστάδων *k*. Η εργασία του Dubes (1987) παρέχει καθοδήγηση για τον καθορισμό του πλήθους των συστάδων. Επίσης, για την εύρεση της καθολικά βέλτιστης λύσης θα έπρεπε να δοκιμαστούν όλοι οι δυνατοί διαχωρισμοί. Ωστόσο, λόγω υπολογιστικού κόστους, αυτό δεν είναι εφικτό. Στην πράξη εφαρμόζεται μια διαδικασία αρχικοποίησης του διαχωρισμού, και στη συνέχεια, μετακίνησης των σημείων.

Οι διαχωριστικές μέθοδοι δημιουργούν ένα σύνολο συστάδων, σε αντίθεση με τις ιεραρχικές μεθόδους, οι οποίες δημιουργούν μια ιεραρχική δομή διαδοχικών επιπέδων, όπου κάθε επίπεδο ορίζει ένα σύνολο συστάδων. Επίσης, είναι υπολογιστικά λιγότερο ακριβές από τις ιεραρχικές μεθόδους, και για τον λόγο αυτό

μπορούν να εφαρμοστούν σε μεγαλύτερα σύνολα δεδομένων. Η πιο γνωστή μέθοδος διαχωριστικής ανάλυσης συστάδων είναι ο αλγόριθμός k-Means.

11.5.1 Η μέθοδος k-Means

Η μέθοδος k-Means προτάθηκε από τον MacQueen (1967), και είναι η πιο γνωστή και διαδεδομένη διαιρετική μέθοδος ΑΣ. Στόχος της είναι να κατανείμει ένα σύνολο αντικειμένων σε έναν προκαθορισμένο αριθμό συστάδων, με τρόπο τέτοιο που να αυξάνει την ομοιότητα εντός των συστάδων. Ο αλγόριθμος περιλαμβάνει μια επαναληπτική διαδικασία, όπου σε κάθε επανάληψη υπολογίζεται το κέντρο της συστάδας (centroid). Τα αντικείμενα εντάσσονται στη συστάδα με το πλησιέστερο κέντρο.

Αναλυτικότερα, ο αλγόριθμος της μεθόδου k-Means έχει ως ακολούθως:

- Αρχικά επιλέγονται τυχαία *k* αντικείμενα. Ο αριθμός *k* είναι το πλήθος των συστάδων που θα προκύψουν και προκαθορίζεται από τον χρήστη. Τα επιλεγμένα σημεία θεωρούνται κέντρα συστάδων.
- Κάθε αντικείμενο κατατάσσεται στη συστάδα, της οποίας το κέντρο είναι πλησιέστερα του. Για τον υπολογισμό της απόστασης συνήθως χρησιμοποιείται η Ευκλείδεια απόσταση.
- Τα κέντρα της κάθε συστάδας επαναϋπολογίζονται. Για κάθε διάσταση το κέντρο έχει τιμή ίση με τη μέση τιμή όλων των αντικειμένων, τα οποία ανήκουν στη συστάδα.

$$m_i = \frac{1}{M_i} \sum_{j=1}^{M_i} x_j$$

(11.18)

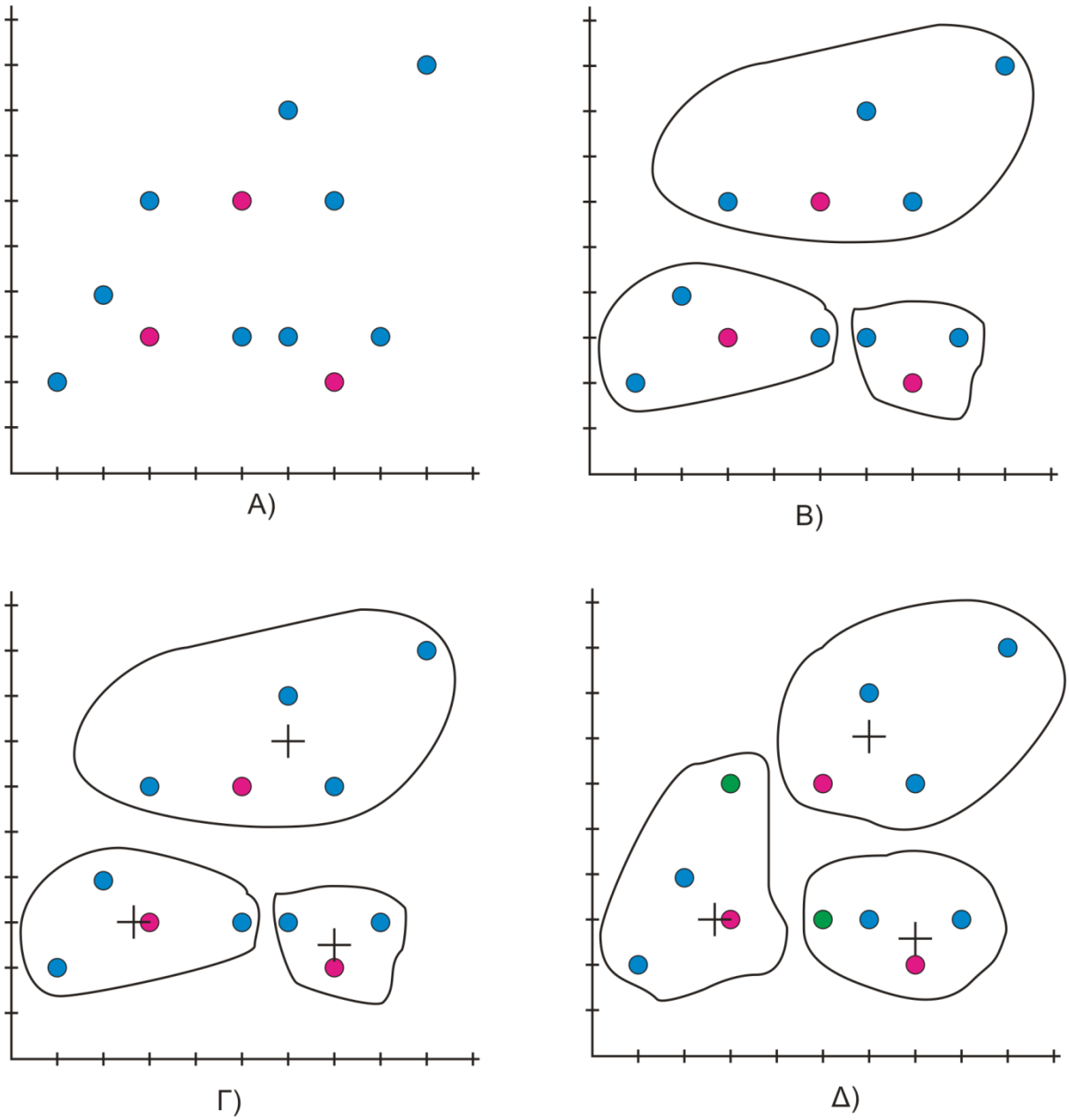
- όπου M_i είναι το πλήθος των αντικειμένων της συστάδας *i*, και m_i είναι το υπολογιζόμενο κέντρο.
- Τα προηγούμενα δύο βήματα επαναλαμβάνονται μέχρι να ικανοποιηθεί η συνθήκη εξόδου. Τυπικά, συνθήκη εξόδου είναι η ελαχιστοποίηση του τετραγωνικού σφάλματος, το οποίο ορίζεται από την Εξίσωση 11.19.

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} (x - m_i)^2$$

(11.19)

- όπου C_i είναι οι συστάδες, *x* είναι τα αντικείμενα και m_i είναι το κέντρο της συστάδας C_i.

Στο Σχήμα 11.5 παρουσιάζεται ο σχηματισμός των συστάδων με τη μέθοδο k-Means. Στο τμήμα Α) παρουσιάζονται τα σημεία. Τα κόκκινα σημεία συμβολίζουν τα αρχικώς επιλεγμένα κέντρα. Στο τμήμα Β) σχηματίζονται οι συστάδες. Κάθε σημείο εντάσσεται στη συστάδα, στης οποίας το κέντρο βρίσκεται πλησιέστερα. Στο τμήμα Γ) υπολογίζονται τα νέα κέντρα των υφιστάμενων συστάδων. Τα νέα κέντρα συμβολίζονται με το σχήμα του σταυρού. Στο τμήμα Δ) επαναϋπολογίζεται η απόσταση των σημείων από τα νέα κέντρα, και τα σημεία επανεντάσσονται στις συστάδες. Τα δύο πράσινα σημεία αλλάζουν συστάδα.



Σχήμα 11.5 *Δημιουργία συστάδων με k-Means*

Ο αλγόριθμος k-Means διαθέτει τα παρακάτω **πλεονεκτήματα**:

- Είναι απλός και κατανοητός.
- Τα αντικείμενα μοιράζονται σε συστάδες με αυτόματο τρόπο.
- Είναι αρκετά γρήγορος, τουλάχιστον σε σχέση με τις ιεραρχικές μεθόδους. Ο χρόνος εκτέλεσης του αλγορίθμου εξαρτάται γραμμικά από τα στοιχεία του προβλήματος, όπως το πλήθος των συστάδων *k*, το πλήθος των αντικειμένων n και το πλήθος των επαναλήψεων *l*. Η υπολογιστική πολυπλοκότητα του αλγορίθμου είναι *O(nkl)*. Για τον λόγο αυτό, είναι πιο κατάλληλος από άλλες μεθόδους για την ομαδοποίηση μεγάλων συνόλων αντικειμένων.

