Κεφάλαιο 1

Βασικες Αρχες της Αναγνωρισης Προτυπων

1.1. Ορισμός

Η *αναγνώριση προτύπων* (*pattern recognition*) είναι ο επιστημονικός κλάδος που ασχολείται με την περιγραφή και κατάταξη αντικειμένων σε ένα αριθμό κατηγοριών. Τα υπό κατάταξη αντικείμενα καλούνται *πρότυπα* (*patterns*). Η αναγνώριση προτύπων θεωρείται ένα βασικό χαρακτηριστικό των ανθρώπων καθώς και άλλων ζώντων οργανισμών. Ένα πρότυπο είναι η περιγραφή ενός αντικειμένου. Καθημερινά αναγνωρίζουμε πρότυπα γύρω μας όπως: αντικείμενα, το πρόσωπο ενός φίλου μέσα στο πλήθος και μπορούμε να καταλάβουμε αν είναι θυμωμένος ή χαρούμενος

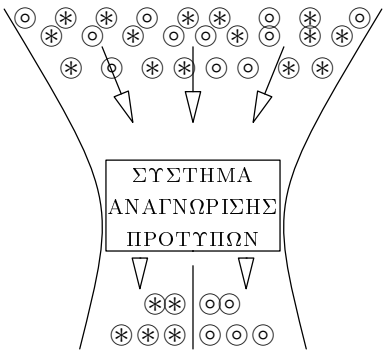
από την έκφρασή του καθώς και να αναγνωρίσουμε τη φωνή του χωρίς να τον βλέπουμε.

Ανάλογα με τη φύση των προτύπων προς αναγνώριση μπορούμε να διαχωρίσουμε τον ανθρώπινο τρόπο αναγνώρισης προτύπων σε δύο κατηγορίες: την αναγνώριση αφενός *σαφών προτύπων* (*concrete items*) όπως είναι χαρακτήρες, εικόνες, αντικείμενα και ήχους και αφετέρου την αναγνώριση *αφηρημένων* (*abstract*) προτύπων όπως η λύση ενός μαθηματικού προβλήματος ή ενός φιλοσοφικού επιχειρήματος. Η αναγνώριση αφηρημένων προτύπων μπορεί να θεωρηθεί σαν εννοιολογική αναγνώριση σε αντιδιαστολή με την οπτική και ακουστική αναγνώριση που είναι αναγνώριση σαφών προτύπων. Στο βιβλίο αυτό θα μελετηθεί μόνο η κατηγορία της αναγνώρισης σαφών προτύπων και συγκεκριμένα ο σχεδιασμός συστημάτων αναγνώρισης προτύπων, χρησιμοποιώντας τον υπολογιστή και εφαρμόζοντας τεχνικές πληροφορικής, στατιστικής και μηχανικής.

Η διαδικασία αναγνώρισης σαφών προτύπων εμπλέκει τον καθορισμό και ταξινόμηση αφενός *προτύπων στο χώρο* (*spatial patterns*) όπως είναι τυπωμένοι χαρακτήρες, άνθρωποι, δακτυλικά αποτυπώματα, εικόνες, χάρτες καιρού, φυσικά αντικείμενα και αφετέρου *προτύπων στο χρόνο* (*temporal patterns*) δηλαδή χρονοσειρές όπως είναι ακουστικές κυματομορφές, ηλεκτροεγκεφαλογραφήματα και σήματα ραντάρ.

Με απλά λόγια η αναγνώριση προτύπων μπορεί να οριστεί ως η κατηγοριοποίηση δεδομένων εισόδου στο σύστημα σε αναγνωρίσιμες κατηγορίες μέσω της εξαγωγής σημαντικών γνωρισμάτων ή χαρακτηριστικών από τα δεδομένα εισόδου, παραλείποντας τις άσχετες πληροφορίες. Μια σχηματική παράσταση της λειτουργίας ενός συστήματος αναγνώρισης προτύπων παρουσιάζεται στο Σχήμα 1.1.

Η πρόγνωση του καιρού από μετεωρολογικούς χάρτες μπορεί να θεωρηθεί σαν ένα πρόβλημα αναγνώρισης προτύπων. Το σύστημα κάνει μια πρόγνωση βασισμένο σε πληροφορίες που εξάγει από τους χάρτες. Η Ιατρική διάγνωση μπορεί να θεωρηθεί σαν ένα πρόβλημα αναγνώρισης προτύπων. Τα συμπτώματα είναι οι είσοδοι και το σύστημα εξάγει μια



Σχήμα 1.1: *Το πρόβλημα της αναγνώρισης προτύπων.*

Πίνακας 1.1: *Εφαρμογές αναγνώρισης προτύπων.*

Οπτική αναγνώριση χαρακτήρων	Εικόνα του χαρακτήρα	Χαρακτήρας
Αναγνώριση ομιλίας	Ηχητικό σήμα	Η λέξη
Αναγνώριση ομιλητή	Φωνή	Όνομα ομιλητή
Πρόγνωση καιρού	Μετεωρολογικοί χάρτες	Πρόβλεψη καιρού
Ιατρική διάγνωση	Συμπτώματα	Ασθένεια
Χρηματοοικονομικές εφαρμογές	Οικονομικά στοιχεία	Χρηματοοικονομική πρόβλεψη

ιατρική γνωμάτευση αναλύοντας τις εισόδους του. Ο Πίνακας 1.1 περιγράφει πολλές εφαρμογές αναγνώρισης προτύπων παρουσιάζοντας ταυτόχρονα τα αντίστοιχα δεδομένα εισόδου και την απόκριση εξόδου για κάθε εφαρμογή.

1.2. Εφαρμογές

Ο τρόπος με τον οποίο οι εφαρμογές του Πίνακα 1.1 χρησιμοποιούνται στη λύση καθημερινών προβλημάτων, αλλά και οι προοπτικές στη λύση πιο σύνθετων και πολύπλοκων θεμάτων καθώς και άλλες εφαρμογές παρουσιάζονται πιο αναλυτικά παρακάτω:

- Ήδη πολλά λειτουργικά συστήματα για προσωπικούς υπολογιστές όπως το System 7 ενός Apple και το OS/2 της IBM διαθέτουν ενσωματωμένο τμήμα αναγνώρισης ομιλίας. Στο μέλλον πιστεύεται ότι το μικρόφωνο θα είναι τόσο σημαντικό εξάρτημα για την επικοινωνία ανθρώπου με υπολογιστή (Human Computer Interaction — HCI) όσο είναι τώρα το ποντίκι και το πληκτρολόγιο.
- **(Optical Character Recognition — OCR):** Δίνει στον υπολογιστή τη δυνατότητα να καταλάβει τυπωμένο κείμενο. Συσκευές όπως ο σαρωτής και η βιντεοκάμερα χρησιμοποιούνται για να εισάγουν την εικόνα στον υπολογιστή και στη συνέχεια το σύστημα αναγνώρισης οπτικών χαρακτήρων μετατρέπει την εικόνα σε κατάλληλη μορφή, συνήθως σε κώδικα ASCII. Ένα παράδειγμα πολύπλοκης εφαρμογής ενός τέτοιου συστήματος είναι η αναγνώριση των αριθμών κυκλοφορίας των αυτοκινήτων σε έναν αυτοκινητόδρομο.
- Είναι ένα πολύ πιο σύνθετο και πολύπλοκο πρόβλημα απ΄ότι η αναγνώριση τυπωμένων χαρακτήρων. Μια απλουστευμένη μορφή της είναι η χρήση ειδικής γραφίδας, που αντικαθιστά το πληκτρολόγιο σε εφαρμογές αλληλεπίδρασης ανθρώπου με υπολογιστή. Ήδη τα σύγχρονα σημειωματάρια όπως το Psion και το Neuton της Apple συνοδεύονται από γραφίδα και ευαίσθητη οθόνη για την εισαγωγή χαρακτήρων.

- Ο υπολογιστής πραγματοποιεί την αναγνώριση ατόμων κυρίως σε εφαρμογές ασφάλειας από διάφορα σωματομετρικά χαρακτηριστικά όπως: δακτυλικά αποτυπώματα, σχήματα ίριδας, χαρακτηριστικά φωνής και τρόπο γραφής.
- Χρησιμοποιείται ως βοηθητικό εργαλείο σε πολλούς ιατρικούς κλάδους. Για παράδειγμα, χρησιμοποιείται για την αυτοματοποιημένη ανάλυση ιατρικής εικόνας (ακτινογραφίας, αξονικής τομογραφίας, υπερηχογραφήματος κλπ.), την ταξινόμηση εγκεφαλογραφημάτων, καρδιογραφημάτων και την ανίχνευση γενετικών ανωμαλιών σε χρωμοσώματα.
- **Geographical Integration Systems — GIS** Χρησιμοποιείται ως κύριο εργαλείο στην αυτοματοποιημένη ανάλυση δορυφορικής φωτογραφίας για την ανίχνευση ασθενειών σε καλλιέργειες, ίχνη αρχαίων οικισμών, των χρήσεων γης, την κατάρτιση του κτηματολογίου, την κατάσταση της ατμόσφαιρας καθώς και για ορυκτολογικές έρευνες.
- Σε συστήματα αυτομάτου ελέγχου παραγωγής, τα οποία ελέγχουν την ποιότητα των παραγόμενων αγαθών σε μια γραμμή παραγωγής. Μια μεταφορική ταινία περνάει μπροστά από μια κάμερα τα προϊόντα και ένα σύστημα αναγνώρισης προτύπων ελέγχει την ποιότητά τους. Για παράδειγμα, σε βιομηχανία κατασκευής πλακετών το αυτόματο υπολογιστικό σύστημα αναγνώρισης προτύπων ελέγχει την καταλληλότητα και ποιότητα της πλακέτας.
- Χρησιμοποιείται για την εύρεση ύποπτων συναλλαγών με πιστωτικές κάρτες, ταξινόμηση αιτήσεων προς δανειοδότηση, διαχείριση αποθεμάτων και την πρόβλεψη των τιμών των μετοχών. Επίσης για τον εντοπισμό πελατών που είναι πιο πιθανό να πραγματοποιήσουν συγκεκριμένες αγορές στο άμεσο μέλλον, ώστε σε αυτούς να σταλούν

διαφημιστικά φυλλάδια, για να μειωθεί έτσι το διαφημιστικό κόστος.

- **(data mining):** Εξόρυξη δεδομένων από μεγάλες βάσεις δεδομένων, δηλαδή να ανιχνευθούν βαθιά κρυμμένα πρότυπα μέσα στις εγγραφές μιας βάσης δεδομένων και να προκύψουν συμπεράσματα, τα οποία δεν μπορούν να βρεθούν με απλά ερωτήματα στη βάση δεδομένων.

1.3. Μεθοδολογίες αναγνώρισης προτύπων

Όπως προαναφέραμε στο Τμήμα 1.1 υπάρχουν δύο είδη προτύπων: πρότυπα στο χώρο και πρότυπα στο χρόνο (χρονοσειρές). Χρησιμοποιώντας ευρέως τον όρο αναγνώριση προτύπων, μπορούμε να ανακαλύψουμε σε κάθε ευφυή δραστηριότητα κάποια μορφή αναγνώρισης προτύπων. Δεν είναι δυνατόν να υπάρχει μόνο ένας τρόπος προσέγγισης της θεωρίας αναγνώρισης προτύπων για τον τεράστιο αριθμό και εύρος των εφαρμογών που υπάρχουν. Γενικά, θα μπορούσαμε να χωρίσουμε τις κύριες μεθοδολογίες αναγνώρισης προτύπων σε δύο μεγάλες κατηγορίες:

(1) *Στατιστική αναγνώριση προτύπων* (*statistical pattern recognition*)

(2) *Συντακτική ή δομημένη αναγνώριση προτύπων* (*syntactic or structural pattern recognition*)

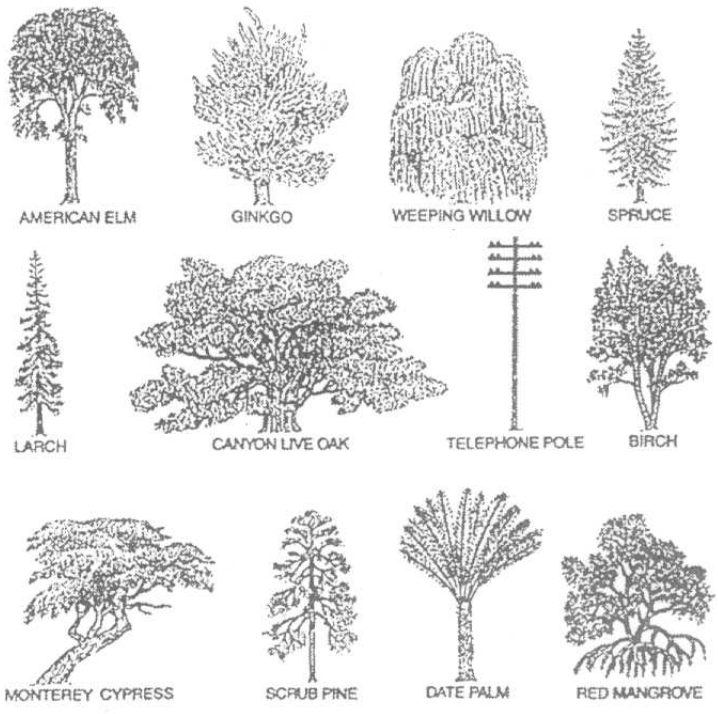
Η *στατιστική αναγνώριση προτύπων* προϋποθέτει μια στατιστική βάση για την ταξινόμηση των αντικειμένων και στηρίζεται στην πιθανοτική φύση των αντικειμένων–προτύπων. Η περιγραφή των αντικειμένων πραγματοποιείται με την εξαγωγή *χαρακτηριστικών γνωρισμάτων* που το περιγράφουν και τη δημιουργία ενός διανύσματος χαρακτηριστικών στοιχείων για κάθε αντικείμενο. Το πρόβλημα ανάγεται στην κατάταξη του διανύσματος χαρακτηριστικών κάθε αντικειμένου στη κατηγορία που ανήκει

χρησιμοποιώντας μαθηματικές – στατιστικές μεθόδους, γραμμική άλγεβρα και θεωρία πιθανοτήτων.

Σε πολλές περιπτώσεις οι σχέσεις μεταξύ των χαρακτηριστικών στοιχείων ενός αντικειμένου φέρουν σπουδαία δομημένη πληροφορία, που μπορεί να χρησιμοποιηθεί για την ταξινόμηση και περιγραφή του αντικειμένου. Η μέθοδος αυτή καλείται *συντακτική ή δομημένη αναγνώριση προτύπων*. Τυπικές προσεγγίσεις της συντακτικής ή δομημένης αναγνώρισης προτύπων είναι η δημιουργία πολύπλοκων ιεραρχικών περιγραφών των προτύπων, τα οποία δημιουργούνται από απλούστερα υπο-πρότυπα. Στο στοιχειώδες επίπεδο του μοντέλου απλά στοιχεία περιγραφής εξάγονται από τα δεδομένα εισόδου.

Τεχνικές οι οποίες χρησιμοποιούνται για την υλοποίηση αλγορίθμων συντακτικής ή δομημένης αναγνώρισης προτύπων είναι δέντρα αποφάσεων (decision trees), λογικοί κανόνες και γραμματικές. Το τελικό αποτέλεσμα είναι μια σειρά κανόνων που περιγράφουν πλήρως τη διαδικασία ταξινόμησης ή μια γραμματική που περιγράφει πλήρως το αντικείμενο.

Οι συντακτικές μεθοδολογίες είναι πολύπλοκες και πολύ ευαίσθητες στην παρουσία θορύβου και γενικά είναι πολύ δύσκολο να ανταποκριθούν αποτελεσματικά σε μικρές παραλλαγές των προτύπων ή και σε ελλιπείς πληροφορίες. Χρησιμοποιούνται κυρίως σε εφαρμογές όπου οι δυνατότητες της στατιστικής αναγνώρισης προτύπων είναι περιορισμένες. Ένα μεγάλο πρόβλημα των συντακτικών μεθοδολογιών είναι ότι δεν είναι πάντα εύκολο να οριστούν πλήρως και χωρίς αμφιβολία τα πρότυπα του προβλήματος. Για παράδειγμα, δεν είναι καθόλου εύκολο να περιγραφεί με λογικούς κανόνες τι είναι ένα δέντρο και να διαχωριστεί αποτελεσματικά από άλλα παρεμφερή αντικείμενα. Η πολυπλοκότητα του προβλήματος απεικονίζεται στο Σχήμα 1.2. Η ανθρώπινη αντίληψη μπορεί εύκολα να διαχωρίσει τα δέντρα και να τα αναγνωρίσει. Αντίθετα, είναι αδύνατη η περιγραφή τους, και συνεπώς ο διαχωρισμός τους, με λογικούς κανόνες.



Σχήμα 1.2: *Το πρόβλημα της συντακτικής αναγνώρισης προτύπων. Πως περιγράφεται ένα δένδρο με συντακτικούς κανόνες;*

Η στατιστική αναγνώριση προτύπων είναι πολύ ισχυρά θεμελιωμένη μαθηματικά ενώ η συντακτική αναγνώριση προτύπων είναι περισσότερο βασισμένη σε λογικούς και διαισθητικούς κανόνες. Το πλήθος και η σειρά στο διάνυσμα των χαρακτηριστικών στοιχείων ενός προτύπου είναι πάντα σταθερό στη στατιστική προσέγγιση, αντίθετα με τη δομημένη προσέγγιση όπου το πλήθος και η σειρά των χαρακτηριστικών στοιχείων μεταβάλλονται από πρότυπο σε πρότυπο.

Το βιβλίο αυτό θα περιοριστεί στη στατιστική αναγνώριση προτύπων. Έτσι, θα εξεταστεί η αναγνώριση προτύπων ως πρόβλημα ταξινόμησης δηλαδή κατάταξης μιας εισόδου σε μια κατηγορία. Βέβαια υπάρχουν

και νέες μεθοδολογίες αναγνώρισης προτύπων οι οποίες χρησιμοποιούν έμπειρα συστήματα, νευρωνικά δίκτυα, γενετικούς αλγορίθμους ή συνδυασμούς αυτών. Η εξέταση όμως αυτών των μεθοδολογιών δεν συμπεριλαμβάνεται στους σκοπούς του βιβλίου.

1.4. Ιστορική αναδρομή

Οι βάσεις της αναγνώρισης προτύπων πρωτοτέθηκαν από τον Πλάτωνα [Bloom91], και τον Αριστοτέλη [Aris96] που πρώτοι έκαναν την διάκριση μεταξύ της *ουσιώδους ιδιότητας* (που μοιράζονται μεταξύ τους τα μέλη μιας κατηγορίας) και της *επουσιώδους ιδιότητας* (που διαφέρει για τα μέλη μιας κατηγορίας). Η Αναγνώριση Προτύπων μπορεί να οριστεί σαν τη διαδικασία που βρίσκει τέτοιες ουσιώδεις ιδιότητες μέσα σε μια κατηγορία αντικειμένων. Οι προβληματισμοί τους παραμένουν βασικοί για την Επιστημονολογία. Ο Αριστοτέλης κατασκεύασε ένα σύστημα ταξινόμησης των ζώων χωρίζοντας τα αρχικά σε αυτά που έχουν κόκκινο αίμα (χονδρικά τα σημερινά σπονδυλωτά) και σε αυτά που δεν έχουν (ασπονδύλωτα). Στην συνέχεια χώρισε τις δύο ομάδες σε μικρότερες χρησιμοποιώντας άλλα χαρακτηριστικά. Στην συνέχεια ο Θεόφραστος έκανε μια ανάλογη ταξινόμηση των φυτών. Η ταξινόμηση ήταν τόσο καλή που μόλις τον 18ο αιώνα ο Carolus Linnaeus κατασκεύασε συστηματικές ταξινομήσεις ζώων, φυτών, πετρωμάτων και ασθενειών χρησιμοποιώντας τις νέες γνώσεις που έφερε ο αιώνας των μεγάλων εξερευνήσεων. Παρόμοια οι Hertzprung και Russell ταξινόμησαν τα άστρα σε κατηγορίες στηριζόμενοι σε δύο μεταβλητές: στην λαμπρότητα και στην θερμοκρασία της επιφάνειάς τους.

Η πρώτη συστηματική προσπάθεια για την μαθηματική μορφοποίηση του προβλήματος έγινε από τον Fisher το 1936 [Fisher36], αλλά έπρεπε να έρθει η ανακάλυψη των υπολογιστών για να θεωρηθεί σαν μια ανεξάρτητη

επιστήμη. Η Αναγνώριση Προτύπων χρησιμοποιείται σε πολλές επιστημονικές περιοχές και σήμερα υπάρχουν δεκάδες επιστημονικά περιοδικά, εκατοντάδες βιβλία και πρακτικά συνεδρίων πάνω σε τομείς της Αναγνώρισης Προτύπων. Κάποιοι επιστημονικοί κλάδοι όπως η Στατιστική [Fuk90], η Μηχανική Μάθηση [Shav90] και τα Τεχνητά Νευρωνικά Δίκτυα [Hertz91] έχουν επεκτείνει κατά πολύ το γνωστικό αντικείμενο της Αναγνώρισης Προτύπων, ενώ ορισμένοι άλλοι, όπως η Υπολογιστική Όραση (Computer Vision) [Fisch87] και η Αναγνώριση Φωνής (Speech recognition) [Rab93] στηρίζονται σχεδόν ολοκληρωτικά σε αυτή. Οι Γνωσιολογική Επιστήμη (Cognitive Science) [Luger94], η Ψυχοβιολογία (Psychobiology)[Uttal73] και η Νευροεπιστήμη (Neuroscience) μελετούν τους μηχανισμούς αναγνώρισης προτύπων των ανθρώπων και των ζώων. Τεχνικές αναγνώρισης προτύπων έχουν εφαρμοστεί έμμεσα ή άμεσα σε όλες σχεδόν τις επιστημονικές περιοχές.