Τα βασικά **μειονεκτήματα** του k-Means είναι τα ακόλουθα:

- Ο αριθμός των συστάδων πρέπει να προκαθοριστεί από τον χρήστη.
- Το τελικό αποτέλεσμα εξαρτάται σε σημαντικό βαθμό από την επιλογή των αρχικών κέντρων. Επιλογή διαφορετικών κέντρων μπορεί να οδηγήσει σε σημαντικά διαφορετικές συστάδες.
- Είναι πολύ ευαίσθητος στην ύπαρξη αντικειμένων με ακραίες τιμές (outliers). Λίγα αντικείμενα με πολύ μεγάλες τιμές μπορούν να επηρεάσουν σημαντικά τον υπολογισμό των νέων

κέντρων και κατά συνέπεια τη διαμόρφωση των τελικών συστάδων.
- Έχει την τάση να δημιουργεί σφαιρικές και ίσου μεγέθους συστάδες. Για τον λόγο αυτό, δεν είναι κατάλληλος για συστάδες με περίπλοκα σχήματα ή με πολύ διαφορετικά μεγέθη.

Για την αντιμετώπιση των προβλημάτων του k-Means έχουν προταθεί διάφορες λύσεις. Ένα βασικό πρόβλημα είναι ο προκαθορισμός του αριθμού των συστάδων. Μια δυνατή λύση σε αυτό το πρόβλημα είναι να εφαρμοστεί αρχικά ιεραρχική ΑΣ. Η ιεραρχική ΑΣ συνίσταται σε μια διαδικασία διαδοχικών συνενώσεων ή διασπάσεων των συστάδων. Με τον τρόπο αυτόν, ο χρήστης μπορεί να εκτιμήσει το πλήθος των συστάδων, και στη συνέχεια να εκτελέσει τον k-Means. Ένα άλλο σημαντικό πρόβλημα είναι ότι ο αλγόριθμος μπορεί να συγκλίνει σε τοπικά βέλτιστα, και δεν υπάρχει εγγύηση για την εύρεση ενός καθολικού βέλτιστου. Το τελικό αποτέλεσμα επηρεάζεται σημαντικά από την επιλογή των αρχικών κέντρων. Ένας τρόπος αντιμετώπισης αυτού του προβλήματος είναι οι διαδοχικές, πολλαπλές εκτελέσεις του αλγορίθμου, με διαφορετικά αρχικά κέντρα κάθε φορά. Πρόσθετες τεχνικές επιδιώκουν τη σύγκλιση σε καθολικό βέλτιστο. Οι Likas, Vlassis and Verbeek (2003) εφαρμόζουν μια αιτιοκρατική διαδικασία καθολικής αναζήτησης. Στη διαδικασία αυτή εκτελούνται πολλαπλές τοπικές αναζητήσεις με τον k-Means για διαρκώς αυξανόμενο πλήθος συστάδων, μέχρι το τελικό επιθυμητό πλήθος συστάδων *M*.

11.5.2 Λοιποί αλγόριθμοι Διαιρετικής Ανάλυσης Συστάδων

11.5.2.1 k-Medoids

Όπως αναφέρθηκε και προηγουμένως, ο αλγόριθμος k-Means είναι ευαίσθητος στην ύπαρξη εξαιρέσεων. Ένας τρόπος αντιμετώπισης αυτού του προβλήματος είναι η χρήση ως κέντρου, όχι ενός υπολογιζόμενου μέσου σημείου, αλλά ενός υπαρκτού σημείου δεδομένων. Ο αλγόριθμος k-Medoids ακολουθεί αυτήν την προσέγγιση. Μια από τις πρώτες εκδοχές του k-Medoids ήταν η μέθοδος Partitioning Around Medoids – PAM (Kaufman & Rousseeuw, 1990).

Οι αλγόριθμοι k-Means και k-Medoids παρουσιάζουν αρκετές ομοιότητες:

- Αρχικά επιλέγονται αυθαίρετα τα κέντρα των συστάδων.
- Σε μια επαναληπτική διαδικασία τα κέντρα επαναπροσδιορίζονται.
- Σε κάθε επανάληψη μειώνεται το κριτήριο.
- Επιλογή διαφορετικών αρχικών κέντρων μπορεί να δώσει διαφορετικά αποτελέσματα.
- Δεν επιτυγχάνουν καθολικά βέλτιστα.

Αναλυτικότερα, στον αλγόριθμο k-Medoids επιλέγονται αρχικά k σημεία ως κέντρα (medoids). Τα υπόλοιπα σημεία κατατάσσονται στη συστάδα του πλησιέστερου κέντρου. Μια συνάρτηση κόστους μετρά το άθροισμα των αποστάσεων όλων των σημείων από το κέντρο της συστάδας τους. Σε μια επαναληπτική διαδικασία, σημεία τα οποία δεν είναι κέντρα δοκιμάζονται ως πιθανά κέντρα. Εάν για ένα σημείο το κόστος γίνεται μικρότερο, τότε το σημείο αυτό γίνεται το νέο κέντρο στη θέση του προηγούμενου.

Ο αλγόριθμος k-Medoids λειτουργεί πιο αποτελεσματικά από τον k-Means, όταν στα δεδομένα υπάρχουν αντικείμενα με ακραίες τιμές. Ωστόσο, το κόστος υπολογισμού των medoids είναι σημαντικά μεγαλύτερο από το κόστος υπολογισμού των μέσων τιμών. Για τον λόγο αυτό, ο k-Medoids υπολείπεται του k-Means ως προς τον χρόνο επεξεργασίας μεγάλων συνόλων δεδομένων.

11.5.2.2 CLARA

Ο αλγόριθμος k-Medoids δεν αποδίδει καλά με μεγάλα σύνολα δεδομένων, λόγω υπολογιστικού κόστους. Μια βελτίωση του αλγορίθμου, η οποία αντιμετωπίζει αυτό το πρόβλημα, είναι η μέθοδος CLARA (Clustering LARge Applcations) (Kaufman & Rousseeuw, 1990). Η μέθοδος CLARA δεν χρησιμοποιεί ολόκληρο το σύνολο δεδομένων. Αντιθέτως, εκτελεί τυχαία δειγματοληψία και επιλέγει ένα υποσύνολο του. Το υποσύνολο δεδομένων υπόκειται σε ανάλυση συστάδων, σύμφωνα με τη μέθοδο PAM. Λόγω της τυχαίας δειγματοληψίας, είναι αρκετά πιθανό, ότι τα medoids που θα υπολογιστούν, θα είναι όμοια με αυτά που θα προέκυπταν από την επεξεργασία ολόκληρου του συνόλου δεδομένων. Ο αλγόριθμος επιλέγει πολλά υποσύνολα δεδομένων και επιστρέφει το καλύτερο αποτέλεσμα.

11.6 Αυτοοργανούμενοι Χάρτες

Οι **Αυτοοργανούμενοι Χάρτες (ΑΟΧ)** (Self Organizing Maps (SOMs)) είναι ένας τύπος Νευρωνικού Δικτύου, ο οποίος προτάθηκε από τον Φιλανδό καθηγητή Kohonen (1982). Λόγω των ιδιαίτερων χαρακτηριστικών και δυνατοτήτων τους, οι Αυτοοργανούμενοι Χάρτες προσέλκυσαν το ενδιαφέρον του επιστημονικού κόσμου, και πλήθος βιβλίων και ερευνητικών εργασιών αναφέρθηκαν σε αυτούς. Ωστόσο, το βιβλίο το οποίο θεωρείται σημείο αναφοράς είναι το Kohonen (2001).

Ο τρόπος εκπαίδευσης των ΑΟΧ είναι μη επιβλεπόμενος. Στη μη-επιβλεπόμενη μάθηση, οι απαντήσεις δεν είναι γνωστές εκ των προτέρων. Οι αλγόριθμοι χρησιμοποιούν για την εκπαίδευση μόνο τα δεδομένα εισόδου, και όχι τις απαντήσεις του δικτύου. Στην περίπτωση των ΑΟΧ, εκτελείται μια επαναληπτική διαδικασία, όπου το μοντέλο τροφοδοτείται με παραδείγματα εκπαίδευσης. Οι νευρώνες του δικτύου προσαρμόζονται, έτσι ώστε να «μοιάζουν» με τα δεδομένα εκπαίδευσης. Ένα πολύ σημαντικό χαρακτηριστικό είναι ότι παρεμφερή παραδείγματα αντιστοιχίζονται σε περιοχές γειτονικών νευρώνων. Με τον τρόπο αυτό, διατηρούνται οι τοπολογικές σχέσεις των δεδομένων εισόδου. Οι ΑΟΧ παρουσιάζουν σημαντικές αναλογίες με τον τρόπο λειτουργίας του ανθρώπινου εγκεφάλου. Ο εγκέφαλος είναι χωρικά οργανωμένος, και συγκεκριμένα τμήματα του είναι υπεύθυνα για συγκεκριμένες εργασίες, όπως πχ τη μνήμη, την ομιλία κλπ. Με αντίστοιχο τρόπο, οι ΑΟΧ αντιστοιχούν παρεμφερείς έννοιες σε γειτονικές περιοχές. Για τον λόγο αυτό, οι ΑΟΧ θεωρούνται ένα από τα πιο ρεαλιστικά μοντέλα του ανθρώπινου εγκεφάλου.