Κεφάλαιο 2

Εισαγωγη στη Στατιστικη Αναγνωριση Προτυπων

2.1. Σύστημα Αναγνώρισης Προτύπων

Ο σχεδιασμός ενός αυτομάτου συστήματος αναγνώρισης προτύπων περιλαμβάνει διάφορα βήματα τα οποία παρουσιάζονται στο Σχήμα 2.1. Ως πρώτο βήμα πραγματοποιείται η αναπαράσταση των προτύπων με μια διαδικασία συλλογής δεδομένων. Όπως αναφέρθηκε στο Τμήμα 1.2, στις περισσότερες εφαρμογές αναγνώρισης προτύπων εμφανίζονται δύο τύποι προτύπων: πρότυπα στο χρόνο (χρονοσειρές) και πρότυπα στο χώρο (γεωμετρικά αντικείμενα). Στην περίπτωση της αναγνώρισης ενός προτύπου που εκφράζεται με μια χρονοσειρά πρέπει αρχικά να γίνει μια

δειγματοληψία του σήματος σε N χρονικές στιγμές όπως φαίνεται στο Σχήμα 2.2.(α) και λαμβάνονται οι τιμές: $X(t_1), X(t_2), \ldots, X(t_N)$. Στην περίπτωση αναγνώρισης γεωμετρικών αντικειμένων, όπως για παράδειγμα αναγνώριση χαρακτήρων, πρέπει να μετρηθούν οι τιμές έντασης των εικονοκυττάρων (pixels) της ψηφιακής αναπαράστασης του αντικειμένου $X_1, X_2, \ldots, X_N$, όπως δείχνει το Σχήμα 2.2.(β). Οι N μετρήσεις δημιουργούν το ακατέργαστο χαρακτηριστικό διάνυσμα του προτύπου $\boldsymbol{X}$. Πρέπει να σημειωθεί ότι οι μετρήσεις ενός προτύπου είναι σε κάθε μέτρηση διαφορετικές έστω και ελάχιστα κάθε φορά, ακόμα και σε κανονικές συνθήκες. Έτσι τα $X(t_i)$ και X_i είναι *τυχαίες μεταβλητές*, το διάνυσμα $\boldsymbol{X}$ ονομάζεται *τυχαίο ακατέργαστο διάνυσμα* του προτύπου και η διαδικασία μέτρησης αποτελεί μια *στοχαστική διαδικασία*.

Όταν οι μετρήσεις είναι στη μορφή πραγματικών αριθμών, είναι συχνά χρήσιμο να βλέπουμε το διάνυσμα του προτύπου σαν ένα σημείο σε ένα N-διάστατο Ευκλείδειο χώρο ο οποίος αποκαλείται *χώρος προτύπων*. Με την υιοθέτηση του διανυσματικού συμβολισμού το γενικό πρόβλημα της αναγνώρισης προτύπων ανάγεται σε ένα γεωμετρικό πρόβλημα, όπου:

$$\boldsymbol{X}_{\text{χώρου}} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{bmatrix}, \quad \boldsymbol{X}_{\text{χρόνου}} = \begin{bmatrix} X(t_1) \\ X(t_2) \\ \vdots \\ X(t_N) \end{bmatrix}. \tag{2.1}$$

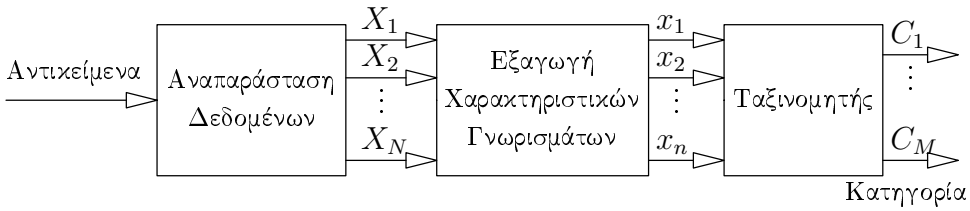
Θα χρησιμοποιείται επίσης η παρακάτω εναλλακτική μορφή, κυρίως μέσα σε κείμενο:

$$\boldsymbol{X}_{\text{χώρου}} = [X_1, \ldots, X_N]^T, \quad \boldsymbol{X}_{\text{χρόνου}} = [X(t_1), \ldots, X(t_N)]^T.$$

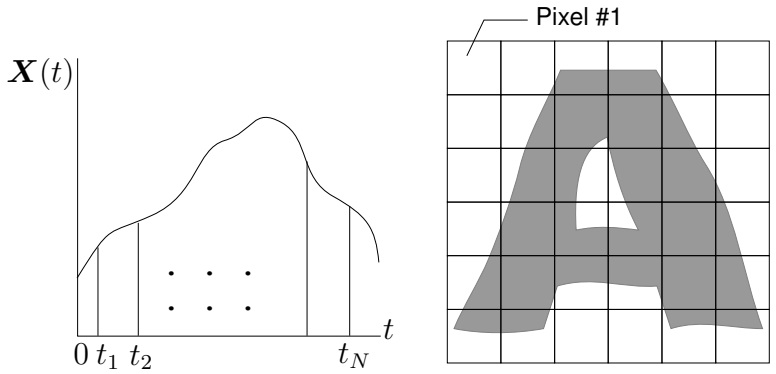
Ο συμβολισμός $\square^T$ συμβολίζει τον ανάστροφο πίνακα.

Στο κείμενο τα διανύσματα θα συμβολίζονται πάντα με έντονα τυπωμένους χαρακτήρες, για παράδειγμα $\boldsymbol{X}$, $\boldsymbol{x}$, $\boldsymbol{w}$, ενώ τα στοιχεία τους με απλούς χαρακτήρες, για παράδειγμα X_i, x_j, w_n.

Το σύνολο των προτύπων τα οποία ανήκουν στην ίδια κατηγορία αντιστοιχούν στα γειτονικά σημεία μέσα σε μια περιοχή του N-διάστατου



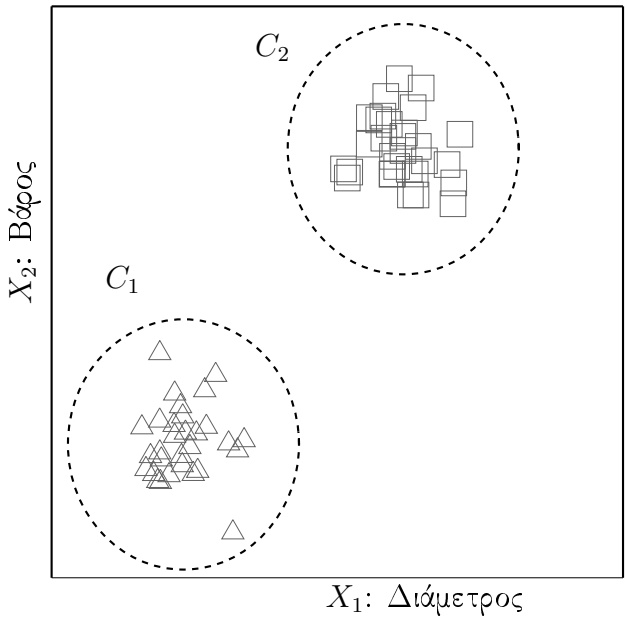
Σχήμα 2.1: *Αυτόματο σύστημα αναγνώρισης προτύπων.*



(α) Χρονοσειρά. *(β) Γεωμετρικό αντικείμενο.*

Σχήμα 2.2: *Δειγματοληψία διανύσματος προτύπου.*

Ευκλείδειου χώρου προτύπων και σχηματίζουν συγκεντρώσεις σημείων οι οποίες καλούνται *ομάδες* (*clusters*). Εκτός από τις διαφορές τιμών που προκύπτουν από τις διαδικασίες μέτρησης, τα μέλη μιας κατηγορίας μπορεί να είναι διαφορετικά για λόγους σχετικούς με την κάθε φορά εφαρμογή. Σε πολλές περιπτώσεις είναι δυνατό να κατασκευαστεί ένα μαθηματικό *αιτιοκρατικό μοντέλο* που να εξηγεί τις διαφορές. Η κατασκευή του μοντέλου απαιτεί μια πολύ βαθιά γνώση της εσωτερικής λειτουργίας του συστήματος καθώς και τη γνώση ενός μεγάλου πλήθους παραμέτρων. Στην πράξη δεν μας ενδιαφέρει να εξηγήσουμε όλες τις λειτουργίες του συστήματος, αλλά απλά να χωρίσουμε όσο το δυνατόν καλύτερα τα πρότυπα σε κατηγορίες. Στα *στοχαστικά μοντέλα* θεωρείται ότι ένα μέρος της παραλλακτικότητας των προτύπων οφείλεται σε τυχαίους λόγους. Ο

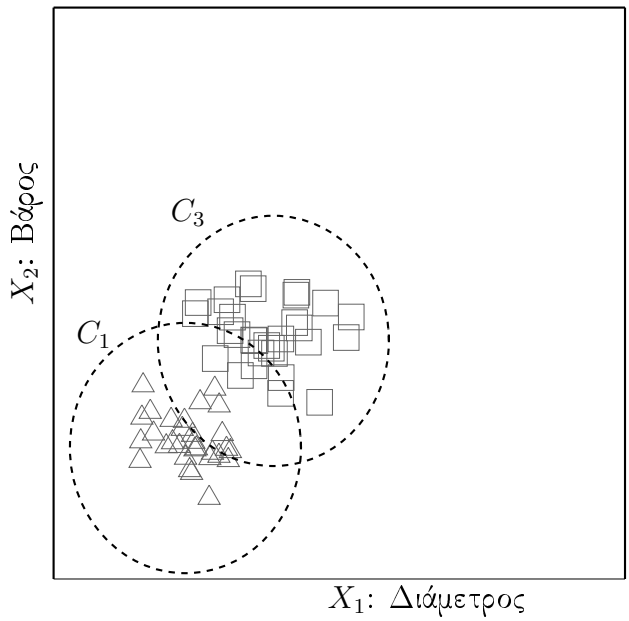


Σχήμα 2.3: *Διαχωρίσιμες κατηγορίες.*

σχεδιαστής του συστήματος πρέπει να βρει στοχαστικές σχέσεις που να περιγράφουν την παραλλακτικότητα, μελετώντας τις στατιστικές ιδιότητες των μελών κάθε κατηγορίας. Άρα, για να μπορέσει να κατασκευαστεί ένα στοχαστικό μοντέλο, θα πρέπει να υπάρχουν διαθέσιμες πολλές μετρήσεις προτύπων κάθε κατηγορίας.

Ένα απλό παράδειγμα παρουσιάζεται στο Σχήμα 2.3 για δύο κατηγορίες προτύπων την C_1 και την C_2, οι οποίες αντιστοιχούν σε μανταρίνια και καρπούζια. Κάθε πρότυπο χαρακτηρίζεται από δύο μετρήσεις: την διάμετρο και το βάρος του φρούτου. Έτσι, τα τυχαία ακατέργαστα διανύσματα των προτύπων έχουν τη μορφή: $\boldsymbol{X} = [X_1, X_2]^T$, όπου το X_1 αντιπροσωπεύει την διάμετρο και το X_2 το βάρος.

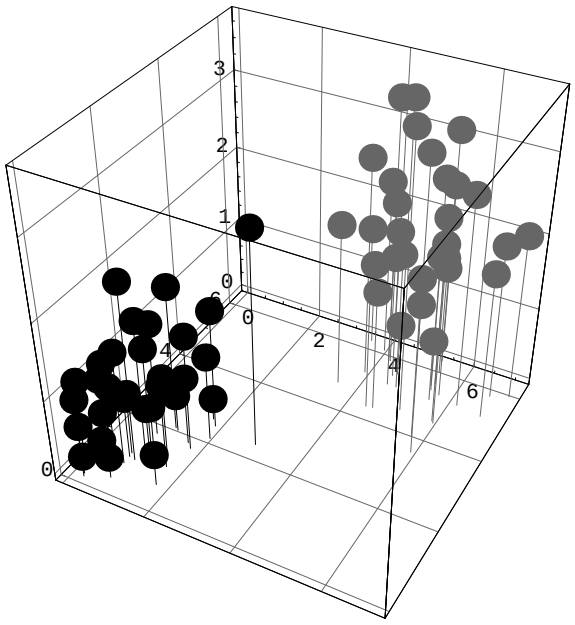
Κάθε τυχαίο διάνυσμα προτύπου φαίνεται σαν ένα σημείο στο δισδιάστατο χώρο του Σχήματος 2.3. Λόγω της φύσης των μετρήσεων οι δύο



Σχήμα 2.4: *Μη διαχωρίσιμες κατηγορίες.*

κατηγορίες του προβλήματος (τα καρπούζια και τα μανταρίνια) δημιουργούν δυο ξεχωριστά και εύκολα διαχωρίσιμα σύνολα στο χώρο προτύπων. Όπως φαίνεται στο Σχήμα 2.3 αυτές οι κατηγορίες δημιουργούν ξεχωριστές μεταξύ τους ομάδες, λόγω της φύσης των μετρήσεων. Όμως, στις περισσότερες πρακτικές περιπτώσεις οι κατηγορίες του προβλήματος δεν σχηματίζουν τόσο ευδιάκριτες ομάδες, αλλά υπάρχει κάποια αλληλοεπικάλυψη μεταξύ τους. Για παράδειγμα, θα υπάρχει κάποια αλληλοεπικάλυψη μεταξύ των κατηγοριών μανταρίνια (C_1) και πορτοκάλια (C_3) όπως φαίνεται και στο Σχήμα 2.4, εάν η διάμετρος και το βάρος χρησιμοποιηθούν ως στοιχεία του διανύσματος προτύπων.

Στο Σχήμα 2.5 παρουσιάζεται μια περίπτωση όπου το τυχαίο ακατέργαστο διάνυσμα προτύπων έχει τρεις συνιστώσες ($N = 3$) και δημιουργούνται δύο ομάδες στον τρισδιάστατο χώρο. Σε περιπτώσεις όπου



Σχήμα 2.5: *Διαχωρίσιμες κατηγορίες στον τρισδιάστατο χώρο.*

$N > 3$, δεν είναι εύκολη η απεικόνιση των πολυδιάσταστων διανυσμάτων προτύπων.

Το δεύτερο βήμα στην αναγνώριση προτύπων είναι η *εξαγωγή χαρακτηριστικών γνωρισμάτων* (*feature extraction*) από τα δεδομένα εισόδου και η ελάττωση της διάστασης των διανυσμάτων προτύπων. Αυτή η διαδικασία ονομάζεται *εξαγωγή χαρακτηριστικών γνωρισμάτων*.

Τα χαρακτηριστικά γνωρίσματα μιας κατηγορίας είναι οι χαρακτηριστικές ιδιότητες οι οποίες είναι κοινές για όλα τα πρότυπα που ανήκουν σε αυτή την κατηγορία και ονομάζονται *ενδοσυνολικά* (*intraset*) γνωρίσματα. Τα χαρακτηριστικά τα οποία αντιπροσωπεύουν τις διαφορές μεταξύ κατηγοριών ονομάζονται *διασυνολικά* (*interset*) γνωρίσματα. Τα στοιχεία των ενδοσυνολικών γνωρισμάτων τα οποία είναι κοινά για όλες τις κατηγορίες δεν περιέχουν καμιά διαχωριστική πληροφορία και μπορούν

να αγνοηθούν. Για παράδειγμα, τα ενδοσυνολικά γνωρίσματα στην κατηγορία πορτοκάλια είναι η διάμετρος, το βάρος και το χρώμα. Παρόμοια είναι και για την κατηγορία μανταρίνια. Το ενδοσυνολικό γνώρισμα–χρώμα είναι κοινό τόσο στα πορτοκάλια όσο και στα μανταρίνια, οπότε δεν περιέχει καμία διαχωριστική πληροφορία. Αντίθετα, τα χαρακτηριστικά γνωρίσματα βάρος και διάμετρος υποδεικνύουν διαφορές μεταξύ των κατηγοριών πορτοκάλια και μανταρίνια και αποτελούν τα διασυνολικά γνωρίσματα του προβλήματος.

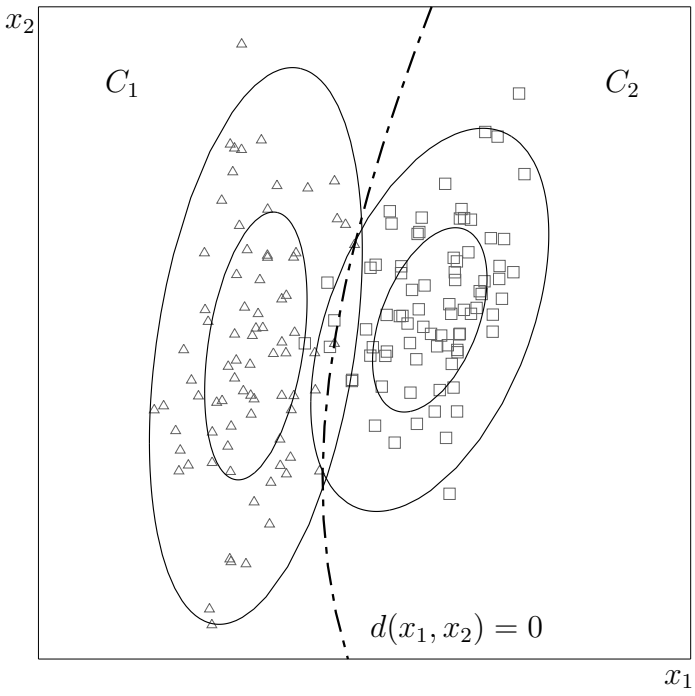
Εάν μπορεί να καθοριστεί ένα πλήρες σύνολο διαχωριστικών γνωρισμάτων για κάθε κατηγορία από τις μετρήσεις των δεδομένων, τότε η αναγνώριση και ταξινόμηση των προτύπων είναι πολύ εύκολη. Όμως, στα περισσότερα πρακτικά προβλήματα αναγνώρισης αυτός ο καθορισμός είναι πολύ δύσκολος, αν όχι αδύνατος. Ευτυχώς, πολλές φορές είναι δυνατόν να βρεθούν ορισμένα διαχωριστικά χαρακτηριστικά γνωρίσματα από τα δεδομένα εισόδου. Σε μια εφαρμογή αναγνώρισης τυπωμένων χαρακτήρων το ακατέργαστο διάνυσμα εισόδου $\boldsymbol{X} = [X_1, X_2, \ \ldots, X_N]^T$ που σχηματίζεται από τις τιμές έντασης των εικονοκυτάρων της ψηφιακής αναπαράστασης του χαρακτήρα, όπως φαίνεται και στο Σχήμα 2.2.(β). Από το $\boldsymbol{X}$ μπορούν να επιλεγούν ως χαρακτηριστικά γνωρίσματα – μεταξύ άλλων – το εμβαδόν του δεξιού τμήματος του χαρακτήρα (x_1), το εμβαδόν του αριστερού τμήματος (x_2) και η περίμετρος του (x_3). Χρησιμοποιώντας αυτά τα χαρακτηριστικά γνωρίσματα μπορεί να εκτιμηθεί η πυκνότητα του χαρακτήρα (z_1) από το λόγο του ολικού εμβαδού του και του τετραγώνου της περιμέτρου του, δηλαδή $z_1 = (x_1 + x_2)/x_3^2$. Επίσης, μπορεί να εκτιμηθεί ο βαθμός συμμετρίας ενός χαρακτήρα γύρω από ένα κάθετο άξονα (z_2) συγκρίνοντας το εμβαδόν του αριστερού με το εμβαδόν του δεξιού τμήματος, δηλαδή $z_2 = x_1/x_2$. Έτσι, το αρχικό διάνυσμα χαρακτηριστικών γνωρισμάτων είναι $\boldsymbol{x} = [x_1, x_2, x_3]^T$. Από αυτό προκύπτει ένα νέο χαρακτηριστικό διάνυσμα $\boldsymbol{z} = [z_1, z_2]^T$ το οποίο περιέχει πιο χρήσιμες πληροφορίες και είναι εκείνο που θα χρησιμοποιηθεί στην ταξινόμηση. Πρέπει να σημειωθεί ότι τα αρχικά στοιχεία x_1, x_2, x_3

αποκαλούνται *ανεξάρτητες μεταβλητές*, ενώ τα z_1 και z_2 αποκαλούνται *εξαρτημένες μεταβλητές* αφού είναι συναρτήσεις των x_1, x_2, x_3.

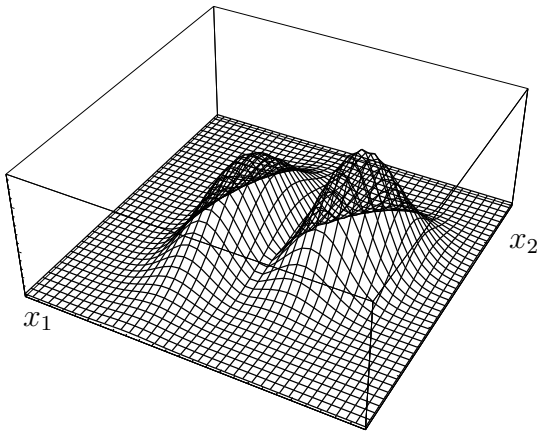
Συνήθως, το διάνυσμα χαρακτηριστικών γνωρισμάτων είναι μικρότερης διάστασης από το ακατέργαστο διάνυσμα εισόδου με συνέπεια την ελάττωση της διάστασης των διανυσμάτων προτύπων. Είναι προφανές ότι η επιλογή των χαρακτηριστικών γνωρισμάτων εξαρτάται από τη συγκεκριμένη εφαρμογή προς επίλυση. Ο σχεδιασμός ενός καλού συνόλου χαρακτηριστικών είναι περισσότερο τέχνη και εμπειρία παρά επιστήμη.

Το τρίτο βήμα στο σχεδιασμό ενός συστήματος αναγνώρισης προτύπων είναι ο καθορισμός βέλτιστων διαδικασιών απόφασης, οι οποίες είναι αναγκαίες για την αναγνώριση και ταξινόμηση. Μετά τη δημιουργία του διανύσματος χαρακτηριστικών γνωρισμάτων του προτύπου, το σύστημα πρέπει να αποφασίσει σε ποια κατηγορία ανήκει το πρότυπο. Έστω ότι το σύστημα αναγνώρισης προτύπων είναι σχεδιασμένο κατά τρόπο τέτοιο, ώστε να αναγνωρίζει M διαφορετικές κατηγορίες προτύπων C_1, C_2, ..., C_M. Τότε ο χώρος προτύπων αποτελείται από M περιοχές, κάθε μια από τις οποίες περιέχει τα σημεία προτύπων αυτής της κατηγορίας. Το πρόβλημα της αναγνώρισης προτύπων ανάγεται στη δημιουργία των *ορίων απόφασης* (*decision boundaries*) τα οποία διαχωρίζουν τις M κατηγορίες βασισμένα στα διανύσματα χαρακτηριστικών γνωρισμάτων των προτύπων.