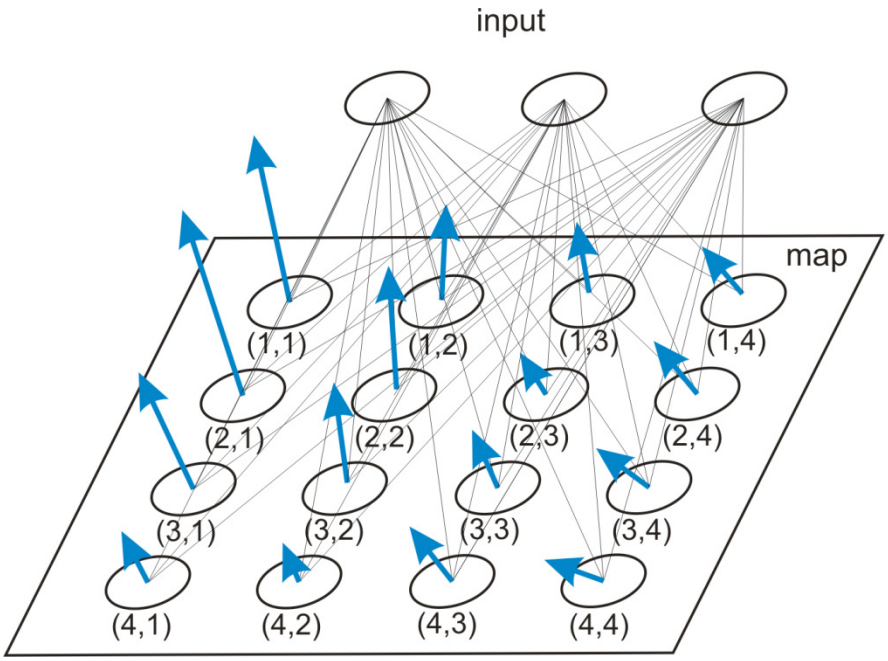
Οι ΑΟΧ παρέχουν έναν τρόπο απεικόνισης πολυδιάστατων δεδομένων, σε έναν χώρο πολύ λιγότερων, τυπικά μίας ή δύο, διαστάσεων. Με τον τρόπο αυτό, οπτικοποιούν σύνθετα δεδομένα, και τα παρουσιάζουν με τρόπο κατανοητό στους ανθρώπους. Ο Kohonen περιγράφει τους ΑΟΧ σαν ένα εργαλείο οπτικοποίησης και ανάλυσης πολυδιάστατων δεδομένων. Ωστόσο, χάρη στα ιδιαίτερα χαρακτηριστικά τους, οι ΑΟΧ μπορούν να χρησιμοποιηθούν για διάφορες εργασίες, όπως για την ανάλυση συστάδων, τη μείωση των διαστάσεων, ακόμα και για κατηγοριοποίηση. Μπορεί κανείς να φανταστεί τους ΑΟΧ σαν μια ελαστική επιφάνεια που διακυμαίνεται, έτσι ώστε να είναι όσο το δυνατό πλησιέστερα στα πρότυπα εκπαίδευσης,

11.6.1 Δομή ΑΟΧ

Ένας ΑΟΧ είναι ένα νευρωνικό δίκτυο ενός επιπέδου και αποτελείται από νευρώνες. Οι νευρώνες είναι διατεταγμένοι σε ένα πλέγμα *n* διαστάσεων. Κατά κανόνα, το πλέγμα είναι 2 διαστάσεων και ορθογώνιο ή εξαγωνικό.

Στο Σχήμα 11.6 απεικονίζεται ένας Αυτοοργανούμενος Χάρτης, ο οποίος διαθέτει ορθογώνιο πλέγμα 2 διαστάσεων. Το δίκτυο αποτελείται από 16 νευρώνες σε διάταξη *4X4*. Επίσης, υπάρχουν 3 νευρώνες εισόδου. Αυτό σημαίνει ότι τα πρότυπα εισόδου έχουν τρεις διαστάσεις, και ότι γίνεται προβολή ενός χώρου 3 διαστάσεων σε έναν χώρο 2 διαστάσεων. Κάθε νευρώνας εισόδου είναι συνδεδεμένος με κάθε νευρώνα του χάρτη. Με τον τρόπο αυτό, οι τιμές των προτύπων εισόδου διαβιβάζονται σε όλους τους νευρώνες του δικτύου. Ένα σημαντικό στοιχείο είναι ότι οι νευρώνες του χάρτη δεν είναι συνδεδεμένοι μεταξύ τους. Επίσης, κάθε νευρώνας του χάρτη διαθέτει δύο χωρικές συντεταγμένες *(i,j)*, οι οποίες καθορίζουν τη θέση του. Οι συντεταγμένες χρησιμεύουν για τον προσδιορισμό του εκάστοτε νευρώνα, καθώς και για τον υπολογισμό των αποστάσεων μεταξύ των νευρώνων.

Τα δεδομένα, με τα οποία θα εκπαιδευτεί και θα χρησιμοποιηθεί το δίκτυο, έχουν N διαστάσεις, είναι δηλαδή της μορφής $x=(x_1, x_2,...,x_N)$. Κάθε πρότυπο εισόδου μπορεί να θεωρηθεί ως ένα διάνυσμα N διαστάσεων. Επίσης, κάθε νευρώνας του δικτύου w περιέχει ένα σύνολο αριθμητικών τιμών $(w_1, w_2,...w_N)$, οι οποίες καλούνται και βάρη του νευρώνα. Το πλήθος των βαρών είναι ίσο με *N*, όσο δηλαδή είναι και το πλήθος των διαστάσεων των διανυσμάτων εισόδου. Επομένως, ο νευρώνας μπορεί να θεωρηθεί ως ένα διάνυσμα στον ίδιο χώρο με τα διανύσματα εισόδου. Τα διανύσματα των νευρώνων συμβολίζονται στο Σχήμα 11.6 με μπλε βέλη.



Σχήμα 11.6 *Αυτοοργανούμενος Χάρτης*

11.6.2 Εκπαίδευση ΑΟΧ

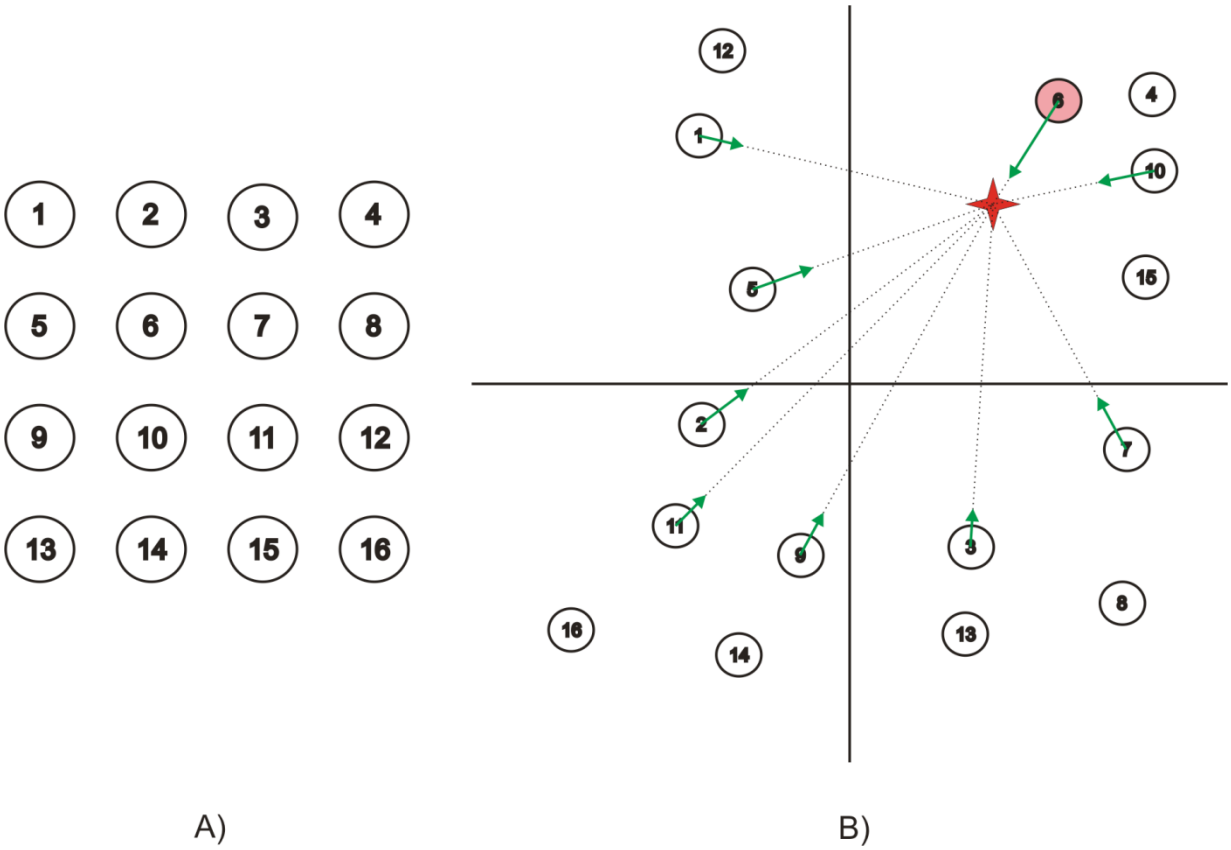
Η εκπαίδευση ενός ΑΟΧ συνίσταται στη ρύθμιση των βαρών των νευρώνων. Εκτελείται μια επαναλαμβανόμενη διαδικασία, όπου σε κάθε επανάληψη το δίκτυο τροφοδοτείται με ένα διάνυσμα εισόδου. Στη συνέχεια, υπολογίζεται η απόσταση του διανύσματος εισόδου από το διάνυσμα του κάθε νευρώνα. Ο νευρώνας, ο οποίος βρίσκεται πλησιέστερα στο διάνυσμα εισόδου, θεωρείται «νικητής», και είναι αυτός που αντιπροσωπεύει το διάνυσμα εισόδου. Ο νικητής νευρώνας ονομάζεται και Best Matching Unit (BMU). Τα βάρη του BMU τροποποιούνται κατάλληλα, έτσι ώστε το διάνυσμα του να μετακινηθεί προς το διάνυσμα εισόδου, και να βελτιωθεί η αντιπροσώπευση. Το μέγεθος της μεταβολής των βαρών εξαρτάται από μια παράμετρο, που ονομάζεται ρυθμός εκπαίδευσης.

Σκοπός των ΑΟΧ είναι η προβολή παρεμφερών προτύπων πολλών διαστάσεων σε μια ορισμένη περιοχή, ενός χώρου λιγότερων διαστάσεων. Για την επίτευξη αυτού του σκοπού, κατά τη διάρκεια της εκπαίδευσης, δεν ρυθμίζονται μόνο τα βάρη του νευρώνα νικητή, αλλά και τα βάρη των τοπογραφικά γειτονικών του νευρώνων. Τα διανύσματα των γειτονικών νευρώνων μετακινούνται και αυτά προς το διάνυσμα εισόδου. Η διαδικασία παρουσίασης ενός διανύσματος εισόδου, και ρύθμισης των βαρών του νικητή νευρώνα και των γειτονικών του νευρώνων επαναλαμβάνεται πολλές φορές. Το πλήθος των επαναλήψεων εξαρτάται από τον αριθμό των εποχών. Μια εποχή είναι η επεξεργασία όλων των διανυσμάτων εισόδου μια φορά. Με τον τρόπο αυτό, κάθε διάνυσμα εισόδου χρησιμοποιείται για εκπαίδευση τόσες φορές όσες είναι το πλήθος των εποχών.

Στο Σχήμα 11.7 παρουσιάζεται σχηματικά η ρύθμιση των βαρών των νευρώνων με τη χρήση ενός διανύσματος εισόδου. Στο αριστερό μέρος του σχήματος (τμήμα Α) απεικονίζεται η τοπολογική διάταξη των νευρώνων στο πλέγμα. Το δίκτυο διαθέτει 16 νευρώνες, σε επίπεδη διάταξη *4X4*. Για καλύτερη οπτική απεικόνιση, θεωρούμε διανύσματα εισόδου 2 διαστάσεων. Κατά συνέπεια, και οι νευρώνες του δικτύου θα είναι δύο διαστάσεων. Ανάλογα με τις τιμές των βαρών τους, οι νευρώνες απεικονίζονται ως σημεία σε έναν δισδιάστατο χώρο, στο δεξιό μέρος του σχήματος (τμήμα Β). Στο δίκτυο παρουσιάζεται ένα διάνυσμα εισόδου, το οποίο συμβολίζεται με το σχήμα του κόκκινου άστρου. Ο πλησιέστερος, από άποψη βαρών, νευρώνας, είναι ο νευρώνας 6. Ο νευρώνας αυτός είναι ο νικητής. Ο νικητής νευρώνας μετακινείται προς το διάνυσμα εισόδου, ώστε να βελτιωθεί η αντιπροσώπευση. Οι πλησιέστεροι τοπολογικά νευρώνες στον νευρώνα 6 είναι οι 1, 2, 3, 5, 7, 9, 10 και 11 (τμήμα Α). Οι νευρώνες αυτοί μετακινούνται επίσης προς το διάνυσμα εισόδου. Η μετακίνηση των νευρώνων γίνεται κατά μήκος των διακεκομμένων γραμμών. Η μεταβολή των βαρών και η αντίστοιχη μετακίνηση συμβολίζεται με τα πράσινα βέλη. Το μέγεθος του βέλους υποδηλώνει το ποσοστό της μεταβολής. Όσο τοπολογικά μακρύτερα βρίσκεται ο εκάστοτε νευρώνας από τον BMU, τόσο λιγότερο μετακινείται.

Πριν την έναρξη της εκπαίδευσης, τα βάρη των νευρώνων αρχικοποιούνται με τυχαίες τιμές. Κατά τη διάρκεια της εκπαίδευσης, το μέγεθος της γειτονιάς του BMU και ο ρυθμός εκπαίδευσης δεν παραμένουν στα-

θεροί, αλλά μειώνονται με τον αριθμό των εποχών. Ο καθορισμός του ρυθμού εκπαίδευσης και της ακτίνας της γειτονιάς αποτέλεσε αντικείμενο μελέτης πολλών ερευνητών. Σύμφωνα με τον Kohonen, η εκπαίδευση χωρίζεται σε δύο στάδια. Κατά το πρώτο στάδιο, το οποίο είναι γνωστό και ως unfolding phase, ο ρυθμός εκπαίδευσης μειώνεται από 0,9 σε 0,1. Η ακτίνα της γειτονιάς αρχικά ισούται με το ήμισυ της διαμέτρου του πλέγματος, και σταδιακά περιορίζεται, με τρόπο που να περιλαμβάνει στο τέλος τους άμεσους γείτονες του BMU. Κατά το δεύτερο στάδιο, γνωστό και ως fine tuning phase, ο ρυθμός εκπαίδευσης μειώνεται σταδιακά από την τιμή 0,1 στην τιμή 0,0, ενώ η ακτίνα της γειτονιάς διατηρεί σταθερή τιμή ίση με 1 και περιλαμβάνει μόνο τον νευρώνα BMU. Πρακτικά, κατά το πρώτο στάδιο καθορίζεται το γενικό σχήμα του δικτύου, και στο δεύτερο στάδιο προσαρμόζονται καλύτερα οι νευρώνες στα διανύσματα εισόδου.



Σχήμα 11.7 *Εκπαίδευση ΑΟΧ*

Μετά την ολοκλήρωση της εκπαίδευσης, κάθε διάνυσμα εισόδου θα αντιστοιχείται σε έναν νευρώνα, δεν θα ταυτίζεται όμως με αυτόν, δηλαδή μεταξύ τους θα υπάρχει μια διαφορά. Η διαφορά αυτή καλείται **σφάλμα κβάντωσης** (quantization error). Αθροίζοντας τη διαφορά κάθε διανύσματος εισόδου με τον νευρώνα στον οποίο αντιστοιχίζεται, λαμβάνουμε το συνολικό σφάλμα κβάντωσης. Το συνολικό σφάλμα κβάντωσης αποτελεί μέτρο της ποιότητας της αντιπροσώπευσης των διανυσμάτων εισόδου από το εκπαιδευμένο δίκτυο.

Ένα άλλο ζήτημα, το οποίο αφορά την ποιότητα του εκπαιδευμένου δικτύου, είναι η διατήρηση της τοπολογίας του χώρου εισόδου, δηλαδή η διατήρηση των σχέσεων γειτονίας που υφίστανται στον χώρο εισόδου. Για τη μέτρηση αυτού του ποιοτικού στοιχείου, ο Kohonen (2001) προτείνει ένα **τοπογραφικό σφάλμα**. Το σφάλμα αυτό μετρά το ποσοστό των διανυσμάτων εισόδου, για τα οποία ο πρώτος και ο δεύτερος νευρώνας αντιπροσώπευσης δεν είναι γειτονικοί. Πρόσθετα μέτρα για την εκτίμηση της διατήρησης της τοπολογίας έχουν προταθεί από τους Bauer and Pawelzik (1992), τους Bezdek and Pal (1995) και τους Uriarte and Martin (2005).

Το τελικό αποτέλεσμα της εκπαίδευσης ενός ΑΟΧ είναι η ρύθμιση των βαρών των νευρώνων. Ωστόσο, ο πίνακας των βαρών δεν θεωρείται ως ο πλέον κατάλληλος, για την οπτική αναπαράσταση και την αναγνώριση των συστάδων. Διάφορες τεχνικές έχουν προταθεί για την καλύτερη οπτική αναπαράσταση του χάρτη. Μια πολύ διαδεδομένη τεχνική είναι η U-Matrix (Ultsch & Siemon, 1990). Σύμφωνα με την τεχνική U-Matrix, υπολογίζονται οι αποστάσεις μεταξύ των διανυσμάτων γειτονικών νευρώνων. Οι αποστάσεις αυτές είναι μια εκτίμηση του βαθμού ανομοιότητας ομάδων των διανυσμάτων εισόδου. Αν οι αποστάσεις των διανυσμάτων μιας ομάδας γειτονικών νευρώνων είναι μικρές, τότε αυτό περιγράφει μια συστάδα δεδομένων εισόδου. Για

την καλύτερη οπτική αναπαράσταση της πληροφορίας, η διαφορά των αποστάσεων συμβολίζεται με διαβαθμίσεις του χρώματος γκρι. Ανοιχτόχρωμες περιοχές υποδηλώνουν την ύπαρξη συστάδας, ενώ σκουρόχρωμες περιοχές θεωρούνται ότι διαχωρίζουν συστάδες. Ορισμένες υλοποιήσεις του αλγορίθμου σε πακέτα λογισμικού χρησιμοποιούν πολλαπλά χρώματα αντί για διαβαθμίσεις του γκρι. Άλλη μέθοδος για την οπτικοποίηση των ΑΟΧ είναι η προβολή Sammon (Sammon, 1969).