Το Σχήμα 2.6 παρουσιάζει ένα απλό δισδιάστατο παράδειγμα δύο κατανομών, οι οποίες αντιστοιχούν σε κανονικές και ελαττωματικές καταστάσεις λειτουργίας μιας μηχανής. Τα δισδιάστατα σημεία στο σχήμα είναι τιμές των διανυσμάτων χαρακτηριστικών γνωρισμάτων, ενώ οι καμπύλες αντιστοιχούν σε *συναρτήσεις πυκνότητας πιθανοτήτων* (*probability density functions*) και περικλείουν το 50% και το 90% των διανυσμάτων χαρακτηριστικών γνωρισμάτων των προτύπων κάθε κατηγορίας. Η συνάρτηση πυκνότητας πιθανοτήτων κάθε κατηγορίας παρουσιάζεται σε τρισδιάστατη μορφή στο Σχήμα 2.7. Εάν οι δύο κατανομές του $\boldsymbol{x}$



Σχήμα 2.6: *Δύο στατιστικές κατανομές προτύπων.*



Σχήμα 2.7: *Δύο στατιστικές κατανομές σε τρισδιάστατη μορφή.*

είναι γνωστές εκ των προτέρων, τότε μπορεί να βρεθεί ένα όριο απόφασης $d(x_1, x_2) = 0$, όπως φαίνεται στο Σχήμα 2.6 το οποίο διαιρεί το δισδιάστατο χώρο σε δύο περιοχές. Μετά τον υπολογισμό του ορίου απόφασης είναι δυνατή η ταξινόμηση του διανύσματος χαρακτηριστικών γνωρισμάτων μιας μηχανής σε μια από τις δύο κατηγορίες, ανάλογα με τη τιμή του $d(x_1, x_2)$. Έτσι, εάν $d(x_1, x_2) < 0$ η κατάσταση λειτουργίας της μηχανής ταξινομείται ως κανονική, ενώ εάν $d(x_1, x_2) > 0$ η μηχανή ταξινομείται ως ελαττωματική.

Τα όρια απόφασης μπορούν να υπολογιστούν με διάφορους τρόπους. Εφόσον υπάρχει *"εκ των προτέρων"* (a priori) ολοκληρωμένη γνώση για τα στατιστικά χαρακτηριστικά των προτύπων τα όρια απόφασης μπορούν να καθοριστούν με ακρίβεια, με βάση αυτές τις πληροφορίες. Όταν όμως είναι διαθέσιμη μόνο ποιοτική γνώση των στατιστικών χαρακτηριστικών των προτύπων, τότε είναι μόνο δυνατόν να εκτιμηθούν λογικές προβλέψεις για τη μορφή των ορίων απόφασης. Στη περίπτωση αυτή, τα όρια απόφασης μπορεί να είναι λανθασμένα οπότε για να σχεδιαστεί σωστά ο ταξινομητής πρέπει να εξεταστούν τα χαρακτηριστικά της στατιστικής κατανομής $P(\boldsymbol{x})$ για κάθε κατηγορία. Αυτή η διαδικασία ονομάζεται *εκμάθηση* ή *εκπαίδευση* και τα δείγματα τα οποία χρησιμοποιούνται για το σχεδιασμό του ταξινομητή ονομάζονται *πρότυπα εκπαίδευσης ή εκμάθησης*. Αρχικά καθορίζονται τυχαίες συναρτήσεις αποφάσεων και κατόπιν, μέσα από μια ακολουθία εκπαιδευτικών βημάτων, οι συναρτήσεις αποφάσεων αναπροσαρμόζονται, ώστε να προσεγγίσουν μια βέλτιστη ή απλά ικανοποιητική μορφή.

2.2. Εφαρμογή: Βιομηχανικό Robot

Το εύρος των προβλημάτων που εμφανίζονται σε εφαρμογές αναγνώρισης προτύπων, θα παρουσιαστεί εξετάζοντας τη ρεαλιστική εφαρμογή της αναγνώρισης προτύπων για τον έλεγχο ενός βιομηχανικού robot, καθώς και η εφαρμογή της γενικής μεθοδολογίας του Σχήματος 2.1 στη συγκεκριμένη περίπτωση. Κατά την ανάπτυξη της επίλυσης του προβλήματος θα εισαχθούν έννοιες κλειδιά της αναγνώρισης προτύπων.

2.2.1. Προδιαγραφές του προβλήματος

Το πρόβλημα είναι η δημιουργία ενός προγράμματος ελέγχου βιομηχανικού robot που θα εκτελεί χρέη εργαζόμενου σε μια βιομηχανία. Ένας κυλιόμενος διάδρομος μεταφέρει τα εξαρτήματα μιας μηχανής. Το robot θα παίρνει ένα εξάρτημα από το διάδρομο, θα εξετάζει εάν το εξάρτημα έχει κάποιο κατασκευαστικό ελάττωμα και θα συναρμολογεί τα μη ελαττωματικά εξαρτήματα. Επίσης, το robot θα δέχεται προφορικές εντολές από τον εξουσιοδοτημένο επιστάτη του και θα απευθύνεται σ' αυτόν για συγκεκριμένες εντολές σε περίπτωση προβλήματος.

2.2.2. Ανάλυση του προβλήματος

Για το σχεδιασμό του προγράμματος ελέγχου του robot, θα πρέπει να λυθούν μια σειρά από προβλήματα αναγνώρισης προτύπων:

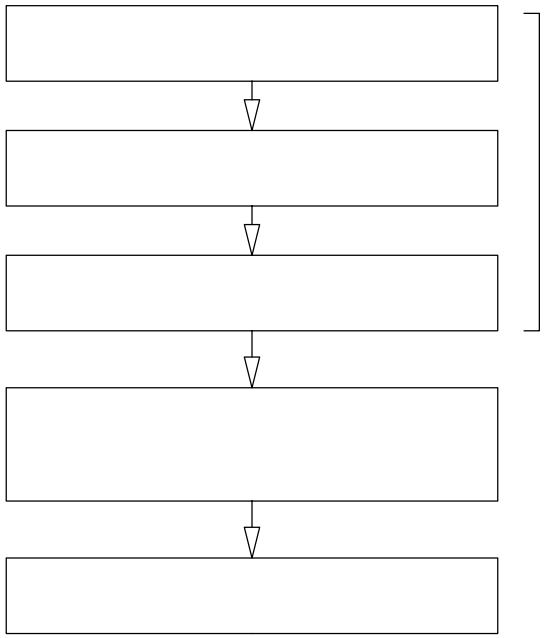
(1) Η αναγνώριση του είδους ενός εξαρτήματος, ανεξάρτητα από τη γωνία παρατήρησης.
(2) Ο έλεγχος του εξαρτήματος για κατασκευαστικά ελαττώματα.
(3) Η κατανόηση των προφορικών εντολών του επιστάτη.
(4) Η αναγνώριση της ταυτότητας του επιστάτη πιθανώς από τη φωνή του.
(5) Η αναγνώριση καταστάσεων όπου απαιτείται ανθρώπινη επέμβαση, όπως για παράδειγμα η αναγνώριση προβλημάτων στη σωστή λειτουργία του ίδιου του robot.

Το καθένα από τα παραπάνω προβλήματα είναι από μόνο του μια εφαρμογή αναγνώρισης προτύπων. Εδώ θα εξεταστεί αναλυτικά μόνο ένα από τα παραπάνω προβλήματα: η αναγνώριση του είδους του εξαρτήματος. Το πρόβλημα εμπίπτει σε μια εξειδικευμένη περιοχή της αναγνώρισης προτύπων η οποία ονομάζεται *αναγνώριση αντικειμένων* (*object recognition*). Τα κυριότερα βήματα για την οπτική αναγνώριση του εξαρτήματος παρουσιάζονται στο Σχήμα 2.8.

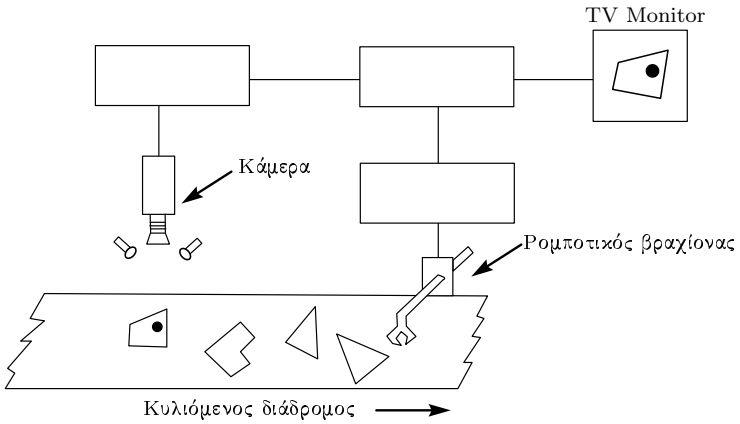
2.2.3. Αναγνώριση του είδους ενός εξαρτήματος

Το Σχήμα 2.9 απεικονίζει το περιβάλλον του προβλήματος. Ένας κυλιόμενος διάδρομος μεταφέρει τα εξαρτήματα. Πάνω από το διάδρομο τοποθετείται ένας οπτικός αισθητήρας ο οποίος συλλαμβάνει την εικόνα του εξαρτήματος. Ανάλογα με τις ειδικές απαιτήσεις της εφαρμογής ο οπτικός αισθητήρας μπορεί να είναι: α) μια σειρά από φωτοαισθητήρες στην πιο απλή περίπτωση, β) μια κάμερα στη πιο συνηθισμένη περίπτωση και γ) ένας συνδυασμός από δύο κάμερες για στερεοσκοπική όραση και τριςδιάστατη ανάλυση σε πιο σύνθετες περιπτώσεις. Στη συγκεκριμένη εφαρμογή χρησιμοποιήθηκε η λύση της απλής κάμερας. Ένας frame grabber μετατρέπει την εικόνα σε ψηφιακή μορφή και εισάγεται στο σύστημα αναγνώρισης εξαρτημάτων. Το σύστημα αναγνώρισης θα αποφασίσει για το είδος του εξαρτήματος.

Τα δεδομένα εισόδου είναι η ψηφιακή εικόνα του εξαρτήματος. Μια πρώτη απλοϊκή προσέγγιση είναι να συγκριθεί η εικόνα με πρότυπες εικόνες εξαρτημάτων, οι οποίες είναι αποθηκευμένες σε μια βάση δεδομένων του συστήματος. Αυτή η προσέγγιση ονομάζεται *ταίριασμα με υποδείγματα* (*template matching*) και χρησιμοποιείται σε ορισμένες χαμηλού κόστους εφαρμογές όπως η οπτική αναγνώριση προτύπων κάτω από ελεγχόμενες συνθήκες. Η προσέγγιση αυτή δεν είναι κατάλληλη στη συγκεκριμένη εφαρμογή για τους παρακάτω λόγους: Η εικόνα του εξαρτήματος συνήθως είναι ελαφρά αλλοιωμένη από τυχαίο θόρυβο ο οποίος προκύπτει κατά τη διαδικασία απεικόνισης και από τη μετατροπή



Σχήμα 2.8: *Λειτουργικό διάγραμμα βιομηχανικού robot.*



Σχήμα 2.9: *Περιβάλλον της αναγνώρισης ενός εξαρτήματος.*

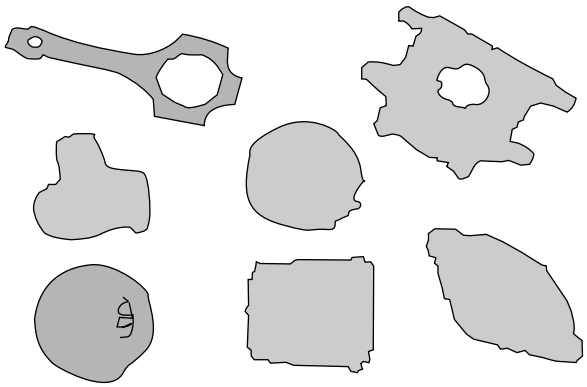
της σε ψηφιακή πληροφορία. Επίσης ο προσανατολισμός του εξαρτήματος πάνω στην μεταφορική ταινία δεν είναι προκαθορισμένος. Έτσι, θα ήταν αναγκαία η αποθήκευση ενός μεγάλου αριθμού υποδειγμάτων για κάθε εξάρτημα ώστε να καλύπτονται όλες οι οπτικές γωνίες. Τέλος, θα πρέπει να ληφθεί υπόψη και το πλήθος της πληροφορίας κάθε εικόνας. Για παράδειγμα αν η διακριτική ανάλυση της ψηφιακής εικόνας είναι 128×128 με βάθος χρώματος 8 bits, τότε απαιτούνται για την αποθήκευση της 16 Kbytes. Είναι φανερό ότι υπάρχει πλεόνασμα πληροφορίας η οποία επιβαρύνει με την επεξεργασία της το σύστημα. Έτσι, για να υλοποιηθεί ένα αποδοτικό σύστημα είναι αναγκαία η εξαγωγή χαρακτηριστικών γνωρισμάτων από τα δεδομένα της εικόνας.

2.2.4. Εξαγωγή χαρακτηριστικών γνωρισμάτων

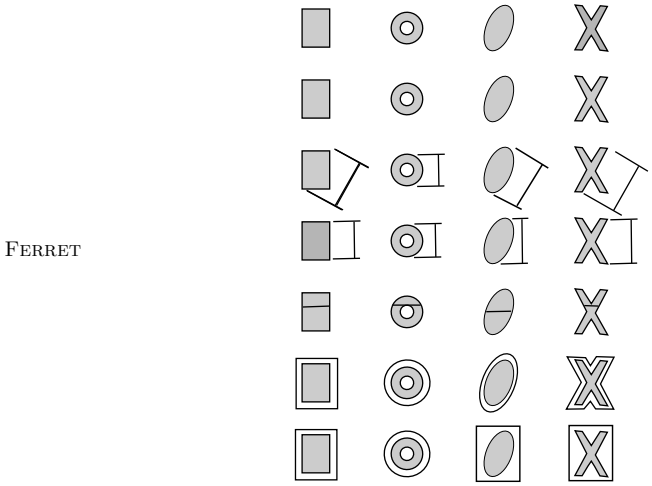
Αρχικά στη ψηφιακή εικόνα θα εφαρμοστούν *μετασχηματισμοί εικόνας*. Δηλαδή διαδικασίες βελτίωσης οι οποίες θα έχουν ως αποτέλεσμα μια νέα ευκρινέστερη εικόνα όπου θα έχουν τονιστεί οι ζητούμενες πληροφορίες. Αυτές οι διαδικασίες περιλαμβάνουν την αφαίρεση του περιθωρίου του εξαρτήματος, τη βελτίωση του χρωματισμού της εικόνας ώστε να παρουσιάζεται με ευκρίνεια το εξάρτημα, τον εντοπισμό των ακμών του εξαρτήματος (edge detection), την αφαίρεση των υπόλοιπων μη αναγκαίων πληροφοριών, ώστε το τελικό αποτέλεσμα να είναι μια δυαδική αναπαράσταση των ακμών του εξαρτήματος.

Ανάλογα με τις ανάγκες κάθε εφαρμογής, πραγματοποιείται περαιτέρω επεξεργασία της εικόνας και εφαρμόζονται αλγόριθμοι *τμηματοποίησης* (segmentation), δηλαδή διαχωρισμός της εικόνας σε περιοχές ενδιαφέροντος. Το αποτέλεσμα της εφαρμογής μετασχηματισμών εικόνας σε ορισμένα εξαρτήματα μηχανής παρουσιάζονται στο Σχήμα 2.10.

Οι μετασχηματισμοί εικόνας (image transformations) και η τμηματοποίηση (segmentation) αποτελούν τμήμα του επιστημονικού κλάδου της *επεξεργασίας εικόνας* (*image processing)* και είναι πέρα από τους σκοπούς αυτού του βιβλίου.



Σχήμα 2.10: *Εξαρτήματα μηχανής σε διάφορες γωνίες.*



Σχήμα 2.11: *Μερικά χαρακτηριστικά γνωρίσματα για βιομηχανικές εφαρμογές.*

Τελικά, από τις ακμές και τα τμήματα του εξαρτήματος υπολογίζονται τα χαρακτηριστικά του γνωρίσματα. Στη συγκεκριμένη εφαρμογή τα χαρακτηριστικά γνωρίσματα κάθε εξαρτήματος είναι:

(1) Το εμβαδόν της προβολής του εξαρτήματος x_1

(2) Η εξωτερική περίμετρος του εξαρτήματος x_2

(3) Η μέγιστη διάσταση του εξαρτήματος x_3

(4) Το πλήθος των οπών του x_4

Το Σχήμα 2.11 παρουσιάζει ένα πλήθος ευρέως χρησιμοποιούμενων χαρακτηριστικών γνωρισμάτων που εξάγονται σε παρόμοιες βιομηχανικές ρομποτικές εφαρμογές καθώς και τη χρησιμότητα τους. Η σωστή επιλογή των χαρακτηριστικών γνωρισμάτων για κάθε εφαρμογή είναι μια διαδικασία κλειδί για την υλοποίηση οποιουδήποτε συστήματος αναγνώρισης προτύπων.

2.2.5. Αναγνώριση του εξαρτήματος

Μετά την εξαγωγή των χαρακτηριστικών γνωρισμάτων χρησιμοποιούνται τυποποιημένοι αλγόριθμοι για την αναγνώρισης προτύπων. Οι αλγόριθμοι αυτοί είναι γενικοί και μπορούν να χρησιμοποιηθούν σε ένα πολύ μεγάλο εύρος εφαρμογών. Η παρουσίαση και ανάπτυξη των πιο διαδεδομένων αλγορίθμων αναγνώρισης προτύπων είναι το κύριο θέμα του βιβλίου.

2.3. Ερευνητικά Θέματα Αναγνώρισης Προτύπων

Έχοντας παρουσιάσει μια γενική εικόνα των θεμάτων που απασχολούν τον επιστημονικό κλάδο της αναγνώρισης προτύπων, συγκεκριμένα ερευνητικά θέματα του κλάδου θα αναφερθούν αναλυτικά. Όταν εμφανίζεται μια νέα εφαρμογή απαιτείται σημαντική ερευνητική και κατασκευαστική προσπάθεια. Καθώς τα περισσότερα προβλήματα αναγνώρισης προτύπων αφορούν συγκεκριμένες γνωστικές περιοχές, η λύση τους εξαρτάται από την γνώση, την εμπειρία και διορατικότητα των ερευνητών. Ωστόσο, ορισμένα από τα προβλήματα είναι γενικά στη φύση τους, συχνά δύσκολα αλλά ταυτόχρονα ενδιαφέροντα και απαιτούν ιδιαίτερη προσοχή.

2.3.1. Εξαγωγή χαρακτηριστικών γνωρισμάτων

Το ιδεατό όριο μεταξύ εξαγωγής χαρακτηριστικών γνωρισμάτων και ταξινόμησης είναι κάπως αυθαίρετο. Ένας ιδανικός εξαγωγέας χαρακτηριστικών δημιουργεί μια αναπαράσταση που κάνει την ταξινόμηση μια απλή διαδικασία. Παρόμοια, ένας ιδανικός ταξινομητής δεν έχει ανάγκη από ένα εξειδικευμένο εξαγωγέα χαρακτηριστικών γνωρισμάτων. Η διάκριση λοιπόν μεταξύ τους είναι περισσότερο θεωρητική παρά πρακτική. Γενικά, η διαδικασία εξαγωγής χαρακτηριστικών γνωρισμάτων εξαρτάται από την συγκεκριμένη εφαρμογή και το πεδίο του προβλήματος, αντίθετα από την διαδικασία της ταξινόμησης η οποία είναι πιο γενική. Ένας ταξινομητής μπορεί να χρησιμοποιηθεί σε πολλές εφαρμογές, ενώ ένας εξαγωγέας χαρακτηριστικών γνωρισμάτων για την αναγνώριση δακτυλικών αποτυπωμάτων θα ήταν άχρηστος σε μία εφαρμογή αναγνώρισης χαρακτήρων.

2.3.2. Θόρυβος

Θόρυβος μπορεί να οριστεί σε γενικές γραμμές κάθε ιδιότητα κατά την διαδικασία μέτρησης ενός προτύπου που δεν σχετίζεται με το πραγματικό του μοντέλο αλλά με την τυχαιότητα στον κόσμο ή στο περιθώριο λάθους των αισθητήρων. Όλες σχεδόν οι πρακτικές εφαρμογές της αναγνώρισης προτύπων σχετίζονται με κάποια μορφή θορύβου στις μετρήσεις των δεδομένων. Ένα σημαντικό ερευνητικό θέμα είναι η ενσωμάτωση γνώσης της πηγής θορύβου στο σύστημα αναγνώρισης προτύπων καθώς και η εξακρίβωση εάν η διαφοροποίηση σε κάποια μέτρηση σχετίζεται με θόρυβο ή άλλους παράγοντες.

2.3.3. Πολυπλοκότητα μοντέλου

Είναι δυνατή η κατασκευή ενός πολύπλοκου συστήματος αναγνώρισης προτύπων το οποίο να πραγματοποιεί τέλεια ταξινόμηση στα πρότυπα εκπαίδευσης αλλά να αποτυγχάνει στην ταξινόμηση πραγματικών δεδομένων; Ορισμένες φορές αυτό συμβαίνει διότι το πολύπλοκο σύστημα αναγνώρισης προτύπων εξαρτάται από τις ιδιομορφίες του τυχαίου συνόλου

δεδομένων εκπαίδευσης και όχι από τις ιδιότητες των πραγματικών δεδομένων. Μία από τις πλέον σημαντικές περιοχές έρευνας στην στατιστική αναγνώριση προτύπων είναι ο καθορισμός της *πολυπλοκότητας* ενός μοντέλου, δηλαδή το μοντέλο να μην είναι τόσο απλό ώστε να μην μπορεί να ξεχωρίσει διαφορές μεταξύ κατηγοριών και ταυτόχρονα όχι τόσο πολύπλοκο ώστε να αποτυγχάνει στην ταξινόμηση πραγματικών δεδομένων.

2.3.4. Επιλογή μοντέλου ταξινομητή

Υπάρχουν πολλά μοντέλα πιθανών ταξινομητών που μπορεί να χρησιμοποιηθούν για να λύσουν ένα πρόβλημα. Πως επιλέγει κάποιος ερευνητής το πλέον κατάλληλο μοντέλο για μία συγκεκριμένη εφαρμογή; Πως αποφασίζει κανείς να απορρίψει ένα μοντέλο και να δοκιμάσει ένα άλλο; Εκτός από την γνωστή μέθοδο της διαδοχικής δοκιμής και απόρριψης, υπάρχει κάποια πιο συστηματική μέθοδος επιλογής μοντέλου;

2.3.5. Μερική έλλειψη τιμών χαρακτηριστικών γνωρισμάτων

Συχνά, για διάφορους λόγους, δεν είναι διαθέσιμες όλες οι τιμές χαρακτηριστικών γνωρισμάτων των προτύπων. Πως μπορεί ο ταξινομητής να πραγματοποιήσει την βέλτιστη απόφαση χρησιμοποιώντας μόνο τα υπάρχοντα δεδομένα; Η απλοϊκή μέθοδος να θεωρούνται μηδέν οι τιμές των ανύπαρκτων δεδομένων ή ο μέσος όρος των υπόλοιπων τιμών των δεδομένων, πιθανά δεν είναι ο βέλτιστος τρόπος καθορισμού των τιμών. Παρόμοια, σε ορισμένες περιπτώσεις, υπάρχει έλλειψη τιμών στα πρότυπα εκπαίδευσης κατά την δημιουργία του συστήματος αναγνώρισης. Πως μπορεί να εκπαιδευτεί ο ταξινομητής όταν υπάρχει έλλειψη ορισμένων τιμών χαρακτηριστικών γνωρισμάτων;

2.3.6. Τμηματοποίηση

Στην εφαρμογή του βιομηχανικού robot θεωρείται ότι τα εξαρτήματα για αναγνώριση είναι μη επικαλυπτόμενα στην ταινία μεταφοράς. Πρακτικά,

είναι δυνατόν ορισμένα από αυτά τα εξαρτήματα να είναι επικαλυπτόμενα. Σε αυτή την περίπτωση τα πρότυπα πρέπει να τμηματοποιηθούν και να αναγνωριστούν από χαρακτηριστικά γνωρίσματα των τμημάτων τους. Η *τμηματοποίηση* είναι από τα πιο δύσκολα ερευνητικά προβλήματα σε ορισμένους κλάδους της αναγνώρισης προτύπων, όπως για παράδειγμα στην αναγνώριση ομιλίας.