11.7 Επιχειρηματικές Εφαρμογές της Ανάλυσης Συστάδων

Η Ανάλυση Συστάδων έχει βρει σημαντικές εφαρμογές στις σύγχρονες επιχειρήσεις. Το πιο γνωστό πεδίο εφαρμογής είναι το **μάρκετινγκ**. Οι διαφημιστικές εκστρατείες, οι οποίες απευθύνονται στο σύνολο του πληθυσμού, είναι ακριβές και έχουν μικρό ποσοστό ανταπόκρισης. Είναι προφανές ότι κάθε προσφορά δεν είναι χρήσιμη για κάθε πελάτη, και κάθε πελάτης δεν ανταποκρίνεται με τον ίδιο τρόπο στο ίδιο διαφημιστικό μήνυμα. Επιθυμία των διαφημιστών είναι να επιμερίσουν τον πληθυσμό σε ομάδες με ομοειδή χαρακτηριστικά. Οι επιχειρήσεις διατηρούν στα μηχανογραφικά τους συστήματα πληροφορίες για τους πελάτες τους. Τα στοιχεία αυτά μπορούν να αναλυθούν, ώστε να εξαχθούν συμπεράσματα χρήσιμα για διαφημιστικούς σκοπούς. Ένας νέος όρος, που περιγράφει αυτήν την πρακτική, είναι ο όρος «data base marketing». Εάν η επιχείρηση δεν διαθέτει στοιχεία, τότε μπορεί να διεξάγει μια έρευνα σε ένα αντιπροσωπευτικό δείγμα του πληθυσμού. Οι πελάτες μπορούν να ομαδοποιηθούν σύμφωνα με διάφορα κριτήρια, όπως γεωγραφικά (πόλη, περιοχή, χώρα κλπ.), ψυχογραφικά (τρόπος ζωής, προσωπικότητα, ηθικές αξίες), δημογραφικά (φύλο, ηλικία κλπ.) και καταναλωτικού προφίλ (προηγούμενες αγορές, τρόπος χρήσης του προϊόντος). Ο επιμερισμός του καταναλωτικού κοινού σε ομογενείς ομάδες είναι γνωστός με τον όρο «**τμηματοποίηση αγοράς**» (market segmentation). Οι διαφημιστές, έχοντας γνώση των τμημάτων που απαρτίζουν την αγορά, μπορούν να οργανώσουν στοχευμένες διαφημιστικές εκστρατείες, εξειδικευμένες για κάθε τμήμα πελατών. Αυτές οι διαφημιστικές εκστρατείες έχουν χαμηλότερο κόστος και καλύτερο ποσοστό ανταπόκρισης. Ο ακριβής και ουσιαστικός καθορισμός της ομάδας αυξάνει περαιτέρω τον βαθμό ανταπόκρισης. Η τμηματοποίηση της αγοράς και η βαθύτερη γνώση του καταναλωτικού προφίλ των υπαρκτών και υποψήφιων πελατών δεν αφορά μόνο τη διαφήμιση. Η εξυπηρέτηση των πελατών, μετά την πώληση, μπορεί να βελτιωθεί, με την ομαδοποίηση των απόψεων των πελατών, τα παράπονα τους, τις αναφορές για τεχνικά προβλήματα και βλάβες κλπ.

Το μάρκετινγκ είναι το πιο γνωστό παράδειγμα εφαρμογής της ΑΣ, δεν είναι όμως το μοναδικό. Κάθε τομέας της επιχειρηματικής δράσης μπορεί να ορίσει πρόσωπα, αντικείμενα ή ενέργειες σε σχέση με γνωρίσματα, και να ωφεληθεί, ανακαλύπτοντας συστάδες. Στη **διαχείριση ανθρωπίνων πόρων**, οι εργαζόμενοι ομαδοποιούνται σύμφωνα με τις αξιολογήσεις τους, την επαγγελματική τους εκπαίδευση, τη συμμετοχή τους σε ομάδες εργασίες, τις ειδικές δεξιότητες τους κλπ. Τα στοιχεία αυτά χρησιμοποιούνται για τον καθορισμό της μισθολογικής πολιτικής, τις προαγωγές των εργαζομένων, τη στελέχωση ομάδων εργασίας κλπ.

Εφαρμογή της ΑΣ γίνεται και στην **παραγωγή**. Κάθε προϊόν ανήκει σε κάποια κατηγορία προϊόντων. Η ένταξη προϊόντων σε κατηγορίες είναι μια παγιωμένη τακτική, ωστόσο με την πάροδο του χρόνου παρουσιάζονται προβλήματα, όπως απαρχαιωμένες κατηγορίες, ένταξη προϊόντων σε λάθος κατηγορίες κλπ. Με τη χρήση της ΑΣ, μπορούν να επανακαθοριστούν οι κατηγορίες με ουσιαστικότερο τρόπο και να επανενταχθούν προϊόντα στην κατάλληλη κατηγορία. Επίσης, η γνώση των ειδικών χαρακτηριστικών τμημάτων του καταναλωτικού κοινού αξιοποιείται και στην παραγωγή, με τον σχεδιασμό εξειδικευμένων προϊόντων για συγκεκριμένες ομάδες, και με τη δημιουργία αντίστοιχων γραμμών παραγωγής. Τέλος, ολόκληρες γραμμές παραγωγής και εργοστάσια μπορούν να ομαδοποιηθούν σύμφωνα με την ταχύτητα, την ποιότητα και το κόστος.

Η ΑΣ συστάδων μπορεί να χρησιμοποιηθεί για έλεγχο και **εντοπισμό απάτης**. Οι περιπτώσεις απάτης είναι λίγες και έχουν ιδιαίτερα χαρακτηριστικά. Με την εφαρμογή της ΑΣ μπορούν να εντοπιστούν μικρές και απομονωμένες συστάδες, οι οποίες είναι ύποπτες για απάτη. Παράδειγμα τέτοιας ανάλυσης είναι η εργασία των Thiprungsri and Vasarhelyi (2011). Οι ερευνητές μελετούν το πρόβλημα των ομαδικών ασφαλίσεων ζωής. Στις ομαδικές ασφαλίσεις, ο πελάτης είναι μια εταιρεία, η οποία ασφαλίζει το προσωπικό της, και όχι ένας μεμονωμένος ιδιώτης. Οι ασφαλιστικές εταιρείες, που παρέχουν τέτοια ασφαλιστικά συμβόλαια, δεν τηρούν πληροφορίες για τα ασφαλισμένα άτομα. Σε αυτόν τον τύπο ασφάλειας έχουν εντοπιστεί περιπτώσεις απάτης. Εφαρμόζοντας τη μέθοδο k-Means, οι ερευνητές εντόπισαν ολιγομελείς συστάδες, οι οποίες χρήζουν ειδικού περαιτέρω ελέγχου.

Η ΑΣ βρίσκει εφαρμογή στη **διαχείριση των Επιχειρηματικών Διαδικασιών**. Οι σύγχρονες επιχειρήσεις οργανώνουν τη λειτουργία τους με τον καθορισμό επιχειρηματικών διαδικασιών. Πρόσφατα, οι επιχειρηματικές διαδικασίες τυποποιούνται, και αυτοματοποιούνται σε εξειδικευμένα πληροφοριακά συστήματα. Οι Jung, Bae and Liu (2009) μετασχηματίζουν τις επιχειρηματικές διαδικασίες σε διανυσματικά μοντέλα, και τις ομα-

δοποιούν, εφαρμόζοντας Συσσωρευτική Ανάλυση Συστάδων. Τα αποτελέσματα μπορούν να χρησιμοποιηθούν για τον καθορισμό νέων ή τον ανασχεδιασμό των υπαρχουσών επιχειρηματικών διαδικασιών.

Η ΑΣ αξιοποιείται και στο **στρατηγικό μάνατζμεντ**. Τα διοικητικά στελέχη αναλύουν στοιχεία του επιχειρηματικού κλάδου, και ομαδοποιούν τις επιχειρήσεις ανάλογα με τα ιδιαίτερα χαρακτηριστικά τους, τα πλεονεκτήματα τους και τις αδυναμίες τους. Με τον τρόπο αυτό, κατανοούν καλύτερα τις συνθήκες του ανταγωνισμού, και μπορούν να σχεδιάσουν δράσεις, που θα τους εξασφαλίσουν το ανταγωνιστικό πλεονέκτημα. Ένα παράδειγμα ανάλυσης στοιχείων επιχειρηματικού κλάδου είναι η εργασία των Mosleh, Nosratabadi and Bahrami (2015), οι οποίοι αναλύουν τα επιχειρηματικά μοντέλα που εφαρμόζονται στα πρακτορεία τουρισμού του Ιράν. Εφαρμόζοντας ιεραρχική ΑΣ, εντοπίζουν ότι εφαρμόζονται τρία διαφορετικά επιχειρηματικά μοντέλα, με δημοφιλέστερο το μοντέλο που βασίζεται στα χρηματοοικονομικά, και ακολουθούμενο από το μοντέλο που βασίζεται στους πελάτες, και το μοντέλο που βασίζεται στις υπηρεσίες.