2.3.7. Αμετάβλητοι μετασχηματισμοί

Για την βέλτιστη υλοποίηση ενός ταξινομητή πρέπει να αντιμετωπιστεί το πρόβλημα της *αμεταβλητότητας*. Στην εφαρμογή του βιομηχανικού robot η θέση των εξαρτημάτων στην ταινία μεταφοράς δεν είναι προκαθορισμένη άρα ο ταξινομητής πρέπει να αναγνωρίζει τα εξαρτήματα ανεξάρτητα της απόλυτης θέσης τους στην ταινία μεταφοράς. Έτσι, ο ταξινομητής πρέπει να είναι αμετάβλητος του μετασχηματισμού *μεταφοράς* σε οποιαδήποτε διεύθυνση. Επιπλέον, ο προσανατολισμός των εξαρτημάτων στην ταινία μεταφοράς δεν είναι προκαθορισμένος. Έτσι, ο ταξινομητής πρέπει να είναι αμετάβλητος στον μετασχηματισμό *περιστροφής*. Σε πολλές περιπτώσεις το μέγεθος της απεικόνισης μπορεί να μην είναι σταθερό και να μην σχετίζεται με την κατηγοριοποίηση. Για παράδειγμα, το μέγεθος της απεικόνισης ενός εξαρτήματος στην ταινία μεταφοράς εξαρτάται από την θέση της κάμερας. Ελαφρές διακυμάνσεις στην τοποθέτηση της κάμερας έχουν σαν αποτέλεσμα την αλλαγή του μεγέθους ενός εξαρτήματος κατά την απεικόνιση του. Έτσι, ο ταξινομητής πρέπει να είναι αμετάβλητος στον μετασχηματισμό *κλιμάκωσης*.

Στην αναγνώριση προτύπων χρησιμοποιούνται πολύπλοκοι μετασχηματισμοί οι οποία σχετίζονται με συγκεκριμένες εφαρμογές. Για παράδειγμα, σε μία εφαρμογή οπτικής αναγνώρισης χειρόγραφων χαρακτήρων, ο ταξινομητής πρέπει να μην είναι ευαίσθητος στο πάχος της γραμμής των χαρακτήρων. Παρόμοια, στην αναγνώριση εικόνας θα πρέπει το σύστημα αναγνώρισης να ανταποκριθεί σε αλλαγές του φωτισμού ή στην παρουσίαση σκιών.

Σε όλες τις περιπτώσεις αμεταβλητότητας υπάρχουν τα εξής θέματα: Πως προκαθορίζεται η παρουσίαση αμεταβλητότητας; Πως ενσωματώνεται αποτελεσματικά τέτοια γνώση στο σύστημα αναγνώρισης; Έστω ότι δεν είναι γνωστή εκ των προτέρων η παρουσίαση μίας συγκεκριμένης αμεταβλητότητας. Πως θα γίνει εκμάθηση στο σύστημα αναγνώρισης αυτής της αμεταβλητότητας;

2.3.8. Κόστος και ρίσκο

Πρέπει να γίνει κατανοητό ότι ένας ταξινομητής αποτελεί μέρος ενός ολοκληρωμένου συστήματος και χρησιμοποιείται σαν συμβουλευτικό όργανο. Στην εφαρμογή του βιομηχανικού robot χρησιμοποιείται για την αναγνώριση τής ταυτότητας ενός εξαρτήματος και στην εξέταση κατασκευαστικών ελαττωμάτων των εξαρτημάτων. Το πιο απλό ρίσκο στην συγκεκριμένη εφαρμογή είναι η λάθος ταξινόμηση ενός εξαρτήματος. Το επόμενο στάδιο ρίσκου είναι η μη αναγνώριση ενός κατασκευαστικού ελαττώματος ενός αναγνωρισμένου εξαρτήματος. Ποιο είναι το ποσοστό των εξαρτημάτων που δεν αναγνωρίζονται σωστά και πιο το ποσοστό των εξαρτημάτων με κατασκευαστικά ελαττώματα που δεν αναγνωρίστηκάν; Οι συμβουλευτικές ενέργειες ενός ταξινομητή έχουν ένα συσχετιζόμενο κόστος και ρίσκο. Ο σχεδιασμός ενός ταξινομητή προϋποθέτει την ελαχιστοποίηση του ολικού κόστους ή ρίσκου. Πως ενσωματώνονται αυτά σε ένα ταξινομητή και πως επηρεάζουν την απόκριση του ταξινομητή; Τελικά, μπορεί να εκτιμηθεί το ολικό ρίσκο ώστε να αποφασιστεί εάν ο ταξινομητής είναι αποδεκτός πριν εφαρμοστεί στον πραγματικό κόσμο; Μπορεί να εκτιμηθεί το ελάχιστο δυνατό ρίσκο οποιουδήποτε ταξινομητή ώστε να υπάρχει σημείο σύγκρισης ενός προτεινόμενου ταξινομητή με τον ιδεατό ή το πρόβλημα είναι γενικά πολύπλοκο;

2.3.9. Υπολογιστική πολυπλοκότητα

Ορισμένα προβλήματα αναγνώρισης προτύπων μπορεί να λυθούν χρησιμοποιώντας αλγορίθμους οι οποίοι όμως δεν είναι πρακτικοί. Για παράδειγμα, στην εφαρμογή του βιομηχανικού robot για την αναγνώριση του είδους ενός εξαρτήματος, τα δεδομένα εισόδου είναι η ψηφιακή απεικόνιση του εξαρτήματος. Όμως, σε αυτή την εφαρμογή, η απευθείας ταξινόμηση της ψηφιακής απεικόνισης δεν είναι κατάλληλη λόγω θορύβου, προσανατολισμού των εξαρτημάτων, χωρητικότητα μνήμης και υπολογιστικής ταχύτητας. Έτσι, η υπολογιστική πολυπλοκότητα διαφόρων αλγορίθμων είναι πολύ σημαντική, κυρίως για πρακτικές εφαρμογές.

Σε ποιο γενικές γραμμές μπορεί να τεθεί το ερώτημα της υπολογιστικής πολυπλοκότητας ενός αλγορίθμου σε σχέση με την διάσταση των δεδομένων εισόδου ή τον αριθμό των προτύπων για ταξινόμηση ή τον αριθμό των κατηγοριών. Ποια είναι η ανταλλαγή μεταξύ υπολογιστικής πολυπλοκότητας και επίδοσης; Σε ορισμένες περιπτώσεις μπορεί να σχεδιαστούν τέλειοι ταξινομητές άλλα να μην έχουν πρακτική εφαρμογή διότι δεν βρίσκονται μέσα στα όρια των μηχανικών και υπολογιστικών περιορισμών. Πως μπορούν να βελτιστοποιηθούν αυτοί οι περιορισμοί;

2.1 Να βρεθεί ένα σύνολο αριθμητικών χαρακτηριστικών γνωρισμάτων ώστε να είναι δυνατή η αναγνώριση μοντέλων αυτοκινήτων από πλάγιες φωτογραφίες τους. Τα μήκη δεν μπορούν να χρησιμοποιηθούν ως χαρακτηριστικά στοιχεία διότι είναι άγνωστη η απόσταση τους από την κάμερα. Υπάρχουν όμως χαρακτηριστικά που παραμένουν αναλλοίωτα με την αλλαγή κλίμακας. Από πέντε διαφορετικά μοντέλα αυτοκινήτων να υπολογίσετε αυτά τα χαρακτηριστικά. **Συμβουλή**: Να εξεταστεί το ύψος σε σχέση με το μήκος ενός αυτοκινήτου.

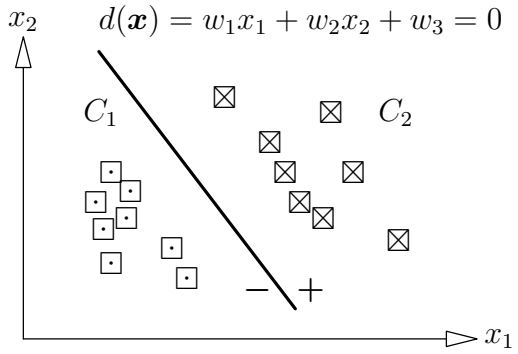
2.2 Να βρεθεί ένα σύνολο αριθμητικών χαρακτηριστικών γνωρισμάτων για την ταξινόμηση εικόνων που περιέχουν μη επικαλυπτόμενα πορτοκάλια, μήλα, μπανάνες και αχλάδια.

Κεφάλαιο 3

Ταξινομηση προτυπων με συναρτησεις αποφασης

3.1. Απλές Γραμμικές Συναρτήσεις Απόφασης

Η κύρια λειτουργία ενός συστήματος αναγνώρισης προτύπων είναι η ταξινόμηση προτύπων σε κατηγορίες. Για να επιτευχθεί αυτό είναι αναγκαίο να καθοριστούν οι κανόνες πάνω στους οποίους το σύστημα θα βασιστεί για να πραγματοποιήσει την ταξινόμηση. Μια πολύ σημαντική προσέγγιση σε αυτό το πρόβλημα είναι η χρησιμοποίηση των *συναρτήσεων απόφασης* (*decision functions*), τις οποίες αρχικά παρουσιάσαμε



Σχήμα 3.1: *Απλή συνάρτηση απόφασης για δύο κατηγορίες.*

στο Κεφάλαιο 2. Το Σχήμα 3.1 παρουσιάζει μια απλή δισδιάστατη περίπτωση, όπου δύο υποθετικές κατηγορίες προτύπων μπορούν πολύ εύκολα να διαχωριστούν με μια ευθεία γραμμή. Η γενική μορφή μιας δισδιάστατης γραμμικής συνάρτησης απόφασης είναι:

$$d(\boldsymbol{x}) = w_1x_1 + w_2x_2 + w_3 = 0, \tag{3.1}$$

όπου τα w_1, w_2, w_3 είναι οι παράμετροι που καθορίζουν την θέση και την κλίση της γραμμής και x_1, x_2 αποτελούν τα στοιχεία του ανύσματος χαρακτηριστικών γνωρισμάτων του προτύπου $\boldsymbol{x} = [x_1, x_2]^T$. Είναι προφανές από το Σχήμα 3.1 ότι οποιοδήποτε πρότυπο $\boldsymbol{x}$, το οποίο ανήκει στην κατηγορία C_2, αντικατασταθεί στην Εξίσωση (3.1) τότε η τιμή της $d(\boldsymbol{x})$ θα είναι θετική. Αντίθετα η τιμή της θα είναι αρνητική για τα πρότυπα $\boldsymbol{x}$ που ανήκουν στην κατηγορία C_1. Έτσι, η $d(\boldsymbol{x})$ μπορεί να χρησιμοποιηθεί σαν μια *συνάρτηση απόφασης* ή *διαχωρισμού* (*decision or discriminant function*), αφού για οποιοδήποτε πρότυπο $\boldsymbol{x}$ μπορεί να αποφασιστεί σε ποια κατηγορία ανήκει ανάλογα με το πρόσημο της. Εάν το πρότυπο βρίσκεται πάνω στη συνάρτηση απόφασης, δηλαδή $d(\boldsymbol{x}) = 0$, τότε έχουμε μια απροσδιόριστη κατάσταση και δεν μπορεί να αποφασιστεί σε ποια κατηγορία ανήκει.

Στην γενική περίπτωση η συνάρτηση απόφασης μπορεί να έχει μια οποιαδήποτε μορφή. Η επιτυχία του παραπάνω τρόπου ταξινόμησης προτύπων εξαρτάται από δύο παράγοντες:

(1) Από την γενική μορφή της συνάρτησης $d(\boldsymbol{x})$.

(2) Από την ικανότητα καθορισμού των συντελεστών της $d(\boldsymbol{x})$.

Το πρώτο πρόβλημα σχετίζεται με τις γεωμετρικές ιδιότητες κάθε κατηγορίας προτύπων. Δυστυχώς, αν δεν υπάρχει κάποια "εκ των προτέρων" γνώση για τα γεωμετρικά χαρακτηριστικά των κατηγοριών τότε ο μόνος τρόπος καθορισμού της μορφής της συνάρτησης απόφασης είναι με μια επαναληπτική διαδικασία δοκιμής και σφάλματος. Από τη στιγμή που θα επιλεγεί η μορφή της συνάρτησης, ή των συναρτήσεων απόφασης εάν υπάρχουν περισσότερες από δύο κατηγορίες, το πρόβλημα ανάγεται στον καθορισμό των συντελεστών της συνάρτησης.

Εάν οι κατηγορίες προτύπων είναι διαχωρίσιμες χρησιμοποιώντας συναρτήσεις απόφασης, στο βιβλίο αυτό θα παρουσιάσουμε αλγόριθμους οι οποίοι χρησιμοποιούν δείγματα προτύπων από κάθε κατηγορία, ώστε να εκτιμηθούν οι τιμές των συντελεστών, οι οποίοι χαρακτηρίζουν αυτές τις συναρτήσεις.

3.2. Γραμμικές Συναρτήσεις Απόφασης

Η απλή δισδιάστατη γραμμική συνάρτηση απόφασης του Σχήματος 3.1 μπορεί να γενικευτεί εύκολα για την n-διάστατη περίπτωση. Η μορφή μιας n-διάστατης γραμμικής συνάρτησης απόφασης είναι:

$$\begin{aligned} d(\boldsymbol{x}) &= w_1 x_1 + w_2 x_2 + \ldots + w_n x_n + w_{n+1} \\ &= \boldsymbol{w}^T \boldsymbol{x} + w_{n+1}. \end{aligned} \tag{3.2}$$

Στην περίπτωση που $n = 2$ η Εξίσωση (3.2) είναι η εξίσωση μιας ευθείας γραμμής που διαχωρίζει τον δισδιάστατο χώρο όπως δείχνει και το Σχήμα 3.1, ενώ όταν $n = 3$ ανάγεται στην εξίσωση ενός επιπέδου που

διαχωρίζει τον τρισδιάστο χώρο. Στην περίπτωση που $n > 3$ ανάγεται στην εξίσωση ενός υπερεπιπέδου διάστασης $n-1$, που χωρίζει στα δύο ένα υπερχώρο διάστασης n.

Η Εξίσωση (3.2) μπορεί να απλοποιηθεί μαθηματικά αν χρησιμοποιηθεί το *επαυξημένο* (*augment*) χαρακτηριστικό άνυσμα του προτύπου $\boldsymbol{x}$, που προκύπτει εάν εισάγουμε ένα επιπλέον στοιχείο που έχει πάντα την τιμή 1. Η Εξίσωση (3.2) παίρνει τότε την απλοποιημένη μορφή:

$$d(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x}, \tag{3.3}$$

όπου $\boldsymbol{x} = [x_1, x_2, \ldots, x_n, 1]^T$ είναι το επαυξημένο χαρακτηριστικό διάνυσμα και $\boldsymbol{w} = [w_1, w_2, \ldots, w_n, w_{n+1}]^T$ ονομάζεται το *διάνυσμα παραμέτρων ή διάνυσμα βαρών*. Η Εξίσωση (3.3) αναπαριστά ένα υπερεπίπεδο το οποίο περνάει από την αρχή των αξόνων.

Σε ένα πρόβλημα δύο κατηγοριών η συνάρτηση απόφασης $d(\boldsymbol{x})$ έχει την ιδιότητα:

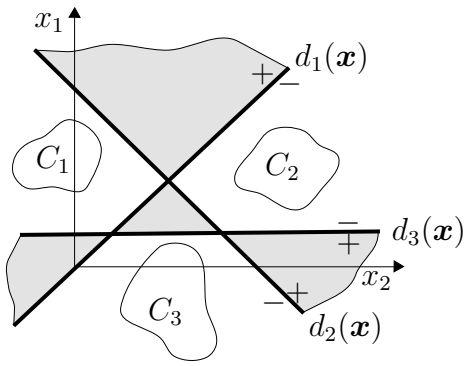
$$d(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} \begin{cases} > 0 & \text{εάν } \boldsymbol{x} \in C_1, \\ < 0 & \text{εάν } \boldsymbol{x} \in C_2. \end{cases} \tag{3.4}$$

Εάν υπάρχουν πάνω από δύο κατηγορίες γραμμικά διαχωρίσιμες, τότε ο γραμμικός διαχωρισμός γίνεται πιο πολύπλοκος και η Εξίσωση (3.4) για M κατηγορίες παίρνει την μορφή:

$$d_i(\boldsymbol{x}) = \boldsymbol{w}_i^T \boldsymbol{x} \begin{cases} > 0 & \text{εάν } \boldsymbol{x} \in C_i, \\ < 0 & \text{διαφορετικά}, \end{cases} \tag{3.5}$$

όπου $i = 1, 2, \ldots, M$ και $\boldsymbol{w}_i = [w_{i1}, w_{i2}, \ldots, w_{in}, w_{i(n+1)}]^T$ είναι το διάνυσμα βαρών για την i-οστή συνάρτηση απόφασης $d_i(\boldsymbol{x})$. Με τον αυτόν τον τρόπο γραμμικού διαχωρισμού δημιουργείται πρόβλημα εάν υπάρχουν σημεία του χώρου για τα οποία περισσότερες από μία συναρτήσεις απόφασης είναι θετικές. Τότε για τα πρότυπα που εμπίπτουν σε αυτές τις περιοχές δεν μπορεί να προσδιορισθεί η κατηγορία τους.

Το Σχήμα 3.2 παρουσιάζει ένα απλό παράδειγμα γραμμικού διαχωρισμού για τρεις κατηγορίες ($M = 3$). Στη συγκεκριμένη περίπτωση



Σχήμα 3.2: *Γραμμικές συναρτήσεις απόφασης με την πρώτη μεθοδολογία.*

υπάρχει μια συνάρτηση απόφασης που διαχωρίζει μια κατηγορία από τις υπόλοιπες. Για τα πρότυπα που εμπίπτουν στις σκιασμένες περιοχές του Σχήματος 3.2, περισσότερες από μία συναρτήσεις απόφασης είναι θετικές και για τα πρότυπα αυτά δεν μπορεί να αποφασιστεί η κατηγορία στην οποία ανήκουν.

Ένας δεύτερος τρόπος γραμμικού διαχωρισμού είναι η εύρεση συναρτήσεων απόφασης, όπου κάθε κατηγορία διαχωρίζεται από κάθε μια από τις υπόλοιπες, δηλαδή οι κατηγορίες είναι αμοιβαία ανά δύο διαχωρίσιμες. Τότε δημιουργούνται $\frac{M(M-1)}{2}$ συναρτήσεις απόφασης, όσες ο συνδυασμός M κατηγοριών ανά δύο, της μορφής:

$$d_{ij}(\boldsymbol{x}) = \boldsymbol{w}_{ij}^T \boldsymbol{x}, \tag{3.6}$$

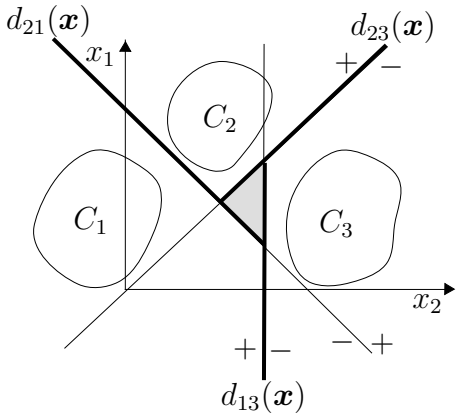
με την ιδιότητα ότι εάν το πρότυπο $\boldsymbol{x}$ ανήκει στην κατηγορία C_i, τότε:

$$d_{ij}(\boldsymbol{x}) = \boldsymbol{w}_{ij}^T \boldsymbol{x} > 0, \quad \forall\, j \neq i. \tag{3.7}$$

Οι συναρτήσεις αυτές έχουν τη συμμετρική ιδιότητα, δηλαδή:

$$d_{ij}(\boldsymbol{x}) = -d_{ji}(\boldsymbol{x}). \tag{3.8}$$

Όμως, και με αυτό τον τρόπο γραμμικού διαχωρισμού, είναι δυνατό να υπάρξουν περιπτώσεις, όπως φαίνεται από την σκιασμένη περιοχή του



Σχήμα 3.3: *Γραμμικές συναρτήσεις απόφασης με την δεύτερη μεθοδολογία.*

Σχήματος 3.3, όπου η συγκυρία της Εξίσωσης (3.7) είναι αδύνατη για τις συγκεκριμένες συναρτήσεις απόφασης της κατηγορίας C_i.