Πολύ ενδιαφέροντα πεδία εφαρμογής βρίσκει η ΑΣ σε ζητήματα που αφορούν την **ανάλυση χωρικής** πληροφορίας. Για παράδειγμα, μπορούν να ομαδοποιηθούν οι κατοικίες ανάλογα με τη γεωγραφική τους θέση, τον τύπο τους και την αξία τους, και να καθοριστούν εξειδικευμένες υπηρεσίες για συγκεκριμένες ομάδες. Επίσης, μπορούν να σχεδιαστούν θεματικοί χάρτες, οι οποίοι αναγνωρίζουν περιοχές με παρόμοια χρήση γης, όπως αγροτικές περιοχές, αστικές περιοχές, βιομηχανικές περιοχές κλπ. Τα στοιχεία αυτά είναι χρήσιμα σε κρατικούς φορείς για την άσκηση πολιτικής και τη δημιουργία υποδομών. Ένα πολύ σημαντικό σχετικό αντικείμενο είναι ο εντοπισμός **συστάδων επιχειρήσεων**. Μια συστάδα επιχειρήσεων είναι ένα σύνολο επιχειρήσεων, οι οποίες βρίσκονται στον ίδιο γεωγραφικό χώρο, και επιπλέον ενώνονται στη βάση κοινών στοιχείων, και λειτουργούν ανταγωνιστικά ή/και συμπληρωματικά. Η πιο γνωστή περίπτωση επιχειρηματικής συστάδας είναι η διαβόητη Silicon Valley. Η γεωγραφική συγκέντρωση παρόμοιων και αλληλοσχετιζόμενων επιχειρήσεων μπορεί να εξασφαλίσει μια σειρά από οφέλη όπως:

- Μείωση εξόδων μεταφοράς.
- Ανάπτυξη εξειδικευμένων υποδομών, χρήσιμων για τον κλάδο.
- Επίτευξη οικονομία κλίμακας.
- Αυξημένη έκθεση στον ανταγωνισμό, η οποία ενισχύει τη βελτίωση της ποιότητας.
- Κοινή αξιοποίηση του τοπικού εργατικού δυναμικού, το οποίο σταδιακά αποκτά εξειδικευμένες γνώσεις και δεξιότητες.
- Διάχυση βέλτιστων πρακτικών και νέων τεχνολογιών.
- Επίτευξη ανταγωνιστικού πλεονεκτήματος.

Ο εντοπισμός επιχειρηματικών συστάδων, ακόμα και εν τη γενέσει τους, είναι ιδιαίτερα χρήσιμος, καθώς οι αρμόδιοι φορείς μπορούν να λάβουν ειδικά μέτρα για την ενίσχυση του φαινομένου. Τέτοια μέτρα περιλαμβάνουν την ενίσχυση των υποδομών, την εκπαίδευση του πληθυσμού, την εξασφάλιση επιδοτήσεων, την προσαρμογή της φορολογικής πολιτικής, τη θέσπιση απαγορευτικών διατάξεων κλπ.

Όπως καταδείχτηκε ανωτέρω, η ΑΣ βρίσκει πεδία εφαρμογής σε πλήθος επιχειρηματικών δραστηριοτήτων και ζητημάτων. Σε κάθε περίπτωση, οι εμπλεκόμενοι αναλυτές και τα διοικητικά στελέχη οφείλουν να γνωρίζουν ότι στην ΑΣ συχνά απαιτείται πειραματισμός με διάφορα γνωρίσματα και μεθόδους, μέχρι να εξαχθούν χρήσιμα αποτελέσματα. Η ΑΣ έχει αρκετά έντονα περιγραφικό και διερευνητικό χαρακτήρα. Τα αποτελέσματα συνδυάζονται με τη γνώμη στελεχών, τα οποία διαθέτουν εξειδικευμένη γνώση στο συγκεκριμένο πεδίο. Τα στελέχη αυτά αξιολογούν τα αποτελέσματα των αλγορίθμων και αποφαίνονται εάν είναι λογικά, και εάν προσφέρουν χρήσιμη πληροφορία.

Βιβλιογραφία/Αναφορές

Aggarwal, C., & Reddy, C. (Eds.). (2014). *Data Clustering: Algorithms and Applications*. Hoboken: CRC Press.

Bauer, H. U., & Pawelzik, K. R. (1992). Quantifying the Neighborhood Preservation of Self-Orrganizing Maps. *IEEE Transactions on Neural Networks, 3*(4), 570-579. doi: 10.1109/72.143371

Bezdek, J. C., & Pal, N. R. (1995). An Index of Topological Preservation for Feature Extraction. *Pattern Recognition, 28*(3), 381-391. doi: 10.1016/0031-3203(94)00111-x

Dubes, R. C. (1987). How many Clusters are Best: An Experiment. *Pattern Recognition, 20*(6), 645-663. doi: 10.1016/0031-3203(87)90034-3

Estivill-Castro, V., & Yang, J.A. (2000). Fast and Robust General Purpose Clustering Algorithm. *Proceedings of the Pacific Rim International Conference on Artificial Intelligence,* 208-218. doi: 10.1007/3-540-44533-1_24

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining Concepts and Techniques*. Waltham, MA: Morgan Kaufmann Publishers.

Jung, J. Y., Bae, J., & Liu, L. (2009). Hierarchical Clustering of Business Process Models. *International Journal of Innovative Computing, Information and Control, 5*(12), 1-10.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. New York, NY: John Wiley and Sons.

King, B. (1967). Step-Wise Clustering Procedures. *Journal of the American Statistical Association, 62*(317), 86-101. doi: 10.2307/2282912

Kohonen, T. (1982). Self Organized Formation of Topologically Correct Feature Maps. *Biological Cybernetics, 43*(1), 59-69. doi: 10.1007/bf00337288

Kohonen, T. (2001). *Self-Organizing Maps*. Berlin: Springer-Verlag.

Kruskal, J. (1977). The Relationship between Multidimensional Scaling and Clustering. In J. Van Ryzin (Ed.), *Classification and Clustering* (pp 17-45). New York, NY: Academic Press Inc.

Likas, A., Vlassis, N., & Verbeek, J. J. (2003). The Global k-Means Clustering Algorithm. *Pattern Recognition, 36*(2), 451-461. doi: 10.1016/S0031-3203(02)00060-2

MacQueen, J. (1967). Some Methods for Classification and Analysis of Multivariate Observations. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability,* 281-297. Berkeley, CA: University of California Press.

Mao, J., & Jain, A. K. (1996). A self-organizing network for hyper ellipsoidal clustering (HEC). *IEEE Transactions on Neural Networks, 7*(1), 16–29. doi: 10.1109/72.478389

Mosleh, A., Nosratabadi, S., & Bahrami, P. (2015). Recognizing the Business Models types in Tourism Agencies: Utilizing the Cluster Analysis. *International Business Research, 8*(2), 173-180. doi: 10.5539/ibr.v8n2p173

Murtagh, F. A. (1984). A Survey of Recent Advances in Hierarchical Clustering Algorithms Which Use Cluster Centers. *The Computer Journal, 26*(4), 354-359. doi: 10.1093/comjnl/26.4.354

Ng, R. T., & Han, J. (1994). Efficient and Effective Clustering Methods for Spatial Data Mining. *Proceedings of the 20th International conference on Very Large Data Bases,* 144-155. San Francisco, CA: Morgan Kaufmann Publishers.

Sammon, J. (1969). A Nonlinear Mapping for data Structure Analysis. *IEEE Transactions on Computers, C-18*(5), 401-409. doi: 10.1109/t-c.1969.222678

Shekhar, S., & Chawla, S. W. (2003). *Spatial Databases: A Tour*. Upper Saddle River, NJ: Prentice Hall.

Sneath, P., & Sokal, R. (1973*). Numerical Taxonomy: The Principles and Practice of Numerical Classification*. San Francisco, CA: WH Freeman Co.

Thiprungsri, S., & Vasarhelyi, M. (2011). Cluster Analysis for Anomaly Detection in Accounting Data: An Audit Approach. *The International Journal of Digital Accounting Research, 11*, 69-84. doi: 10.4192/1577-8517-v11_4

Ultsch, A., & Siemon, H. P. (1990). Kohonen's Self Organizing Feature Maps for Exploratory Data Analysis. *Proceedings of the ICNN'90 International Neural Network Conference,* 305-308.

Uriarte, A., & Martin, D. (2005). Topology Preservation in SOM. *International Journal of Applied Mathematics and Computer Sciences, 1*(1), 19-22.

Ward, J. (1963). Hierarchical Grouping to Optimize and Objective Function. *Journal of the American*

Statistical Association, 58(301), 236-244. doi: 10.2307/2282967

Κριτήρια Αξιολόγησης

Άσκηση Υπολογισμών 11.1

Δίνονται τα ακόλουθα οκτώ αντικείμενα, κάθε ένα από τα οποία αποτελείται από δύο αριθμητικές τιμές: A(1,3), B(1,7), C(4,6), D(5,6), E(6,8), F(7,5), G(7,7), H(7,9). Εφαρμόστε τη μέθοδο k-Means για τον διαμοιρασμό των αντικειμένων σε δύο συστάδες. Χρησιμοποιήστε την Ευκλείδεια απόσταση. Υλοποιήστε τα ακόλουθα βήματα:
Θεωρήστε ως αρχικά κέντρα τα σημεία C και G.
Υπολογίστε τις αποστάσεις των υπόλοιπων έξι σημείων από τα σημεία C και G.
Κατανείμετε τα σημεία σε συστάδες ανάλογα με τις αποστάσεις τους από τα σημεία C και G.
Υπολογίστε τα νέα κέντρα των συστάδων C1 και C2.
Υπολογίστε τις αποστάσεις όλων των σημείων από τα νέα κέντρα C1 και C2.
Κατανείμετε τα σημεία σε συστάδες ανάλογα με τις αποστάσεις τους από τα νέα κέντρα C1 Και C2.