Τελικά, ένας τρίτος διαχωριστικός κανόνας είναι να υπάρχουν M συναρτήσεις απόφασης:

$$d_i(\boldsymbol{x}) = \boldsymbol{w}_i^T \boldsymbol{x}, \quad i = 1, 2, \ldots, M, \tag{3.9}$$

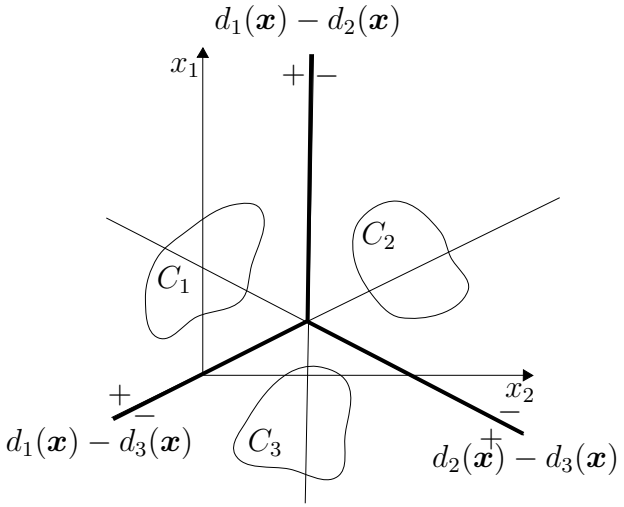
με την ιδιότητα ότι εάν το $\boldsymbol{x}$ ανήκει στην κατηγορία C_i, τότε:

$$d_i(\boldsymbol{x}) > d_j(\boldsymbol{x}), \quad \forall\, i \neq j. \tag{3.10}$$

Η μεθοδολογία είναι ουσιαστικά μια υποπερίπτωση της Μεθοδολογίας 2 μιας και μπορεί να οριστούν συναρτήσεις απόφασης της μορφής:

$$\begin{aligned} d_{ij}(\boldsymbol{x}) &= d_i(\boldsymbol{x}) - d_j(\boldsymbol{x}) \\ &= (\boldsymbol{w}_i - \boldsymbol{w}_j)^T \boldsymbol{x} \\ &= \boldsymbol{w}_{ij}^T \boldsymbol{x}, \end{aligned} \tag{3.11}$$

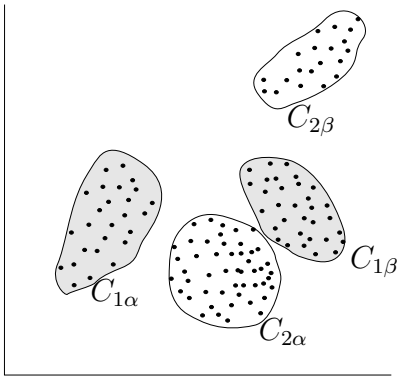
όπου $\boldsymbol{w}_{ij} = \boldsymbol{w}_i - \boldsymbol{w}_j$. Εύκολα φαίνεται ότι αν $d_i(\boldsymbol{x}) > d_j(\boldsymbol{x})$ για κάθε $j \neq i$, τότε $d_{ij}(\boldsymbol{x}) > 0$ για κάθε $j \neq i$, άρα οι κατηγορίες είναι διαχωρίσιμες υπό τις προϋποθέσεις της Μεθοδολογίας 2. Το αντίθετο γενικά δεν ισχύει.



Σχήμα 3.4: *Γραμμικές συναρτήσεις απόφασης με την τρίτη μεθοδολογία.*

Πρέπει να σημειωθεί ότι με την τρίτη μέθοδο, όπως φαίνεται και στο Σχήμα 3.4, δεν υπάρχουν απροσδιόριστες περιοχές, εκτός βέβαια από τα σημεία που βρίσκονται πάνω στις συναρτήσεις αποφάσεων.

Εάν οι κατηγορίες προτύπων σε οποιαδήποτε περίπτωση μπορούν να ταξινομηθούν με γραμμικές συναρτήσεις αποφάσεων χρησιμοποιώντας μια από τις παραπάνω μεθόδους γραμμικού διαχωρισμού, τότε οι κατηγορίες ονομάζονται *γραμμικά διαχωρίσιμες*. Το βασικό όμως πρόβλημα παραμένει, δηλαδή μετά τον καθορισμό των συναρτήσεων απόφασης πρέπει να υπολογιστούν οι συντελεστές τους. Συνήθως αυτοί οι συντελεστές υπολογίζονται χρησιμοποιώντας διαθέσιμα δείγματα προτύπων. Όταν οι συντελεστές κάθε συνάρτησης απόφασης έχουν υπολογιστεί, τότε αυτές οι συναρτήσεις μπορούν να χρησιμοποιηθούν σαν βάση για την ταξινόμηση προτύπων.

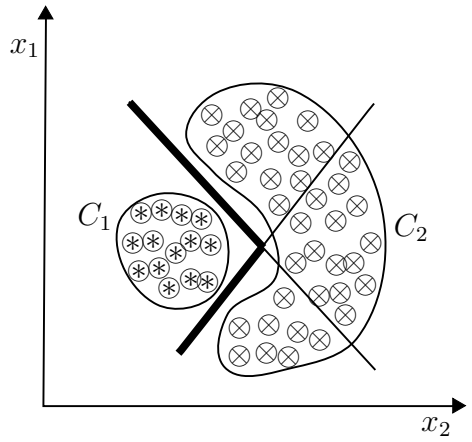


Σχήμα 3.5: *Το πρόβλημα του Sebestyen.*

3.3. Γενικευμένες Συναρτήσεις Απόφασης

Συναρτήσεις απόφασης μπορούν πάντοτε να καθοριστούν μεταξύ κατηγοριών προτύπων οι οποίες είναι διαχωρίσιμες. Στην πιο απλή περίπτωση οι συναρτήσεις απόφασης είναι γραμμικές αλλά ανάλογα με τη γεωμετρία και την τοπολογία των κατηγοριών μπορεί να γίνουν πολύπλοκες. Το Σχήμα 3.5 παρουσιάζει ένα απλό φανταστικό παράδειγμα, το οποίο παρουσίασε ο Sebestyen [Sebes62] και αποτελείται από δύο κατηγορίες, στην κάθε μια από τις οποίες ανήκουν δύο υποκατηγορίες. Εύκολα διαπιστώνεται ότι δεν υπάρχει γραμμική συνάρτηση απόφασης η οποία να μπορεί να διαχωρίσει αυτές τις κατηγορίες. Το Σχήμα 3.6 παρουσιάζει μια άλλη περίπτωση δύο κατηγοριών που δεν είναι γραμμικά διαχωρίσιμες. Σε αυτές τις περιπτώσεις υπάρχει η έννοια της δημιουργίας μη γραμμικών συναρτήσεων απόφασης. Ένας απλός τρόπος για την γενίκευση των συναρτήσεων απόφασης γραμμικού τύπου είναι ο ορισμός συναρτήσεων απόφασης της μορφής:

$$\begin{aligned} d(\boldsymbol{x}) &= w_1 f_1(\boldsymbol{x}) + w_2 f_2(\boldsymbol{x}) + \ldots + w_k f_k(\boldsymbol{x}) + w_{k+1} \\ &= \sum_{i=1}^{k+1} w_i f_i(\boldsymbol{x}). \end{aligned} \tag{3.12}$$



Σχήμα 3.6: *Μη γραμμικά διαχωρίσιμες κατηγορίες.*

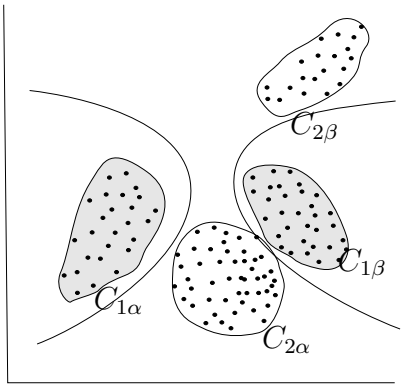
όπου $\{f_i(\boldsymbol{x})\}$, $i = 1, 2, \ldots, k$ είναι πραγματικές συναρτήσεις του διανύσματος προτύπου $\boldsymbol{x}$ και $f_{k+1}(\boldsymbol{x}) = 1$. Οι συναρτήσεις $\{f_i(\boldsymbol{x})\}$ ονομάζονται *συναρτήσεις βάσης* (*basis functions*). Η Εξίσωση (3.12) αντιπροσωπεύει μια απεριόριστη ποικιλία συναρτήσεων απόφασης ανάλογα με τις επιλογές της μορφής των συναρτήσεων $f_i(\boldsymbol{x})$.

Μια από τις πιο συχνά χρησιμοποιούμενες γενικευμένες συναρτήσεις απόφασης είναι η πολυωνυμική μορφή της Εξίσωσης (3.12). Στην απλούστερη τους μορφή οι αυτές οι συναρτήσεις είναι γραμμικές, δηλαδή $f_i(\boldsymbol{x}) = x_i$ και $k = n$. Σε αυτή την περίπτωση έχουμε:

$$d(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + w_{n+1}. \tag{3.13}$$

Στο επόμενο επίπεδο πολυπλοκότητας έχουμε την γενική δευτερο-βάθμια συνάρτηση απόφασης. Για την απλή δισδιάστατη περίπτωση $\boldsymbol{x} = [x_1, x_2]^T$ και η συνάρτηση απόφασης παίρνει την μορφή:

$$d(\boldsymbol{x}) = w_{11}x_1^2 + w_{12}x_1x_2 + w_{22}x_2^2 + w_1x_1 + w_2x_2 + w_3. \tag{3.14}$$



Σχήμα 3.7: *Υπερβολική παραβολή στο πρόβλημα του Sebestyen.*

Η Εξίσωση (3.14) μπορεί εύκολα να εκφραστεί σε γραμμική μορφή $d(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x}$, αν οριστούν τα $\boldsymbol{w}$ και $\boldsymbol{x}$ ως:

$$\begin{aligned} \boldsymbol{x} &= \left[x_1^2, x_1x_2, x_2^2, x_1, x_2, 1\right]^T, \\ \boldsymbol{w} &= \left[w_{11}, w_{12}, w_{22}, w_1, w_2, w_3\right]^T. \end{aligned} \tag{3.15}$$

Αυτός ο τρόπος προσέγγισης δυστυχώς δεν αλλάζει τίποτα στην συνδυαστική φύση του προβλήματος. Στη γενική περίπτωση του $n-$διάστατου χώρου ο απαιτούμενος αριθμός των συντελεστών w που πρέπει να υπολογιστούν αυξάνει πολύ ραγδαία σε σχέση με το n, καθιστώντας δύσκολο τον υπολογισμό των συναρτήσεων απόφασης σε χώρους υψηλής διαστασιμότητας. Το φαινόμενο την εκθετικής αύξησης των παραμέτρων, αλλά και των απαιτούμενων παραδειγμάτων εκπαίδευσης για τον υπολογισμό τους, με την αύξηση της διαστασιμότητας του χώρου προτύπων ονομάστηκε *κατάρα της διαστασιμότητας* (*curse of dimensionality*) [Bellm56].

Εάν παρατηρήσουμε πιο προσεκτικά το πρόβλημα του Sebestyen διαπιστώνουμε ότι μπορεί να λυθεί πολύ εύκολα με μια γνωστή δευτεροβάθμια συνάρτηση απόφασης την υπερβολική παραβολή (hyperbole parabola) όπως φαίνεται στο Σχήμα 3.7 και ορίζεται από τη σχέση:

$$\frac{(u+\alpha)^2}{c} - \frac{(v+b)^2}{d} = k, \tag{3.16}$$

όπου $u = m\,x_1 + n\,x_2, \quad v = p\,x_1 + q\,x_2$ και $\alpha,\ b,\ c,\ d,\ k,\ m$,$n,\ p,\ q$ σταθερές.

Αλλά, ακόμα και στο σχεδιασμό δευτεροβάθμιων συναρτήσεων απόφασης, ο υπολογισμός ενός τόσου μεγάλου αριθμού παραμέτρων είναι προβληματικός. Όπως σημειώνουν οι Duda and Hart [DuHa73] στο κλασικό βιβλίο τους:

“

$$\begin{array}{c} n \\ (n+1)(n+2)/2 \end{array} \qquad\qquad n$$

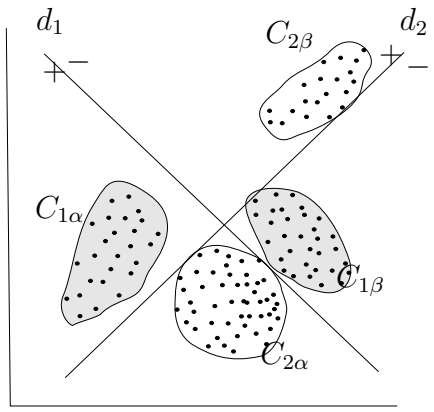
$$n \;=\; 50$$

...”

3.4. Τμηματικός Γραμμικός Διαχωρισμός.

Το Σχήμα 3.8 παρουσιάζει μια αποτελεσματική μέθοδο διαχωρισμού των κατηγοριών του προβλήματος του Sebestyen με δύο γραμμικές συναρτήσεις απόφασης. Οι δύο γραμμικές συναρτήσεις διαχωρίζουν μεταξύ τους όλες τις ξεχωριστές υποκατηγορίες. Εάν κάθε γραμμική συνάρτηση απόφασης δίνει μια δυαδική απόφαση “+” ή “−”, όπως φαίνεται και στο Σχήμα 3.8, η αναπαράσταση τους μπορεί να πραγματοποιηθεί με λογικές μεταβλητές όπου “+” το 1 και “−” το 0. Τότε οι κατηγορίες μπορούν να διαχωριστούν με τις παρακάτω λογικές εξισώσεις:

$$\begin{aligned} C_1 &= C_{1\alpha} + C_{1\beta} = d_1 \cdot d_2 \;+\; \bar{d}_1 \cdot \bar{d}_2, \\ C_2 &= C_{2\alpha} + C_{2\beta} = d_1 \cdot \bar{d}_2 \;+\; \bar{d}_1 \cdot d_2. \end{aligned} \tag{3.17}$$



Σχήμα 3.8: *Τμηματικός γραμμικός διαχωρισμός στο πρόβλημα του Sebestyen*

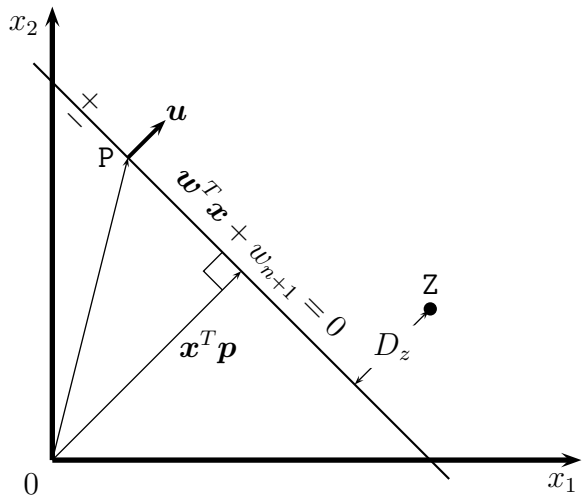
όπου $+$ η λογική πράξη OR, $\cdot$ η λογική πράξη AND και $\overline{\square}$ η λογική πράξη NOT. Ο παραπάνω τρόπος διαχωρισμού ονομάζεται *τμηματικός γραμμικός διαχωρισμός* (*piecewise - linear discrimination*).

Η περίπτωση του Σχήματος 3.6, παρότι και σε αυτό μπορεί να εφαρμοστεί η μέθοδος του τμηματικού γραμμικού διαχωρισμού, έχει τελείως διαφορετικά χαρακτηριστικά από το πρόβλημα του Sebestyen. Οι συναρτήσεις απόφασης περνούν μέσα από τις κατηγορίες και τις διχοτομούν.

Από τα παραπάνω παραδείγματα φαίνεται ξεκάθαρα πόσο αναγκαία είναι η γνώση της τοπολογίας του χώρου που καταλαμβάνουν οι κατηγορίες.

3.5. Γεωμετρικές ιδιότητες υπερεπιπέδων.

Στην ενότητα αυτή θα εξεταστούν σημαντικές ιδιότητες των γραμμικών επιφανειών απόφασης. Μια γενική συνάρτηση απόφασης ορίζεται



Σχήμα 3.9: *Γεωμετρικές ιδιότητες ενός υπερεπιπέδου.*

από την σχέση:

$$\begin{aligned} d(\boldsymbol{x}) &= w_1 x_1 + w_2 x_2 + \cdots + w_n x_n + w_{n+1} \\ &= \boldsymbol{w}^T \boldsymbol{x} + w_{n+1} = 0, \end{aligned} \tag{3.18}$$

όπου $\boldsymbol{w} = [w_1, w_2, \ldots, w_n]^T$. Να σημειωθεί ότι στην Εξίσωση (3.18) δεν χρησιμοποιείται η επαυξημένη μορφή του $\boldsymbol{x}$ μιας και η τιμή του w_{n+1} παίζει όπως θα δειχθεί σημαντικό ρόλο. Το Σχήμα 3.9 δείχνει μια σχηματική αναπαράσταση ενός υπερεπιπέδου.

Έστω $\boldsymbol{u}$ μοναδιαίο διάνυσμα κάθετο στο υπερεπίπεδο σε ένα τυχαίο σημείο P και δείχνει προς την θετική πλευρά του υπερεπιπέδου. Για ένα τυχαίο σημείο του υπερεπιπέδου X το άνυσμα $(\boldsymbol{x} - \boldsymbol{p})$ θα είναι πάνω στο επίπεδο, άρα και κάθετο στο $\boldsymbol{u}$, όπου $\boldsymbol{x}$ και $\boldsymbol{p}$ τα αντίστοιχα διανύσματα των σημείων. Άρα για κάθε σημείο του υπερεπιπέδου ισχύει η σχέση:

$$\boldsymbol{u}^T (\boldsymbol{x} - \boldsymbol{p}) = 0, \tag{3.19}$$

ή

$$\boldsymbol{u}^T \boldsymbol{x} = \boldsymbol{u}^T \boldsymbol{p}. \tag{3.20}$$

Διαιρώντας και τα δύο μέλη της Εξίσωσης (3.18) με το μέτρο του $\boldsymbol{w}$ που δίδεται από την $\|\boldsymbol{w}\| = \sqrt{w_1^2 + w_2^2 + \cdots + w_n^2}$ προκύπτει η εξίσωση:

$$\frac{\boldsymbol{w}^T \boldsymbol{x}}{\|\boldsymbol{w}\|} = -\frac{w_{n+1}}{\|\boldsymbol{w}\|}. \tag{3.21}$$

Συγκρίνοντας τις Εξισώσεις (3.20) και (3.21) προκύπτει ότι το κάθετο μοναδιαίο άνυσμα δίδεται από την εξίσωση:

$$\boldsymbol{u} = \frac{\boldsymbol{w}^T}{\|\boldsymbol{w}\|} \tag{3.22}$$

και

$$\boldsymbol{u}^T \boldsymbol{p} = -\frac{w_{n+1}}{\|\boldsymbol{w}\|}. \tag{3.23}$$

Από το Σχήμα 3.9 φαίνεται ότι η απόλυτη τιμή του $\boldsymbol{u}^T\boldsymbol{p}$, δηλαδή η προβολή του διανύσματος $\boldsymbol{p}$ πάνω στον άξονα που ορίζει το $\boldsymbol{u}$, είναι η απόσταση του υπερεπιπέδου από την αρχή των αξόνων. Αντίστοιχα η απόσταση D_z του σημείου Z δίδεται από τη σχέση:

$$\begin{aligned} D_z &= |\boldsymbol{u}^T \boldsymbol{z} - \boldsymbol{u}^T \boldsymbol{p}| \\ &= \left| \frac{\boldsymbol{w}^T \boldsymbol{z}}{\|\boldsymbol{w}\|} + \frac{w_{n+1}}{\|\boldsymbol{w}\|} \right| \\ &= \left| \frac{\boldsymbol{w}^T \boldsymbol{z} + w_{n+1}}{\|\boldsymbol{w}\|} \right|. \end{aligned} \tag{3.24}$$

Το μοναδιαίο άνυσμα $\boldsymbol{u}$ προσδιορίζει μοναδικά την διεύθυνση του υπερεπιπέδου. Αν κάποιο στοιχείο του είναι μηδέν, τότε είναι παράλληλο στον αντίστοιχο άξονα. Τέλος το w_{n+1} προσδιορίζει την απόσταση του υπερεπιπέδου από την αρχή των αξόνων. Αν $w_{n+1} = 0$ τότε το υπερεπίπεδο περνάει από την αρχή των αξόνων.

3.1 Μπορούν τα χαρακτηριστικά διανύσματα της κατηγορίας C_1 με τιμές $(2,3)$,$(3,5)$,$(4,2)$,$(2,7)$ να διαχωριστούν με μια γραμμική συνάρτηση απόφασης από τα ανύσματα $(6,2)$,$(5,4)$,$(5,6)$,$(3,7)$ της κατηγορίας C_2; Εάν ναι, σχεδιάστε μια τέτοια συνάρτηση απόφασης και υπολογίστε τις παραμέτρους της συνάρτησης. Διαφορετικά, προσπαθήστε να τα διαχωρίσετε γραφικά όσο καλύτερα μπορείτε.

3.2 Να επαναληφθεί η Άσκηση 3.1 χρησιμοποιώντας τα χαρακτηριστικά διανύσματα $(1,1)$,$(1,2)$, $(2,2)$,$(2,3)$,$(3,0)$ για τη κατηγορία C_1 και τα $(3,1)$,$(4,1)$, $(5,2)$ για τη κατηγορία C_2.

3.3 Σε ένα πρόβλημα διαχωρισμού 10 κατηγοριών, εάν 3 από αυτές ικανοποιούν τις προϋποθέσεις της πρώτης μεθοδολογίας και οι υπόλοιπες ικανοποιούν τις προϋποθέσεις της δεύτερης μεθοδολογίας, ποίο είναι το ελάχιστο πλήθος γραμμικών συναρτήσεων απόφασης που απαιτούνται για να λυθεί το πρόβλημα;

3.4 Σε ένα πρόβλημα διαχωρισμού 3 κατηγοριών έχουν βρεθεί οι παρακάτω γραμμικές συνάρτησης απόφασης:

$$d_1(\boldsymbol{x}) = -x_1, \;\; d_2(\boldsymbol{x}) = x_1 + x_2 + 1, \;\; d_3(\boldsymbol{x}) = x_1 - x_2 - 1,$$

οι οποίες ικανοποιούν τις προϋπόθεσης της πρώτης μεθοδολογίας. Σχεδιάστε τις συναρτήσεις απόφασης, τις περιοχές κάθε κατηγορίας και τις απροσδιόριστες περιοχές, εάν υπάρχουν.

3.5 Σε ένα πρόβλημα διαχωρισμού 3 κατηγοριών έχουν βρεθεί οι παρακάτω γραμμικές συνάρτησης απόφασης:

$$d_{12}(\boldsymbol{x}) = -x_1, \;\; d_{13}(\boldsymbol{x}) = x_1 + x_2 + 1, \;\; d_{23}(\boldsymbol{x}) = x_1 - x_2 - 1,$$

οι οποίες ικανοποιούν τις προϋπόθεσης της δεύτερης μεθοδολογίας. Σχεδιάστε τις συναρτήσεις απόφασης, τις περιοχές κάθε κατηγορίας και τις απροσδιόριστες περιοχές, εάν υπάρχουν.

3.6 Σε ένα πρόβλημα διαχωρισμού 3 κατηγοριών έχουν βρεθεί οι παρακάτω γραμμικές συνάρτησης απόφασης:

$$d_1(\boldsymbol{x}) = -x_1, \;\; d_2(\boldsymbol{x}) = x_1 + x_2 + 1, \;\; d_3(\boldsymbol{x}) = x_1 - x_2 - 1,$$

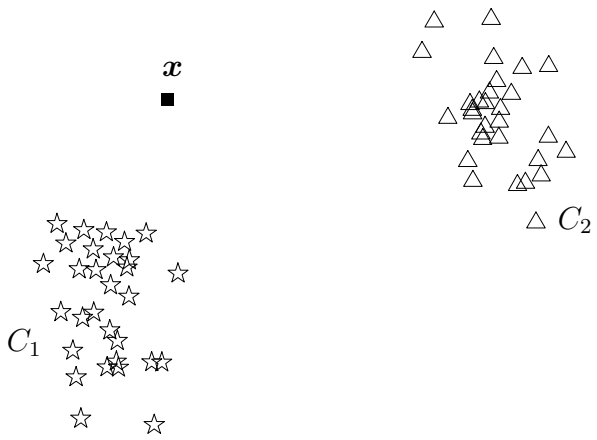
οι οποίες ικανοποιούν τις προϋπόθεσης της τρίτης μεθοδολογίας. Σχεδιάστε τις συναρτήσεις απόφασης και τις περιοχές κάθε κατηγορίας.

Κεφάλαιο 4

Ταξινομηση προτυπων με συναρτησεις αποστασης

4.1. Απλοί ταξινομητές

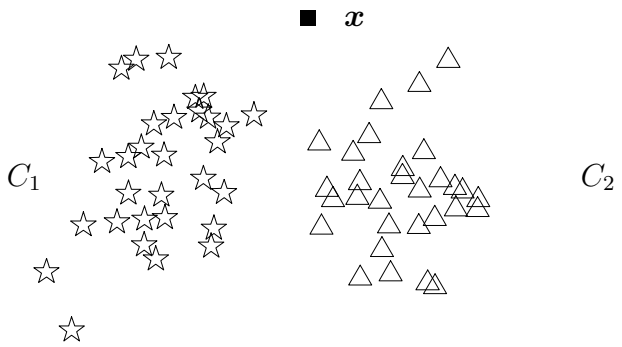
Υπάρχουν τουλάχιστον δύο προσεγγίσεις για το σχεδιασμό ενός ταξινομητή. Η πρώτη προσέγγιση είναι η *θεωρητική*. Αρχικά, δημιουργείται ένα μαθηματικό μοντέλο του προβλήματος και στην συνέχεια βάση του μοντέλου σχεδιάζεται ένας βέλτιστος ταξινομητής. Η δεύτερη προσέγγιση είναι η *πρακτική εφαρμογή*. Αρχικά πραγματοποιείται η υπόθεση



Σχήμα 4.1: *Κατάταξη δια μέσου της εγγύτητας.*

μιας πιθανής λύσης, σύμφωνα με τα δείγματα των κατηγοριών του προβλήματος και κατόπιν πραγματοποιείται η βέλτιστη αναπροσαρμογή της λύσης σύμφωνα με τα πραγματικά δεδομένα του προβλήματος. Η δεύτερη προσέγγιση είναι εμπειρική μέθοδος, χρησιμοποιείται ευρέως σε πρακτικές εφαρμογές αναγνώρισης προτύπων και είναι η προσέγγιση που θα ακολουθήσουμε. Θα ξεκινήσουμε με μια απλή λύση του προβλήματος, θα αναλύσουμε τα χαρακτηριστικά της, θα εντοπίσουμε τις αδυναμίες της και θα την κάνουμε σταδιακά όσο πολύπλοκη χρειαστεί.

Η πιο απλή και εμπειρική προσέγγιση στο πρόβλημα της αναγνώρισης προτύπων είναι η ιδέα της ταξινόμησης με τη χρήση συναρτήσεων απόστασης. Η χρησιμοποίηση συναρτήσεων απόστασης είναι φυσικό επακόλουθο του γεγονότος ότι ο πιο εμφανής τρόπος καθορισμού ενός μέτρου ομοιότητας μεταξύ των προτύπων, τα οποία μπορούν να θεωρηθούν σαν σημεία στον Ευκλείδειο χώρο, είναι ο υπολογισμός της απόστασης μεταξύ τους. Η έννοια της ταξινόμησης με χρήση συναρτήσεων απόστασης προκύπτει διαισθητικά από το Σχήμα 4.1. Τείνει να κατατάξει κάποιος το πρότυπο $\boldsymbol{x}$ στην κατηγορία C_1 γιατί βρίσκεται εγγύτερα της στο χώρο προτύπων.



Σχήμα 4.2: *Κατηγορίες μη-κατάταξιμες δια μέσου της έννοιας της εγγύτητας.*

Για να είναι αποτελεσματική η ταξινόμηση με χρήση συναρτήσεων απόστασης θα πρέπει οι κατηγορίες του προβλήματος να σχηματίζουν ευδιάκριτες ομάδες στο χώρο προτύπων. Το Σχήμα 4.2 δείχνει μια περίπτωση που αυτό δεν ισχύει και η κατάταξη με χρήση της έννοιας της εγγύτητας δεν είναι ασφαλής.

Οι παραπάνω έννοιες θα γενικευθούν και θα αναπτυχθούν σε ένα συμπαγές μαθηματικό πλαίσιο στις ενότητες που ακολουθούν. Η σημαντική έννοια των ομάδων προτύπων θα αναπτυχθεί αναλυτικά στο επόμενο Κεφάλαιο. Εφόσον η απόσταση μεταξύ ενός άγνωστου προτύπου και των προτύπων μιας κατηγορίας θα χρησιμοποιηθεί σαν μέτρο ταξινόμησης, ο όρος *ταξινομητές ελάχιστης απόστασης* (*minimal distance classifiers*) θα χρησιμοποιηθεί για τον χαρακτηρισμό αυτής της προσέγγισης.

4.2. Ταξινομητές Ελάχιστης Απόστασης

Η αναγνώριση προτύπων χρησιμοποιώντας συναρτήσεις απόστασης είναι μια από τις πρωτοπόρες τεχνικές που αναπτύχθηκαν στον τομέα της αναγνώρισης προτύπων. Είναι βασικά μια απλή τεχνική ταξινόμησης,

η οποία όμως είναι αποτελεσματική σε προβλήματα που οι κατηγορίες τους είναι σχετικά "καλά" διαχωρίσιμες. Αρχικά θα εξεταστεί η περίπτωση που κάθε κατηγορία μπορεί να αντιπροσωπευθεί από μόνο ένα *χαρακτηριστικό πρότυπο*.

Έστω ότι κάθε κατηγορία C_i, $i = 1, \ldots, M$ εκπροσωπείται ικανοποιητικά με ένα μόνο αντιπροσωπευτικό διάνυσμα $\boldsymbol{z}_i$. Για παράδειγμα, το αντιπροσωπευτικό διάνυσμα $\boldsymbol{z}_i$ θα μπορούσε να είναι το μέσο διάνυσμα όλων των προτύπων της κατηγορίας C_i.

Η έννοια της Ευκλείδειας απόστασης μεταξύ ενός τυχαίου προτύπου $\boldsymbol{x}$ και ενός αντιπροσωπευτικού προτύπου $\boldsymbol{z}_i$ ορίζεται από τη σχέση:

$$D_i(\boldsymbol{x}) = \|\boldsymbol{x} - \boldsymbol{z}_i\| = \sqrt{(\boldsymbol{x} - \boldsymbol{z}_i)^T(\boldsymbol{x} - \boldsymbol{z}_i)}\,, \tag{4.1}$$

όπου το σύμβολο $\|\boldsymbol{u}\|$ αποκαλείται μετρική ή νόρμα του διανύσματος $\boldsymbol{u}$ και είναι μια γενικευμένη έννοια του μήκους του.

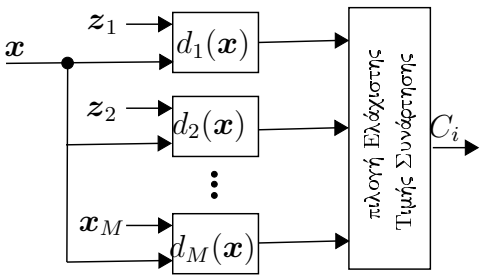
Ένας ταξινομητής ελάχιστης απόστασης υπολογίζει την απόσταση ενός προτύπου $\boldsymbol{x}$, άγνωστης ταξινόμησης, με το αντιπροσωπευτικό πρότυπο $\boldsymbol{z}_i$ κάθε κατηγορίας και καταχωρεί το πρότυπο $\boldsymbol{x}$ στην κατηγορία που έχει την ελάχιστη απόσταση από το $\boldsymbol{x}$. Αν εκφραστεί μαθηματικά σημαίνει ότι το $\boldsymbol{x}$ κατατάσσεται στην κατηγορία C_i εάν ισχύει:

$$D_i(\boldsymbol{x}) < D_j(\boldsymbol{x}) \quad \forall \quad j \neq i. \tag{4.2}$$

Η Εξίσωση (4.1) μπορεί να εκφραστεί σε μια πιο χρήσιμη μορφή, εάν υψωθούν στο τετράγωνο και οι δύο πλευρές της:

$$\begin{aligned} D_i^2(\boldsymbol{x}) = \|\boldsymbol{x} - \boldsymbol{z}_i\|^2 &= (\boldsymbol{x} - \boldsymbol{z}_i)^T(\boldsymbol{x} - \boldsymbol{z}_i) \\ &= \boldsymbol{x}^T\boldsymbol{x} - 2\,\boldsymbol{x}^T\boldsymbol{z}_i + \boldsymbol{z}_i^T\boldsymbol{z} \\ &= \boldsymbol{x}^T\boldsymbol{x} - 2\,\left(\boldsymbol{x}^T\boldsymbol{z}_i - 0.5\,\boldsymbol{z}_i^T\boldsymbol{z}\right). \end{aligned} \tag{4.3}$$

Η επιλογή της ελάχιστης $D_i^2(\boldsymbol{x})$ είναι ισοδύναμη με την επιλογή της ελάχιστης $D_i(\boldsymbol{x})$, δεδομένου ότι οι τιμές της απόστασης είναι εξ'ορισμού πάντα θετικές. Επιπλέον, εφόσον το πρώτο μέρος της Εξίσωσης (4.3) εξαρτάται μόνο από το $\boldsymbol{x}$, δηλαδή είναι το ίδιο για όλες τις αποστάσεις $D_i(\boldsymbol{x})$, η επιλογή της ελάχιστης $D_i(\boldsymbol{x})$ ανάγεται στο υπολογισμό του



Σχήμα 4.3: *Υλοποίηση ενός ταξινομητή ελάχιστης απόστασης.*

δεύτερου μέρους της Εξίσωσης (4.3): $\boldsymbol{x}^T\boldsymbol{z}_i - 0.5\,\boldsymbol{z}_i^T\boldsymbol{z}$. Επομένως, μπορούν να οριστούν οι παρακάτω M συναρτήσεις απόφασης:

$$\begin{aligned} d_i(\boldsymbol{x}) &= \boldsymbol{x}^T\boldsymbol{z}_i - 0.5\,\boldsymbol{z}_i^T\boldsymbol{z}_i \\ &= \boldsymbol{x}^T\boldsymbol{z}_i - 0.5\,\|\boldsymbol{z}_i\|^2, \quad i = 1, 2, \ldots, M \end{aligned} \tag{4.4}$$

Ένα πρότυπο $\boldsymbol{x}$ κατατάσσεται στην κατηγορία C_i εάν $d_i(\boldsymbol{x}) > d_j(\boldsymbol{x})$ για κάθε $i \neq j$. Πρέπει να σημειωθεί ότι το $d_i(\boldsymbol{x})$ είναι μια γραμμική συνάρτηση απόφασης και μπορεί να εκφραστεί σε γραμμική μορφή ως:

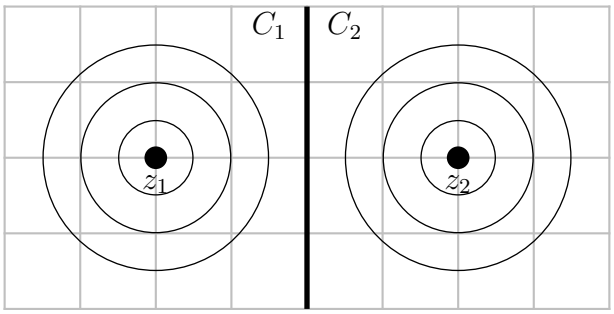
$$d_i(\boldsymbol{x}) = \boldsymbol{w}_i^T\boldsymbol{x}, \tag{4.5}$$

όπου:

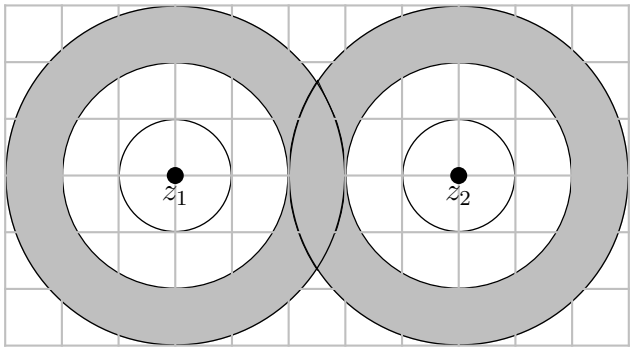
$$\boldsymbol{x} = [x_1, x_2, \ldots, x_n, 1]^T, \tag{4.6}$$

$$\boldsymbol{w} = [z_{i1}, z_{i2}, \ldots, z_{in}, -0.5\,\boldsymbol{z}_i^T\boldsymbol{z}_i]^T. \tag{4.7}$$

Η υλοποίηση ενός ταξινομητή ελάχιστης απόστασης παρουσιάζεται στο Σχήμα 4.3. Τα όρια απόφασης ενός προβλήματος με δύο κατηγορίες όπου κάθε κατηγορία χαρακτηρίζεται από ένα αντιπροσωπευτικό διάνυσμα $\boldsymbol{z}_i$, παρουσιάζονται στο Σχήμα 4.4. Επειδή σαν μέτρο απόστασης χρησιμοποιήθηκε η Ευκλείδεια μετρική οι ισομετρικές αποστάσεις, δηλαδή τα σημεία που απέχουν εξίσου από το αντιπροσωπευτικό



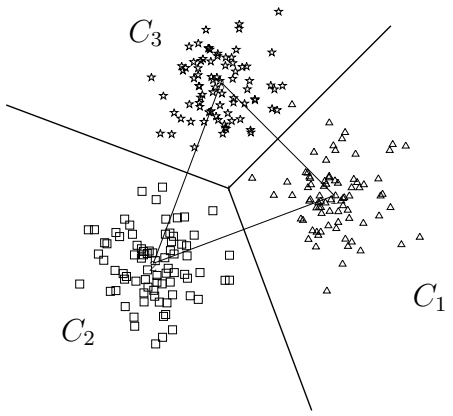
Σχήμα 4.4: *Όρια απόφασης και ισομετρικές για την Ευκλείδεια απόσταση.*



Σχήμα 4.5: *Όρια απόφασης για την Ευκλείδεια απόσταση με επικαλυπτόμενες κατηγορίες.*

διάνυσμα $\boldsymbol{z}_i$ για μια κατηγορία είναι κύκλοι. Εάν τα πρότυπα κάθε κατηγορίας βρίσκονται κοντά στο αντιπροσωπευτικό διάνυσμα τότε ο γραμμικός διαχωρισμός είναι πολύ απλός. Αντίθετα, στην περίπτωση που υπάρχει μεγάλη διασπορά των προτύπων γύρω από το αντιπροσωπευτικό διάνυσμα κάθε κατηγορίας, οι ισομετρικές αποστάσεις των κατηγοριών επικαλύπτονται όπως φαίνεται από τη σκιώδη περιοχή του Σχήματος 4.5.

Τα όρια απόφασης ενός προβλήματος με τρεις κατηγορίες παρουσιάζονται στο Σχήμα 4.6. Οι συναρτήσεις αποφάσεων είναι γραμμικές και είναι οι μεσοκάθετες των ευθειών που συνδέουν τα αντιπροσωπευτικά ανύσματα προτύπων των κατηγοριών.



Σχήμα 4.6: *Όρια απόφασης για την Ευκλείδεια απόσταση και τρεις κατηγορίες.*

4.3. Μέτρα απόστασης

Διαισθητικά μπορεί να εκτιμηθεί ότι αντικείμενα τα οποία βρίσκονται "κοντά" στο χώρο προτύπων είναι όμοια μεταξύ τους, ενώ αντικείμενα τα οποία βρίσκονται "μακριά" είναι ανόμοια. Για την μαθηματική έκφραση της απόστασης μεταξύ αντικειμένων στον χώρο προτύπων πρέπει να οριστεί το μέτρο ομοιότητας μεταξύ των προτύπων. Ο πιο προφανής τρόπος μέτρησης της απόστασης μεταξύ δύο σημείων είναι η γνωστή μέθοδος της Ευκλείδεια απόστασης. Η Ευκλείδεια απόσταση μεταξύ των προτύπων $\boldsymbol{x}$ και $\boldsymbol{z}$ διάστασης n, δίδεται από τη σχέση:

$$\begin{aligned} D_2(\boldsymbol{x},\boldsymbol{z}) = \|\boldsymbol{x}-\boldsymbol{z}\|_2 &= \sqrt{(\boldsymbol{x}-\boldsymbol{z})^T(\boldsymbol{x}-\boldsymbol{z})} \\ &= \left[\sum_{i=1}^{n}(x_i-z_i)^2\right]^{\frac{1}{2}}. \end{aligned} \tag{4.8}$$

Για παράδειγμα, εάν $\boldsymbol{x} = [1,2,2,0]^T$ και $\boldsymbol{z} = [2,1,2,2]^T$ τότε η Ευκλείδεια απόσταση τους είναι:

$$D_2(\boldsymbol{x},\boldsymbol{z}) = \sqrt{(1-2)^2+(2-1)^2+(2-2)^2+(0-2)^2} = \sqrt{6}.$$

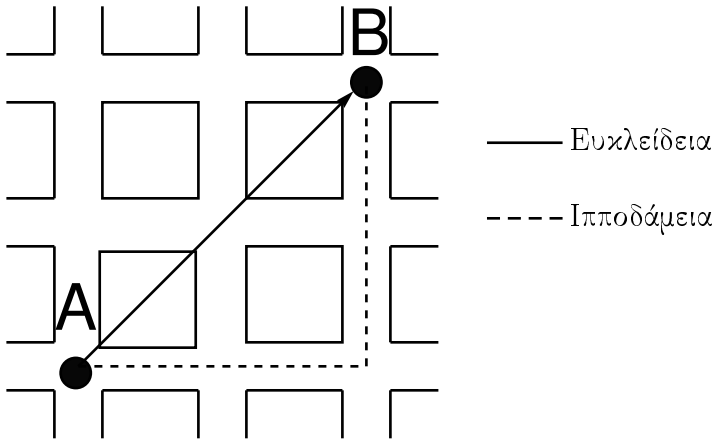
Στην Ευκλείδεια απόσταση οι ισομετρικές αποστάσεις, δηλαδή τα σημεία που απέχουν εξίσου από ένα σημείο, είναι κύκλοι. Το Σχήμα 4.4 παρουσιάζει ορισμένες ισομετρικές αποστάσεις γύρω από τα δύο αντιπροσωπευτικά διανύσματα $\boldsymbol{z}_1$ και $\boldsymbol{z}_2$ των κατηγοριών C_1 και C_2 αντίστοιχα, καθώς και ένα απλοϊκό τρόπο δημιουργίας ομάδων.

Η Ευκλείδεια απόσταση είναι πιθανά το πιο ευρέως χρησιμοποιούμενο μέτρο απόστασης, για μετρήσεις ανομοιότητας μεταξύ χαρακτηριστικών διανυσμάτων, αλλά δεν είναι πάντα κατάλληλη για όλες τις εφαρμογές. Είναι χρήσιμη όταν το πρόβλημα εμπεριέχει πρότυπα με συνεχείς τιμές. Όμως, σε μια ιατρική εφαρμογή όπου το ένα χαρακτηριστικό γνώρισμα αντιπροσωπεύει το επίπεδο ζαχάρου στο αίμα και το άλλο το βάρος του ασθενούς, δεν φαίνεται λογικό να σχηματιστεί διάνυσμα και να χρησιμοποιηθεί η Ευκλείδεια απόσταση για τον διαχωρισμό κατηγοριών ασθενών.

Πολλά προβλήματα περιλαμβάνουν ακέραιες τιμές, δηλαδή τα δεδομένα είναι της μορφής: Μήνας $(1 \ldots 12)$, πλήθος οπών, κλπ. Το γεγονός ότι οι διαφορές των χαρακτηριστικών διανυσμάτων υψώνονται στο τετράγωνο έχει σαν αποτέλεσμα να υπερτονίζονται αυτές, ενώ απαιτούνται και επιπλέον υπολογισμοί. Μια απλή προσέγγιση είναι να αθροιστούν οι απόλυτες διαφορές μεταξύ των στοιχείων των χαρακτηριστικών διανυσμάτων. Το μέτρο απόστασης που προκύπτει ονομάζεται *Ιπποδάμεια μετρική* ή μετρική πρώτης τάξης και ορίζεται από τη σχέση:

$$D_1(\boldsymbol{x}, \boldsymbol{z}) = \|\boldsymbol{x} - \boldsymbol{z}\|_1 = \sum_{i=1}^{n} |x_i - z_i|. \tag{4.9}$$

Το μέτρο απόστασης πήρε το όνομα του από τον πολεοδόμο Ιππόδαμο τον Μιλήσιο ($5^{ος}$ αιώνας π.χ.) ο οποίος εισήγαγε το τετραγωνικό σύστημα δόμησης. Στη διεθνή βιβλιογραφία είναι γνωστό με τα ονόματα Manhatan distance ή City block distance ή Taxi distance. Η Ευκλείδεια απόσταση μεταξύ δύο αντίθετων γωνιών ενός οικοδομικού τετράγωνου με πλευρά μήκους 1 είναι $\sqrt{2}$, ενώ η πραγματική απόσταση που πρέπει να



Σχήμα 4.7: *Σύγκριση Ευκλείδειας και Ιπποδάμειας απόστασης.*

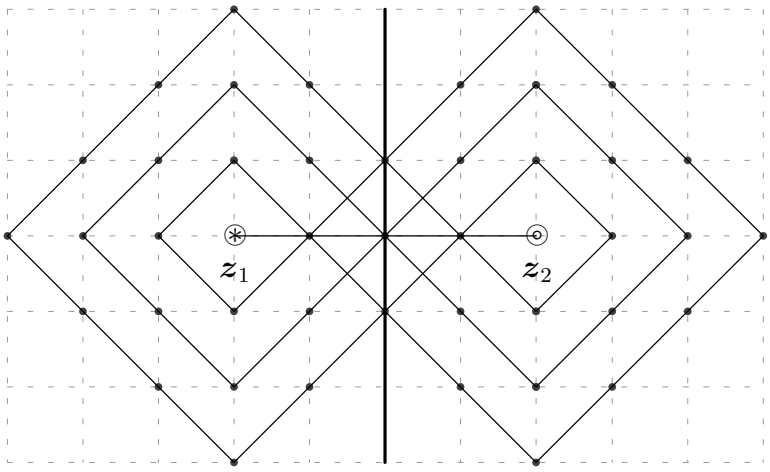
διανύσει κάποιος είναι η Ιπποδάμεια απόσταση ίση με 2, όπως φαίνεται στο Σχήμα 4.7.