Λύση

Για τον υπολογισμό της απόστασης των έξι αντικειμένων από τα αντικείμενα C και G χρησιμοποιήστε την Εξίσωση 11.1. Οι αποστάσεις δίνονται στον Πίνακα 11.2.

Αντικείμενο	X	Y	Dist to C	Dist to G
A	1	3	4,2426	7,211103
B	1	7	3,1623	6
D	5	6	1	2,236068
E	6	8	2,8284	1,414214
F	7	5	3,1623	2
H	7	9	4,2426	2

Πίνακας 11.2 *Αποστάσεις αντικειμένων από τα αντικείμενα C και G*

Σύμφωνα με τα στοιχεία του πίνακα, τα αντικείμενα A, B, D βρίσκονται πλησιέστερα στο C, ενώ τα αντικείμενα E,F,H βρίσκονται πλησιέστερα στο G. Η πρώτη συστάδα θα αποτελείται από τα αντικείμενα A, B, C, D και η δεύτερη συστάδα από τα αντικείμενα E, F, G, H.

Υπολογίστε τα νέα κέντρα των δύο συστάδων C1 και C2 σύμφωνα με την Εξίσωση 11.18. Τα νέα κέντρα είναι τα σημεία C1(2,75, 5,5) και C2(6,75, 7,25).

Υπολογίστε τις αποστάσεις των οκτώ σημείων από τα νέα κέντρα C1 και C2 σύμφωνα με την Εξίσωση 11.1. Οι αποστάσεις παρατίθενται στον Πίνακα 11.3.

Αντικείμενο	X	Y	Dist to C1	Dist to C2
A	1	3	3,0516	7,150175
B	1	7	2,3049	5,755432
C	4	6	1,3463	3,020761
D	5	6	2,3049	2,150581
E	6	8	4,1003	1,06066
F	7	5	4,2793	2,263846
G	7	7	4,5069	0,353553
H	7	9	5,5057	1,767767

Πίνακας 11.3 *Αποστάσεις σημείων από τα νέα κέντρα C1 και C2*

Σύμφωνα με τα στοιχεία του πίνακα, τα αντικείμενα A, B, C βρίσκονται πλησιέστερα στο C1, ενώ τα αντικείμενα D,E,F,G,H βρίσκονται πλησιέστερα στο C2. Η πρώτη συστάδα θα αποτελείται από τα αντικείμενα A,

B, C και η δεύτερη συστάδα από τα αντικείμενα D, E, F, G, H. Το αντικείμενο D αλλάζει συστάδα.

Άσκηση Εφαρμογής 11.2

Χρησιμοποιήστε το αρχείο «Wholesale Customers Data.csv». Θα το βρείτε στην ιστοσελίδα UCI Machine Learning Repository με το όνομα «Wholesale Customers». Το αρχείο περιέχει στοιχεία πωλήσεων σε πελάτες ενός διανομέα χονδρικής. Αποτελείται από οκτώ στήλες και 440 γραμμές. Η πρώτη στήλη αφορά το κανάλι διανομής, η δεύτερη τις περιοχές, και οι υπόλοιπες έξι στήλες αφορούν τις κατηγορίες προϊόντων. Οι τιμές αναφέρονται σε συνολικές ετήσιες πωλήσεις ανά κατηγορία προϊόντος. Σύμφωνα με την ιστοσελίδα UCI Machine Learning Repository, το συγκεκριμένο σύνολο δεδομένων είναι κατάλληλο για Ανάλυση Συστάδων και Κατηγοριοποίηση.
Πραγματοποιήστε με το WEKA Ιεραρχική Ανάλυση Συστάδων. Χρησιμοποιήστε διαδοχικά τύπο σύνδεσης, Single, Complete, Mean, Centroid και Ward. Για κάθε περίπτωση ορίστε να δημιουργηθούν τρεις συστάδες. Συγκρίνετε τα αποτελέσματα και τα δενδρογράμματα της κάθε περίπτωσης.

Λύση

Βήμα 1. Προμηθευτείτε το αρχείο από την ιστοσελίδα UCI Machine Learning Repository.

Το αρχείο είναι σε μορφότυπο Comma Separated Values (CSV). Πρέπει να μετατραπεί σε μορφότυπο ARFF. Ανοίξτε το αρχικό αρχείο στο Excel και αποθηκεύστε σε μορφότυπο XLS. Χρησιμοποιήστε την εφαρμογή μετατροπής αρχείων Excel σε αρχεία ARFF. Τη συγκεκριμένη εφαρμογή θα τη βρείτε στην ιστοσελίδα Excel to Arff Converter download. Μετατρέψτε το αρχείο, μεταφέρετε τα αποτελέσματα με αντιγραφή και επικόλληση σε κάποιον κειμενογράφο και αποθηκεύστε τα σε αρχείο κειμένου. Αλλάξτε την κατάληξη του αρχείου από TXT σε ARFF.

Βήμα 2. Εκκινήστε το WEKA και ανοίξτε το αρχείο «Wholesale Customers Data.csv», πιέζοντας το κουμπί «Open file».

Στο tab «Preprocess» μελετήστε τα γνωρίσματα και τις κατανομές τιμών. Μπορείτε να δείτε τα δεδομένα αναλυτικά κάνοντας κλικ στο κουμπί «Edit».

Βήμα 3. Μεταβείτε στο tab «Cluster».

Στο πεδίο «Clusterer» κάντε κλικ στο κουμπί «Choose» και επιλέξτε weka/clusterer/HierarchicalClusterer.

Κάντε κλικ στα περιεχόμενα του πεδίου «Clusterer» και στο όνομα HierarchicalClusterer. Θα ανοίξει το παράθυρο ρύθμισης παραμέτρων. Βεβαιωθείτε ότι ως συνάρτηση απόστασης έχει οριστεί η Ευκλείδεια απόσταση και ότι ο τρόπος σύνδεσης είναι τύπου Single. Ορίστε να εξαχθούν τρεις συστάδες (πεδίο «numClusters»). Κάντε κλικ στο κουμπί «OK» για να επιστέψετε στο παράθυρο του WEKA και στη συνέχεια κάντε κλικ στο κουμπί «Start» για να εκτελέσετε τον αλγόριθμο.

Στο πεδίο «Clusterer output» εμφανίζονται τα αποτελέσματα. Η κατανομή των παρατηρήσεων σε συστάδες δεν είναι τόσο επιτυχημένη, καθώς η πρώτη συστάδα περιέχει 420 παρατηρήσεις, η δεύτερη μια παρατήρηση και η τρίτη δεκαεννέα παρατηρήσεις. Στο πεδίο «Results list», κάντε δεξί κλικ στην εγγραφή και επιλέξτε «Visualize tree» για να δείτε το δενδρόγραμμα.

Βήμα 4. Κάντε κλικ στα περιεχόμενα του πεδίου «Clusterer» και στο όνομα HierarchicalClusterer. Θα ανοίξει το παράθυρο ρύθμισης παραμέτρων. Αλλάξτε τον τρόπο σύνδεσης σε Complete. Ορίστε να εξαχθούν τρεις συστάδες. Εκτελέστε τον αλγόριθμο.

Στο πεδίο «Clusterer output» εμφανίζονται τα αποτελέσματα. Οι τρεις συστάδες περιέχουν 138, 298 και 4 παρατηρήσεις. Στο πεδίο «Results list» κάντε δεξί κλικ στην εγγραφή και επιλέξτε «Visualize tree» για να δείτε το δενδρόγραμμα.

Βήμα 5. Κάντε κλικ στα περιεχόμενα του πεδίου «Clusterer» και στο όνομα HierarchicalClusterer. Θα ανοίξει το παράθυρο ρύθμισης παραμέτρων. Αλλάξτε τον τρόπο σύνδεση σε Mean. Ορίστε να εξαχθούν τρεις συστάδες. Εκτελέστε τον αλγόριθμο.

Στο πεδίο «Clusterer output» εμφανίζονται τα αποτελέσματα. Οι τρεις συστάδες περιέχουν 120, 259 και 61 παρατηρήσεις. Η κατανομή των παρατηρήσεων σε συστάδες έγινε λιγότερο ανισομερής. Στο πεδίο «Results list» κάντε δεξί κλικ στην εγγραφή και επιλέξτε «Visualize tree» για να δείτε το δενδρόγραμμα.

Βήμα 6. Κάντε κλικ στα περιεχόμενα του πεδίου «Clusterer» και στο όνομα HierarchicalClusterer. Θα ανοίξει το παράθυρο ρύθμισης παραμέτρων. Αλλάξτε τον τρόπο σύνδεσης σε Centroid. Ορίστε να εξαχθούν τρεις συστάδες. Εκτελέστε τον αλγόριθμο.