Η Ιπποδάμεια απόσταση μεταξύ των διανυσμάτων $\boldsymbol{x} = [1, 2, 2, 0]^T$ και $\boldsymbol{z} = [2, 1, 2, 2]^T$ είναι:

$$D_1(\boldsymbol{x}, \boldsymbol{z}) = |1 - 2| + |2 - 1| + |2 - 2| + |0 - 2| = 4.$$

Στην Ιπποδάμεια απόσταση οι ισομετρικές αποστάσεις γύρω από ένα σημείο είναι τετράγωνα. Το Σχήμα 4.8 παρουσιάζει ορισμένες ισομετρικές αποστάσεις γύρω από τα δύο αντιπροσωπευτικά διανύσματα $\boldsymbol{z}_1$ και $\boldsymbol{z}_2$ αντίστοιχα, καθώς και την συνάρτηση απόφασης που προκύπτει. Τα διανύσματα της Ιπποδάμειας μετρικής περιλαμβάνουν μόνο ακέραιες τιμές και η αναπαράσταση τους στον χώρο γίνεται μόνο σε συγκεκριμένα σημεία. Στο Σχήμα 4.8 οι τελείες δείχνουν τις πιθανές θέσεις των ακέραιων διανυσμάτων στον δισδιάστατο χώρο.

Σε ορισμένες περιπτώσεις δεν έχουμε αριθμητικά δεδομένα αλλά χαρακτηριστικά του αντικειμένου τα οποία περιγράφονται ποιοτικά. Πολλά από τα ποιοτικά δεδομένα μπορούν να εκφραστούν σε δυαδική μορφή όπως "ΝΑΙ-ΟΧΙ", "ΠΑΡΟΝ-ΑΠΟΝ". Το διάνυσμα χαρακτηριστικών



Σχήμα 4.8: *Ισομετρικές στην Ιπποδάμεια απόσταση.*

που προκύπτει παίρνει τιμές 1 ή 0. Το πιο γνωστό δυαδικό μέτρο απόστασης είναι η *απόσταση Hamming*, η οποία μετράει των αριθμό των θέσεων του δυαδικού διανύσματος όπου τα πρότυπα διαφέρουν.

Για το μαθηματικό ορισμό της απόστασης Hamming πρέπει να εισάγουμε τον γνωστό δυαδικό τελεστή Exclusice OR, του οποίου το σύμβολο είναι $\oplus$ και ορίζεται ως:

$$0 \oplus 0 = 0,\ 1 \oplus 1 = 0,\ 0 \oplus 1 = 1,\ 1 \oplus 0 = 1.$$

Η απόσταση Hamming μεταξύ δύο διανυσμάτων ορίζεται ώς:

$$D_H(\boldsymbol{x}, \boldsymbol{z}) = \|\boldsymbol{x} - \boldsymbol{z}\|_1 = \sum_{i=1}^{n} x_i \oplus z_i \tag{4.10}$$

Η απόσταση Hamming είναι μια υποπερίπτωση της Ιπποδάμειας απόστασης και ταυτίζονται όταν τα διανύσματα είναι δυαδικά.

Η απόσταση Hamming μεταξύ των διανυσμάτων $\boldsymbol{x} = [1, 0, 0, 1, 1]^T$ και $\boldsymbol{z} = [1, 1, 0, 1, 0]^T$ είναι:

$$D_H(\boldsymbol{x}, \boldsymbol{z}) = |1 - 1| + |0 - 1| + |0 - 0| + |1 - 1| + |1 - 0| = 2.$$

Όλες οι μετρικές που εξετάστηκαν μέχρι τώρα είναι ειδικές περιπτώσεις της απόστασης του Minkowsky η οποία ορίζεται ως:

$$D_s(\boldsymbol{x}, \boldsymbol{z}) = \|\boldsymbol{x} - \boldsymbol{z}\|_s = \left[\sum_{i=1}^{n} |x_i - z_i|^s\right]^{\frac{1}{s}} \tag{4.11}$$

Με $s = 2$ σχηματίζεται η Ευκλείδεια μετρική και με $s = 1$ η Ιπποδάμεια. Στην περίπτωση που $s = \infty$ έχουμε την ειδική περίπτωση της απόστασης του Chebysher η οποία ορίζεται ως:

$$D_\infty(\boldsymbol{x}, \boldsymbol{z}) = \max_i \{|x_i - z_i|\}, \tag{4.12}$$

δηλαδή είναι η μέγιστη απόσταση μεταξύ των αντιστοίχων στοιχείων των δυο διανυσμάτων.

Η Chebysher απόσταση μεταξύ των διανυσμάτων $\boldsymbol{x} = [1, 2, 2, 0]^T$ και $\boldsymbol{z} = [2, 1, 2, 2]^T$ είναι ίση με:

$$\begin{aligned} D_\infty(\boldsymbol{x}, \boldsymbol{z}) &= \max\big\{|1-2|, |2-1|, |2-2|, |0-2|\big\} = \\ &= \max\big\{1, 1, 0, 2\big\} = 2. \end{aligned}$$

Ορισμένες από τις χρήσιμες ιδιότητες της απόστασης του Minkowsky είναι:

$$\|\boldsymbol{x}\|_s \geq 0, \tag{4.13}$$

$$\|\lambda\boldsymbol{x}\|_s = |\lambda| \, \|\boldsymbol{x}\|_s, \tag{4.14}$$

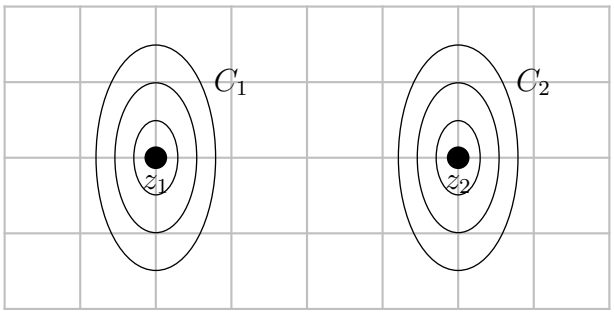
$$\|\boldsymbol{x} + \boldsymbol{y}\|_s \leq \|\boldsymbol{x}\|_s + \|\boldsymbol{y}\|_s, \tag{4.15}$$

όπου $\boldsymbol{x}$ και $\boldsymbol{y}$ διανύσματα και λ πραγματικός αριθμός.

Η Ιπποδάμεια μετρική και η μετρική του Chebysher μπορούν να θεωρηθούν σαν μια υπερεκτίμηση και υποεκτίμηση της Ευκλείδιας μετρικής αντίστοιχα. Μπορεί να αποδειχτεί ότι:

$$D_\infty(\boldsymbol{x}, \boldsymbol{z}) \leq D_2(\boldsymbol{x}, \boldsymbol{z}) \leq D_1(\boldsymbol{x}, \boldsymbol{z}). \tag{4.16}$$

Για παράδειγμα, εάν $\boldsymbol{x} = [1, 2, 3]^T$ και $\boldsymbol{z} = [0, 1, 3]^T$ τότε:



Σχήμα 4.9: *Ισομετρικές αποστάσεις στην απόσταση Mahalanobis.*

$$D_1(\boldsymbol{x}, \boldsymbol{z}) = 2,$$
$$D_2(\boldsymbol{x}, \boldsymbol{z}) = 1.41,$$
$$D_\infty(\boldsymbol{x}, \boldsymbol{z}) = 1,$$

δηλαδή:

$$1 \leq 1.41 \leq 2$$

και η ανισότητα 4.16 ισχύει.

Εκτός από τις παραπάνω αναφερόμενες μετρικές συναρτήσεις υπάρχουν και άλλα χρήσιμα μέτρα απόστασης. Ένα πολύ χρήσιμο μέτρο απόστασης το οποία λαμβάνει υπόψη του στατιστικούς δείκτες είναι η απόσταση του Mahalanobis:

$$D_M(\boldsymbol{x}, \boldsymbol{z}) = (\boldsymbol{x} - \boldsymbol{z})^T \boldsymbol{C}^{-1} (\boldsymbol{x} - \boldsymbol{z}), \tag{4.17}$$

όπου $\boldsymbol{C}$ είναι ο πίνακας συνδιακύμανσης της κατηγορίας και $\boldsymbol{z}$ το μέσο χαρακτηριστικό άνυσμα της κατηγορίας. Στην περίπτωση που ο πίνακας συνδιακύμανσης είναι ο μοναδιαίος πίνακας, $\boldsymbol{C} = \boldsymbol{I}$, η μετρική του Mahalanobis ταυτίζεται με την Ευκλείδια μετρική.

Το Σχήμα 4.9 παρουσιάζει ορισμένες ισομετρικές αποστάσεις της μετρικής Mahalanobis γύρω από δύο αντιπροσωπευτικά διανύσματα $\boldsymbol{z}_1$ και $\boldsymbol{z}_2$ των κατηγοριών C_1 και C_2 αντίστοιχα. Οι ισομετρικές αποστάσεις

Πίνακας 4.1: *Συνοπτικός πίνακας μέτρων απόστασης.*

Ιπποδάμεια	$D_1(\boldsymbol{x}, \boldsymbol{z}) = \sum_{i=1}^{n} \lvert x_i - z_i \rvert$	Ακέραια στοιχεία
Ευκλείδεια	$D_2(\boldsymbol{x}, \boldsymbol{z}) = [\sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{z}_i)^2]^{\frac{1}{2}}$	Συνεχή στοιχεία
Hamming	$D_H(\boldsymbol{x}, \boldsymbol{z}) = \sum_{i=1}^{n} x_i \oplus z_i$	Δυαδικά στοιχεία
Minkowsky	$D_s(\boldsymbol{x}, \boldsymbol{z}) = [\sum_{i=1}^{n} \lvert x_i - z_i \rvert^s]^{\frac{1}{s}}$	Γενικός τύπος
Mahalanobis	$D_M(\boldsymbol{x}, \boldsymbol{z}) = (\boldsymbol{x} - \boldsymbol{z})^T \boldsymbol{C}^{-1} (\boldsymbol{x} - \boldsymbol{z})$	Στατιστική

της μετρικής του Mahalanobis μπορούν να συγκριθούν με τις αντίστοιχες ισομετρικές της Ευκλείδειας μετρικής στο Σχήμα 4.6 και τις ισομετρικές αποστάσεις της Ιπποδάμειας μετρικής στο Σχήμα 4.8 ώστε να κατανοηθεί η διαφορά μεταξύ των τριών μετρικών συστημάτων.

Ο Πίνακας 4.1 παρουσιάζει συνοπτικά τα διάφορα μέτρα απόστασης τα οποία παρουσιάστηκαν, τους μαθηματικούς ορισμούς τους, καθώς και παρατηρήσεις πάνω στην χρησιμότητα τους.

4.4. Μέτρα ομοιότητας

Εκτός από τα μέτρα απόστασης που μας δείχνουν πόσο διαφορετικά είναι δύο πρότυπα, υπάρχουν και τα *μέτρα ομοιότητας* που δείχνουν πόσο όμοια είναι δύο πρότυπα μεταξύ τους. Εάν για δύο πρότυπα το μέτρο ομοιότητας τους είναι μεγάλο, τότε η απόσταση τους θα είναι μικρή και αντίστροφα.

Ένα από τα πιο γνωστά μέτρα ομοιότητας είναι το *εσωτερικό γινόμενο* δύο διανυσμάτων, το οποίο ορίζεται ως:

$$S_I(\boldsymbol{x}, \boldsymbol{z}) = \boldsymbol{x}^T \boldsymbol{z} = \sum_{i=1}^{n} x_i z_i \tag{4.18}$$

Στις περισσότερες περιπτώσεις το εσωτερικό γινόμενο χρησιμοποιείται όταν τα διανύσματα $\boldsymbol{x}$ και $\boldsymbol{z}$ είναι κανονικοποιημένα, δηλαδή έχουν μήκος 1. Τότε, το εσωτερικό γινόμενο $S_I(\boldsymbol{x}, \boldsymbol{z})$ εξαρτάται μόνο από την γωνία που σχηματίζουν τα διανύσματα $\boldsymbol{x}$ και $\boldsymbol{z}$. Τα όρια του εσωτερικού γινομένου είναι:

$$-1 \leq S_I(\boldsymbol{x}, \boldsymbol{z}) \leq +1 \tag{4.19}$$

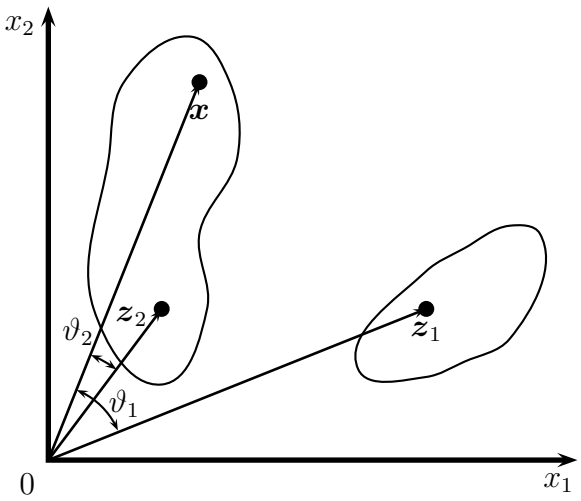
Όταν τα διανύσματα $\boldsymbol{x}$ και $\boldsymbol{z}$ δεν είναι κανονικοποιημένα, τότε μπορεί να χρησιμοποιηθεί το *συνημίτονο της γωνίας* μεταξύ των διανυσμάτων, το οποίο δίδεται από τη σχέση:

$$S_C(\boldsymbol{x}, \boldsymbol{z}) = \frac{\boldsymbol{x}^T \boldsymbol{z}}{\|\boldsymbol{x}\|_2 \, \|\boldsymbol{z}\|_2} \tag{4.20}$$

Το εσωτερικό γινόμενο εκφράζει την *συσχέτιση* (*correlation*) μεταξύ των διανυσμάτων $\boldsymbol{x}$ και $\boldsymbol{z}$. Η συσχέτιση δύο διανυσμάτων παίρνει την μέγιστη της τιμή όταν τα $\boldsymbol{x}$ και $\boldsymbol{z}$ έχουν την ίδια κατεύθυνση. Αυτό το μέτρο ομοιότητας είναι χρήσιμο όταν υπάρχουν ομάδες οι οποίες αναπτύσσονται κατά μήκος των *πρωταρχικών αξόνων*. Ένα τέτοιο παράδειγμα απεικονίζεται στο Σχήμα 4.10 όπου φαίνεται καθαρά ότι το διάνυσμα $\boldsymbol{x}$ είναι πιο κοντά στο αντιπροσωπευτικό διάνυσμα $\boldsymbol{z}_2$ από ότι στο $\boldsymbol{z}_1$. Το συνημίτονο της γωνίας ϑ_1 μεταξύ του $\boldsymbol{x} = [2, 5]^T$ και του αντιπροσωπευτικού διανύσματος $\boldsymbol{z}_1 = [5, 2]^T$ είναι:

$$S_C(\boldsymbol{x}, \boldsymbol{z}_1) = \cos\vartheta_1 = \frac{\boldsymbol{x}^T \boldsymbol{z}_1}{\|\boldsymbol{x}\|_2 \, \|\boldsymbol{z}_1\|_2}, \; = \frac{20}{\sqrt{29}\sqrt{29}} = 0.69$$

Παρόμοια, το συνημίτονο της γωνίας ϑ_2 μεταξύ του διανύσματος $\boldsymbol{x}$ και του αντιπροσωπευτικού διανύσματος $\boldsymbol{z}_2 = [1.5, 2]^T$ είναι:



Σχήμα 4.10: *Δείκτης απόστασης συνημίτονου.*

$$S_C(\boldsymbol{x}, \boldsymbol{z}_2) = \cos\vartheta_2 = \frac{\boldsymbol{x}^T\,\boldsymbol{z}_2}{\|\boldsymbol{x}\|_2\,\|\boldsymbol{z}_2\|_2}, = \frac{13}{\sqrt{29}\;2.5} = 0.97$$

δηλαδή $S_C(\boldsymbol{x}, \boldsymbol{z}_2) > S_C(\boldsymbol{x}, \boldsymbol{z}_1)$ και όπως φαίνεται στο Σχήμα 4.10 το διάνυσμα $\boldsymbol{x}$ κατατάσσεται στην κατηγορία C_2.

Το εσωτερικό γινόμενο ή συνημίτονο της γωνίας είναι κατάλληλα μέτρα ομοιότητας όταν οι ομάδες απέχουν αρκετή απόσταση τόσο μεταξύ τους, όσο και από την αρχή των αξόνων.

Ένα άλλο σημαντικό μέτρο ομοιότητας που χρησιμοποιείται ευρέως είναι η μετρική *Tanimoto* [Tan58]. Μπορεί να χρησιμοποιηθεί τόσο για πραγματικές τιμές όσο και για διακριτές και ορίζεται ως:

$$S_T(\boldsymbol{x}, \boldsymbol{z}) = \frac{\boldsymbol{x}^T\,\boldsymbol{z}}{\boldsymbol{x}^T\,\boldsymbol{x} + \boldsymbol{z}^T\,\boldsymbol{z} - \boldsymbol{x}^T\,\boldsymbol{z}} \tag{4.21}$$

Στο παράδειγμα του Σχήματος 4.10 η μετρική *Tanimoto* μεταξύ του $\boldsymbol{x} = [2, 5]^T$ και του αντιπροσωπευτικού διανύσματος $\boldsymbol{z}_1 = [5, 2]^T$ είναι:

$$S_T(\boldsymbol{x}, \boldsymbol{z}_1) = \frac{\boldsymbol{x}^T \boldsymbol{z}_1}{\boldsymbol{x}^T \boldsymbol{x} + \boldsymbol{z}_1^T \boldsymbol{z}_1 - \boldsymbol{x}^T \boldsymbol{z}_1} = \frac{20}{29 + 29 - 20} = 0.53$$

Παρόμοια, η μετρική *Tanimoto* μεταξύ του διανύσματος $\boldsymbol{x}$ και του αντιπροσωπευτικού διανύσματος $\boldsymbol{z}_2 = [1.5, 2]^T$ είναι:

$$S_T(\boldsymbol{x}, \boldsymbol{z}_2) = \frac{\boldsymbol{x}^T \boldsymbol{z}_2}{\boldsymbol{x}^T \boldsymbol{x} + \boldsymbol{z}_2^T \boldsymbol{z}_2 - \boldsymbol{x}^T \boldsymbol{z}_2} = \frac{13}{29 + 2.5 - 13} = 0.70$$

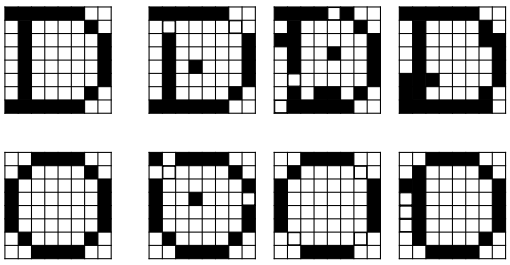
δηλαδή $S_T(\boldsymbol{x}, \boldsymbol{z}_2) > S_T(\boldsymbol{x}, \boldsymbol{z}_1)$ και όπως φαίνεται στο Σχήμα 4.10 το διάνυσμα $\boldsymbol{x}$ κατατάσσεται στην κατηγορία C_2.

Η μετρική Tanimoto χρησιμοποιείται κυρίως για διακριτές τιμές και βασίζεται στην σύγκριση δύο συνόλων. Εφόσον πρόκειται για διακριτές τιμές η Εξίσωση 4.21 αναπαρασταίνει τον λόγο του αριθμού των κοινών στοιχείων δύο διανυσμάτων δια του αριθμού των στοιχείων που διαφέρουν.

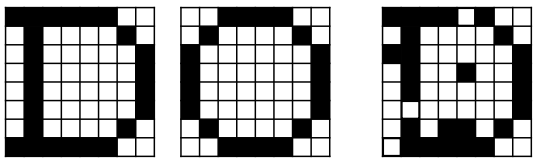
4.5. Ταίριασμα με υποδείγματα

Το *ταίριασμα με υποδείγματα* (*template matching*) είναι μια φυσική προσέγγιση στο πρόβλημα της αναγνώρισης προτύπων χρησιμοποιώντας μέτρα απόστασης ή μέτρα ομοιότητας. Προϋπόθεση είναι να υπάρχουν πρότυπα *υποδείγματα* (*templates*), δηλαδή τα αντιπροσωπευτικά διανύσματα των κατηγοριών, και όταν παρουσιαστεί ένα άγνωστο πρότυπο απλά συγκρίνεται με κάθε ένα από τα υποδείγματα και κατατάσσεται στην κατηγορία του υποδείγματος που ταιριάζει καλύτερα.

Για την ταξινόμηση των χαρακτήρων του Σχήματος 4.11 σε δύο κατηγορίες οι χωρίς θόρυβο χαρακτήρες στην αριστερή πλευρά του σχήματος μπορούν να χρησιμοποιηθούν σαν *υποδείγματα* (*templates*). Τα



Σχήμα 4.11: *Ταίριασμα με υποδείγματα.*



Σχήμα 4.12: *Ταίριασμα με υποδείγματα ενός χαρακτήρα.*

υποδείγματα αποτελούν τα αντιπροσωπευτικά διανύσματα των δυο κατηγοριών. Για την ταξινόμηση ενός χαρακτήρα που περιέχει θόρυβο, ο οποίος συνήθως προκύπτει κατά την διαδικασία σάρωσης ή ψηφιοποίησης συγκρίνεται με κάθε ένα από τα υποδείγματα. Για την σύγκριση μπορεί να χρησιμοποιηθεί ένα από τα μέτρα απόστασης του Πίνακα 4.1 όπως η Ευκλείδεια ή η Ιπποδάμεια απόσταση. Το ποιο κατάλληλο μέτρο απόστασης για να χρησιμοποιηθεί στη συγκεκριμένη εφαρμογή είναι ή Ιπποδάμεια απόσταση, δηλαδή η απόσταση Hamming, διότι οι τιμές είναι δυαδικές. Για παράδειγμα, η απόσταση του αλλοιωμένου χαρακτήρα στη δεξιά πλευρά του Σχήματος 4.12 από τα υποδείγματα είναι:

$$D_H(\boldsymbol{x} - \boldsymbol{z}_D) = 7$$
$$D_H(\boldsymbol{x} - \boldsymbol{z}_O) = 13$$

Άρα ο χαρακτήρας θα ταξινομηθεί σαν "D".