Στο πεδίο «Clusterer output» εμφανίζονται τα αποτελέσματα. Οι τρεις συστάδες περιέχουν 142, 298 και 1

παρατηρήσεις. Δείτε με γραφικό τρόπο το δενδρόγραμμα.

Βήμα 7. Κάντε κλικ στα περιεχόμενα του πεδίου «Clusterer» και στο όνομα HierarchicalClusterer. Θα ανοίξει το παράθυρο ρύθμισης παραμέτρων. Αλλάξτε τον τρόπο σύνδεσης σε Ward. Ορίστε να εξαχθούν τρεις συστάδες. Εκτελέστε τον αλγόριθμο.

Στο πεδίο «Clusterer output» εμφανίζονται τα αποτελέσματα. Οι τρεις συστάδες περιέχουν 142, 239 και 59 παρατηρήσεις. Δείτε με γραφικό τρόπο το δενδρόγραμμα.

Άσκηση Εφαρμογής 11.3

Χρησιμοποιήστε το αρχείο «Wholesale Customers Data.csv». Θα το βρείτε στην ιστοσελίδα UCI Machine Learning Repository με το όνομα «Wholesale Customers». Περισσότερες πληροφορίες για τα δεδομένα υπάρχουν στην εκφώνηση της Άσκησης 11.2
Πραγματοποιήστε με το WEKA Ανάλυση Συστάδων με τη μέθοδο k-Means. Δημιουργήστε τρεις συστάδες χρησιμοποιώντας την Ευκλείδεια απόσταση και την απόσταση Manhattan. Δείτε με γραφικό τρόπο την κατανομή των παρατηρήσεων σε συστάδες ανάλογα με τις τιμές του κάθε γνωρίσματος. Δημιουργήστε στα δεδομένα μια καινούργια στήλη, στην οποία θα αναφέρεται η συστάδα στην οποία ανήκει, σύμφωνα με τη μέθοδο k-Means με απόσταση Manhattan και πλήθος συστάδων ίσο με τρία.

Λύση

Βήμα 1. Προμηθευτείτε το αρχείο από την ιστοσελίδα UCI Machine Learning Repository.

Το αρχείο είναι σε μορφή Comma Separated Values (CSV). Πρέπει να μετατραπεί σε μορφή ARFF. Οδηγίες για τη μετατροπή θα βρείτε στη λύση της Άσκησης 11.2.

Εκκινήστε το WEKA και ανοίξτε το αρχείο «Wholesale Customers Data.arff» πιέζοντας το κουμπί «Open file».

Βήμα 2. Μεταβείτε στο tab «Cluster».

Στο πεδίο «Clusterer» κάντε κλικ στο κουμπί «Choose» και επιλέξτε weka/clusterer/SimpleKMeans.

Κάντε κλικ στα περιεχόμενα του πεδίου «Clusterer» και στο όνομα SimpleKMeans. Θα ανοίξει το παράθυρο ρύθμισης παραμέτρων. Βεβαιωθείτε ότι ως συνάρτηση απόστασης έχει οριστεί η Ευκλείδεια απόσταση. Ορίστε να εξαχθούν τρεις συστάδες (πεδίο «numClusters»). Κάντε κλικ στο κουμπί «OK» για να επιστέψετε στο παράθυρο του WEKA και στη συνέχεια κάντε κλικ στο κουμπί «Start» για να εκτελέσετε τον αλγόριθμο.

Στο πεδίο «Clusterer output» εμφανίζονται τα αποτελέσματα. Οι τρεις συστάδες περιέχουν 298, 132 και 10 παρατηρήσεις.

Βήμα 3. Στο πεδίο «Results list» κάντε δεξί κλικ στην εγγραφή και επιλέξτε «Visualize cluster assignments». Ανοίγει το παράθυρο οπτικοποίησης του WEKA, όπου οι παρατηρήσεις απεικονίζονται σε δισδιάστατο χώρο. Στον άξονα *x* είναι διατεταγμένες οι παρατηρήσεις, ενώ ο άξονας *y* αντιστοιχεί σε ένα από τα γνωρίσματα. Επιλέξτε διαδοχικά τα γνωρίσματα και παρατηρήστε την κατανομή των παρατηρήσεων και την ένταξη τους σε κλάσεις. Ο διαχωρισμός των κλάσεων είναι ιδιαίτερα εμφανής σύμφωνα με τις τιμές του πεδίου «Detergents_Paper».

Βήμα 4. Κάντε κλικ στα περιεχόμενα του πεδίου «Clusterer» και στο όνομα SimpleKMeans. Θα ανοίξει το παράθυρο ρύθμισης παραμέτρων. Επιλέξτε ως συνάρτηση απόστασης την απόσταση Manhattan και ορίστε να εξαχθούν τρεις συστάδες. Εκτελέστε τον αλγόριθμο.

Στο πεδίο «Clusterer output» εμφανίζονται τα αποτελέσματα. Οι τρεις συστάδες περιέχουν 298, 89 και 53 παρατηρήσεις.

Βήμα 5. Στο πεδίο «Results list» κάντε δεξί κλικ στην εγγραφή και επιλέξτε «Visualize cluster assignments». Ανοίγει το παράθυρο οπτικοποίησης του WEKA. Επιλέξτε διαδοχικά τα γνωρίσματα για τον άξονα *y* και παρατηρήστε την κατανομή των παρατηρήσεων και την ένταξη τους σε κλάσεις. Ξανά ο διαχωρισμός των κλάσεων είναι ιδιαίτερα εμφανής σύμφωνα με τις τιμές του πεδίου «Detergents_Paper» και σε μικρότερο βαθμό για τις τιμές των πεδίων «Grocery» και «Milk».

Βήμα 6. Μεταβείτε στο tab «Preprocess». Κάντε κλικ στο κουμπί «Choose» του πεδίου «Filter» και επιλέξτε weka/filters/unsupervised/attribute/AddCluster. Κάντε κλικ στα περιεχόμενα του πεδίου «Filter» και στο όνομα «AddCluster». Ανοίγει το παράθυρο ρύθμισης παραμέτρων. Κάντε κλικ στο κουμπί «Choose» του πεδίου «Clusterer» και επιλέξτε weka/clusterer/SimpleKMeans. Κάντε κλικ στα περιεχόμενα του πεδίου «clusterer» και στο όνομα «SimpleKMeans». Ανοίγει το παράθυρο ρύθμισης παραμέτρων του SimpleKMeans. Στο πεδίο «distanceFunction» επιλέξτε την απόσταση Manhattan και εισάγετε στο πεδίο «numClusters» την

τιμή 3. Κλείστε τα παράθυρα ρύθμισης παραμέτρων πατώντας το πλήκτρο «OK» και εκτελέστε τον αλγόριθμο κάνοντας κλικ στο κουμπί «Apply». Θα προστεθεί στα δεδομένα μια νέα στήλη, όπου θα αναγράφεται η συστάδα στην οποία ανήκει η εκάστοτε παρατήρηση.

Άσκηση Εφαρμογής 11.4

Χρησιμοποιήστε το αρχείο «Wholesale Customers Data.csv». Θα το βρείτε στην ιστοσελίδα UCI Machine Learning Repository με το όνομα «Wholesale Customers». Περισσότερες πληροφορίες για τα δεδομένα υπάρχουν στην εκφώνηση της Άσκησης 11.2
Πραγματοποιήστε με το WEKA Ανάλυση Συστάδων με τη μέθοδο k-Means. Δημιουργήστε τρεις συστάδες χρησιμοποιώντας την απόσταση Manhattan. Δημιουργήστε στα δεδομένα μια καινούργια στήλη, στην οποία θα αναφέρεται η συστάδα στην οποία ανήκει, σύμφωνα με τη μέθοδο k-Means με απόσταση Manhattan και πλήθος συστάδων ίσο με τρία.
Χρησιμοποιώντας ως τιμή κλάσης τη συστάδα στην οποία ανήκουν οι παρατηρήσεις, εκτελέστε κατηγοριοποίηση με Δένδρα Αποφάσεων C4.5.

Λύση

Βήμα 1. Προμηθευτείτε το αρχείο από την ιστοσελίδα UCI Machine Learning Repository.

Το αρχείο είναι σε μορφή Comma Separated Values (CSV). Πρέπει να μετατραπεί σε μορφή ARFF. Οδηγίες για τη μετατροπή θα βρείτε στη λύση της Άσκησης 11.2.

Εκκινήστε το WEKA και ανοίξτε το αρχείο «Wholesale Customers Data.arff» πιέζοντας το κουμπί «Open file».

Βήμα 2. Μεταβείτε στο tab «Cluster» και εκτελέστε την Ανάλυση Συστάδων σύμφωνα με τα στοιχεία της εκφώνησης και τις οδηγίες της Άσκησης 11.3. Προσθέστε τη στήλη με το όνομα της συστάδας σύμφωνα με τα στοιχεία της εκφώνησης και τις οδηγίες της Άσκησης 11.3.

Βήμα 3. Μεταβείτε στο tab «Classify».

Επιλέξτε κατηγοριοποιητή J48 και επικύρωση τύπου cross-validation. Εκτελέστε τον αλγόριθμο.

Στο πεδίο «Classifier Output» μελετήστε τα αποτελέσματα. Το Δένδρο Αποφάσεων επιτυγχάνει υψηλότατη ακρίβεια της τάξης του 97%. Επίσης, το δένδρο περιγράφει ένα κατανοητό τρόπο κατανομής των παρατηρήσεων στις συστάδες.