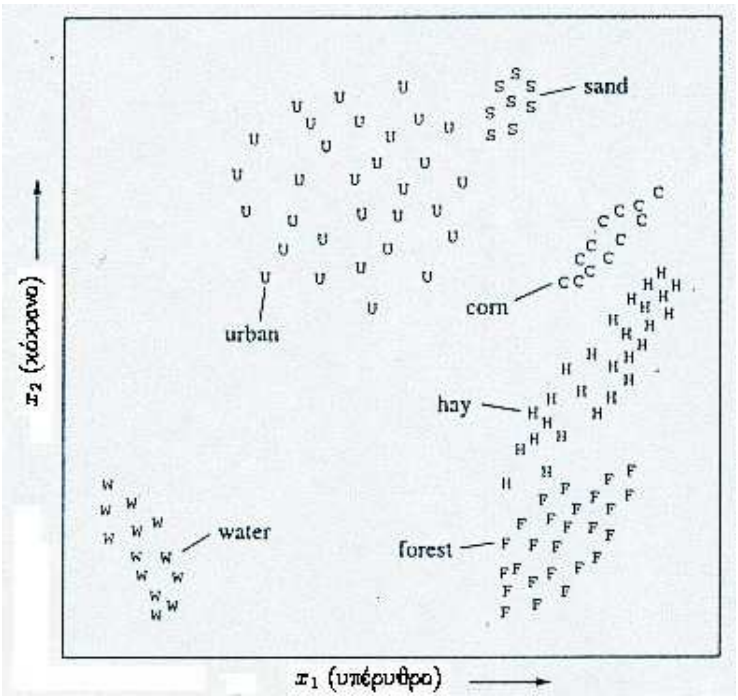
Η τεχνική του ταιριάσματος με υποδείγματα χρησιμοποιείται όταν η παραλλακτικότητα μεταξύ των κατηγοριών οφείλεται σε θόρυβο. Στη παραπάνω εφαρμογή μπορεί να χρησιμοποιηθεί διότι δεν υπάρχουν άλλες παραμορφώσεις των χαρακτήρων, παραμορφώσεις που μπορεί να προέρχονται από μετάθεση, περιστροφή, ψαλίδισμα, τύλιγμα, συστολή, διαστολή, επικάλυψη, αλλαγή κλίμακας κλπ. Τυπικές εφαρμογές που χρησιμοποιείται είναι σε αναγνώριση ομιλίας και απλά προβλήματα τεχνητής όρασης.

4.6. Σύστημα Αναγνώρισης Δορυφορικής Εικόνας

Θα παρουσιαστεί ένα απλό σύστημα αναγνώρισης προτύπων για την επεξεργασία δορυφορικής εικόνας βασισμένο στις αρχές που αναπτύχθηκαν στις προηγούμενες ενότητες. Παρόμοια συστήματα έχουν αναπτυχθεί σε πολλούς τομείς, όπως η πρόβλεψη καιρού, η ανίχνευση ασθενειών σε καλλιέργειες και η κατάστρωση του Εθνικού Κτηματολογίου. Το συγκεκριμένο σύστημα προσδιορίζει τις υπάρχουσες χρήσεις γης με την αναγνώριση προτύπων από δορυφορικές εικόνες. Η όλη διαδικασία βασίζεται στον διαφορετικό τρόπο με τον οποίο απορροφά και αντανακλά το ηλιακό φως το έδαφος, σε διάφορες περιοχές του φάσματοσ[**?**]. Το Σχήμα 4.13 παρουσιάζει ένα δισδιάστατο διάγραμμα προτύπων, για την αναγνώριση χρήσεις γης από δορυφορικές εικόνες. Οι άξονες του σχήματος είναι οι παρακάτω δύο χρωματικές λωρίδες φάσματος:

- x_1 (υπέρυθρο) – Έχει μεγάλη αντανάκλαση σε περιοχές με νερό.
- x_2 (κόκκινο) – Έχει μεγάλη απορρόφηση σε περιοχές με βλάστηση.

Τα x_1 και x_2 αποτελούν τα χαρακτηριστικά γνωρίσματα των προτύπων. Οι πιθανές κατηγορίες ταξινόμησης είναι:

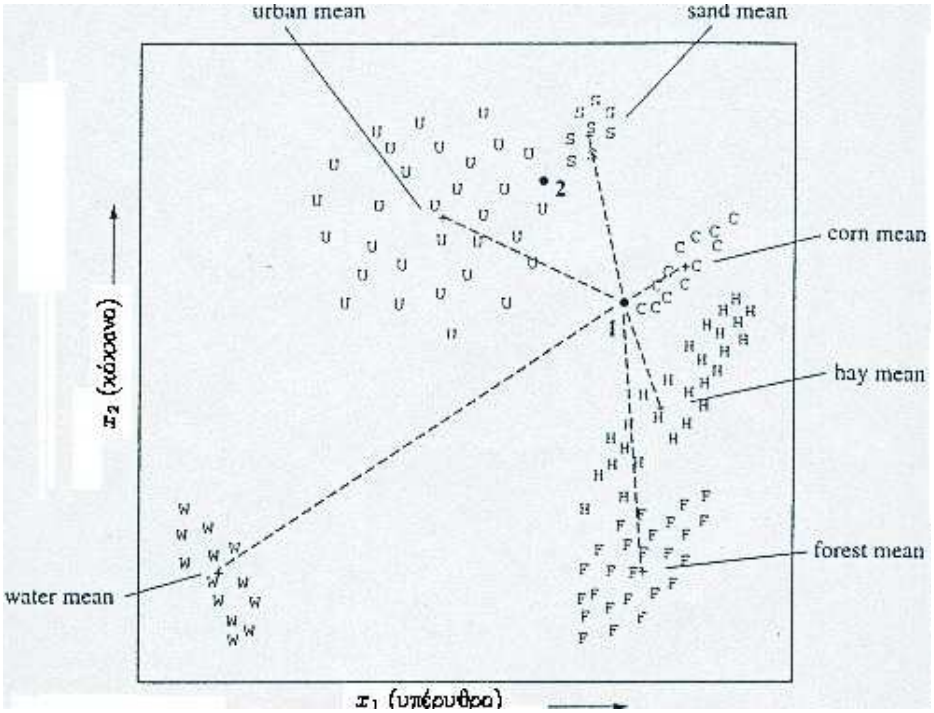


Σχήμα 4.13: *Δισδιάστατο διάγραμμα προτύπων για την αναγνώριση χρήσεις γης από δορυφορικές εικόνες.*

- $\diamond S$ – Sand (Αμμώδης περιοχή)
- $\diamond H$ – Hay (Καλλιέργειες σανού)
- $\diamond W$ – Water (Νερό)
- $\diamond U$ – Urban (Αστική περιοχή)
- $\diamond C$ – Corn (Καλλιέργειες καλαμποκιού)
- $\diamond F$ – Forest (Δασική περιοχή)

Όπως φαίνεται από το Σχήμα 4.13, η κατηγορία W (νερό) είναι τελείως διαχωρίσιμη από τις υπόλοιπες. Όμως, ορισμένες κατηγορίες, όπως η H (καλλιέργειες σανού) και η F (δάση) δεν είναι τόσο καλά διαχωρίσιμες.

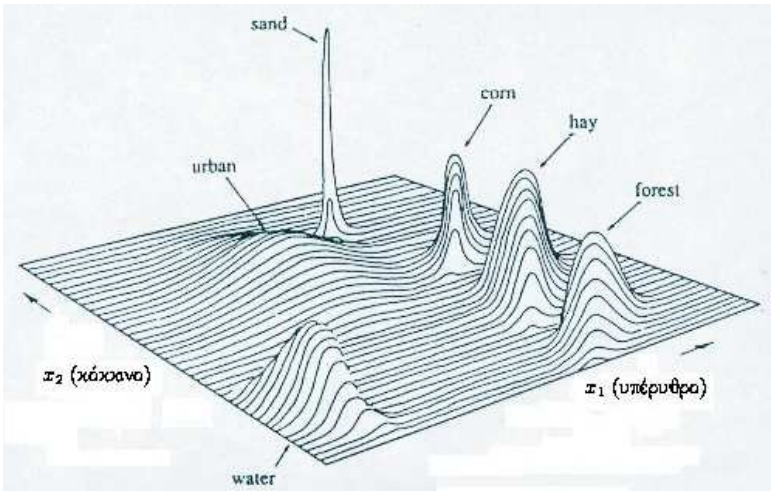
Ο στόχος σε αυτή την εφαρμογή είναι η αναγνώριση της σωστής κατηγορίας για κάθε εικονοκύταρο μίας δορυφορικής φωτογραφίας. Ο



Σχήμα 4.14: *Ταξινομητής ελάχιστης απόστασης.*

αριθμός των εικονοκυτάρων είναι περίπου μισό εκατομμύριο. Το σύστημα αναγνώρισης βασίζεται σε μερικές εκατοντάδες μετρήσεις εδάφους που αντιστοιχούν σε συγκεκριμένα πρότυπα της εικόνας τα οποία φαίνονται στο Σχήμα 4.13. Επειδή οι μετρήσεις εδάφους είναι χρονοβόρες και πολυδάπανες το σύστημα αναγνώρισης προσπαθεί να βελτιστοποιήσει τις πληροφορίες που περιέχουν.

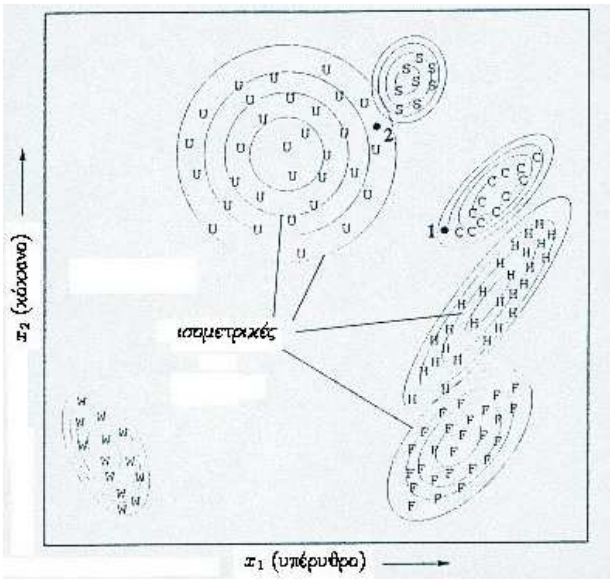
Το αντιπροσωπευτικό διάνυσμα κάθε κατηγορίας συμβολίζεται με μία μαύρη στρογγυλή ένδειξη στο Σχήμα 4.14 και είναι ο μέσος όρος των γνωστών προτύπων, δηλαδή των μετρήσεων εδάφους κάθε κατηγορίας. Για την ταξινόμηση ενός αγνώστου εικονοκυτάρου του οποίου τα χαρακτηριστικά γνωρίσματα x_1, x_2 είναι το σημείο 1 στο Σχήμα 4.14, υπολογίζεται η απόσταση του από κάθε αντιπροσωπευτικό πρότυπο κάθε



Σχήμα 4.15: *Στατιστική κατανομή των κατηγοριών σε τρισδιάστατη μορφή.*

κατηγορίας, όπως φαίνεται στο Σχήμα 4.14. Η απόσταση του 1 από το αντιπροσωπευτικό διάνυσμα της κατηγορίας C είναι η μικρότερη και έτσι το 1 κατατάσσεται στην κατηγορία καλλιέργειες καλαμποκιού. Παρόμοια, το 2 θα ταξινομηθεί στην κατηγορία αμμώδης περιοχή S δίοτι βρίσκεται πλησιέστερα στο αντιπροσωπευτικό πρότυπο της κατηγορίας S.

Αυτή η προσέγγιση είναι απλή και χωρίς πολύπλοκους υπολογισμούς. Όμως δεν λαμβάνονται υπόψη οι ιδιότητες της στατιστικής κατανομής των προτύπων. Για παράδειγμα, τα πρότυπα των αστικών περιοχών (U) είναι περισσότερο "απλωμένα" από ότι τα πρότυπα τα οποία ανήκουν σε αμμώδεις περιοχές (S). Το Σχήμα 4.15 δείχνει τις στατιστικές κατανομές των παραπάνω κατηγοριών σε τρεις διαστάσεις. Επανεξετάζοντας προσεκτικά το Σχήμα 4.14 φαίνεται ότι το 2 έχει ταξινομηθεί σε λάθος κατηγορία. Η λάθος κατάταξη οφείλεται στο γεγονός ότι οι ταξινομητές ελάχιστης απόστασης δεν λαμβάνουν υπόψη τις στατιστικές



Σχήμα 4.16: *Ισομετρική κατανομή των κατηγοριών.*

κατανομές των δεδομένων. Στο Σχήμα 4.16 φαίνεται καθαρά ότι λόγω της στατιστικής κατανομής των προτύπων το πρότυπο 2 ανήκει στην κατηγορία των αστικών περιοχών (U).

4.7. Ασκήσεις

4.1 Για τις κατηγορίες:

$$C_1 = \{[2,1],[2,2],[3,2],[3,1]\}$$
$$C_2 = \{[-2,-1],[-2,-2],[-3,-2]\}$$
$$C_3 = \{[1,-1],[1,-2],[2,-2]\}$$
$$C_4 = \{[-2,2],[-3,2]]\}$$

α) Να υπολογιστούν τα αντιπροσωπευτικά διανύσματα (μέσο διάνυσμα) $\boldsymbol{z_1}$, $\boldsymbol{z_2}$, $\boldsymbol{z_3}$ και $\boldsymbol{z_4}$, των κατηγοριών C_1, C_2, C_3, και C_4 αντίστοιχα.

β) Να σχεδιαστεί ένας ταξινομητής ελάχιστης απόστασης και να ταξινομηθούν τα άγνωστα πρότυπα:

$$[1,1],[7,1],[-1,1],[3,-1],[0,0],[-3,-1]$$

4.2 Να υπολογιστεί η Ευκλείδεια, Ιπποδάμεια, και η Chebyshev απόσταση των διανυσμάτων:

$$\boldsymbol{x} = [1.5, -2.2, 4.4, 3.2]^T$$
$$\boldsymbol{y} = [0.5, 1.2, -2.6, -6.5]^T$$
$$\boldsymbol{z} = [-2.1, 3.9, 12.1, 3.9]^T$$

4.3 Να υπολογιστεί η Ιπποδάμεια, η απόσταση Hamming και η μετρική Tanimoto των διανυσμάτων:

$$\boldsymbol{x} = [0,1,1,0,1,1,0,0,1,0]^T$$
$$\boldsymbol{y} = [0,0,0,0,1,0,1,1,1,1]^T$$

4.4 Τα αντιπροσωπευτικά διανύσματα των κατηγοριών C_1, C_2, και C_3 είναι:

$$\boldsymbol{z_1} = [2,3]^T$$
$$\boldsymbol{z_2} = [3,-2]^T$$
$$\boldsymbol{z_3} = [-2,2]^T$$

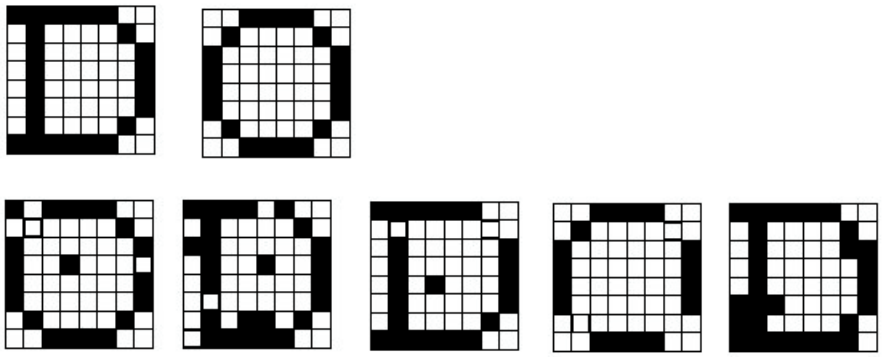
Να ταξινομηθούν τα άγνωστα πρότυπα:

$$[3,1],[1,-1],[2,-2],[-2,1],[0,3],[3,0]$$

χρησιμοποιώντας:

α) Ευκλείδεια απόσταση.

β) Εσωτερικό γινόμενο (συνημίτονο της γωνίας).



Σχήμα 4.17: *Υποδείγματα και αλλοιωμένοι χαρακτήρες.*

γ) Μετρική Tanimoto

4.5 Να ταξινομηθούν οι αλλοιωμένοι χαρακτήρες του Σχήματος 4.17 στις κατηγορίες D ή O χρησιμοποιώντας την μέθοδο ταίριασμα με υποδείγματα.

4.6 Να αποτυπωθούν με χάρακα οι συντεταγμένες των προτύπων κάθε κατηγορίας του Σχήματος 4.14. Να σχεδιαστεί ένας ταξινομητής ελάχιστης απόστασης και τα ταξινομηθούν τα άγνωστα πρότυπα 1 και 2 του Σχήματος 4.14.

Κεφάλαιο 5

Ομαδες

5.1. Δημιουργία ομάδων

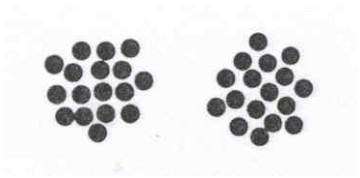
Σε ένα μεγάλο αριθμό επιστημονικών κλάδων χρειάζεται να ταξινομηθούν δεδομένα βάση της ομοιότητας που υπάρχει στα δεδομένα. Σε πολλές περιπτώσεις υπάρχει πολύ μικρή γνώση για την μορφή και τους νόμους που διέπουν τα δεδομένα και πρέπει να γίνουν όσο το δυνατόν λιγότερες αυθαίρετες υποθέσεις για αυτά. Αρχικά, απαιτείται μια πρώτη εξέταση των εσωτερικών σχέσεων μεταξύ των δεδομένων, έτσι ώστε να γίνουν κάποιες πρώτες υποθέσεις για τη δομή τους.

Η *ομαδοποίηση* (*clustering*) είναι εργαλείο διερευνητικής ανάλυσης δεδομένων που προσπαθεί να εκτιμήσει τις σχέσεις που υπάρχουν μεταξύ

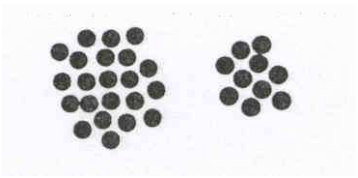
προτύπων οργανώνοντας τα σε *ομάδες* (*clusters*) ή *κατηγορίες* με τέτοιο τρόπο ώστε πρότυπα που ανήκουν σε μια ομάδα να έχουν περισσότερη ομοιότητα μεταξύ τους από ότι πρότυπα που ανήκουν σε διαφορετικές ομάδες. Τα αποτελέσματα της ομαδοποίησης μπορεί να χρησιμοποιηθούν για την εξαγωγή υποθέσεων που αφορούν τα δεδομένα για την ταξινόμηση νέων δεδομένων, για τον έλεγχο ομοιογένειας των δεδομένων καθώς και για συμπίεση δεδομένων.

Το Σχήμα 5.1 παρουσιάζει ορισμένα σύνολα δεδομένων στον δισδιάστατο χώρο. Αν και οπτικά είναι εμφανές ότι κάθε σύνολο δεδομένων αποτελείται από δύο ομάδες δεν υπάρχει μια μοναδική τεχνική ομαδοποίησης η οποία θα μπορούσε να αποκαλύψει όλες αυτές τις δομές. Η πλειονότητα των τεχνικών τείνουν να δημιουργούν ομάδες συγκεκριμένης μορφής. Η ανθρώπινη αντίληψη είναι η καλύτερη τεχνική ομαδοποίησης στον δισδιάστατο και τρισδιάστατο χώρο, αλλά δυστυχώς τα περισσότερα πραγματικά προβλήματα αφορούν ομαδοποιήσεις σε υψηλότερες διαστάσεις. Άλλωστε τα πραγματικά δεδομένα σπάνια ακολουθούν τις ιδανικές δομές που παρουσιάζονται στο Σχήμα 5.1. Έτσι εξηγείται η πληθώρα των μεθόδων που έχουν προταθεί και συνεχίζουν να εμφανίζονταί στην διεθνή βιβλιογραφία. Οι νέοι αλγόριθμοι παρουσιάζουν καλύτερα αποτελέσματα από τους προϋπάρχοντες για συγκεκριμένες, μορφές δεδομένων.

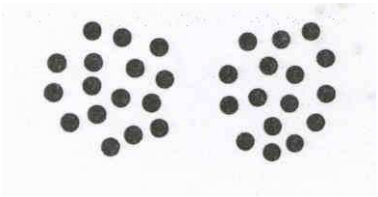
Οι αλγόριθμοι ομαδοποίησης τείνουν να δημιουργούν ομάδες συγκεκριμένης μορφής. Έτσι, από το ίδιο σύνολο δεδομένων δημιουργούνται διαφορετικές ομαδοποιήσεις ανάλογα με την μέθοδο που κάθε φορά χρησιμοποιείται. Στην πραγματικότητα οι αλγόριθμοι θα δημιουργήσουν ομάδες ακόμα και εάν τα δεδομένα είναι εντελώς τυχαία. Αυτό οφείλεται στο ότι οι ομάδες δεν είναι ένα χαρακτηριστικό των δεδομένων, το οποίο μπορεί να ελεγχθεί με ένα αντικειμενικό κριτήριο, αλλά μια οπτική του ερευνητή πάνω στα δεδομένα. Πόσες ομάδες υπάρχουν στο Σχήμα 5.2; δύο ή τέσσερις; Η απάντηση εξαρτάται από το τι αντιπροσωπεύουν τα πρότυπα και ποια έννοια έχει η ομάδα για την εφαρμογή.



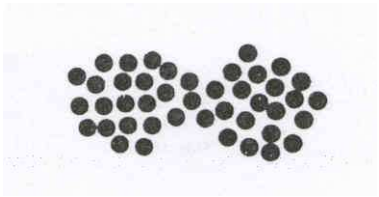
(α) Συμπαγείς και καλά διαχωρίσιμες



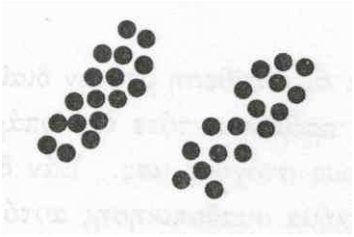
(β) Ανισομεγέθεις ομάδες



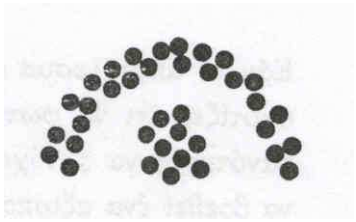
(γ) Μη συμπαγείς



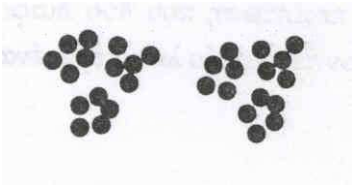
(δ) Εφαπτόμενες ομάδες



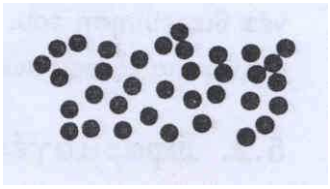
(ε) Επιμήκεις ομάδες.



(ζ) Μη γραμμικά διαχωρίσιμες.



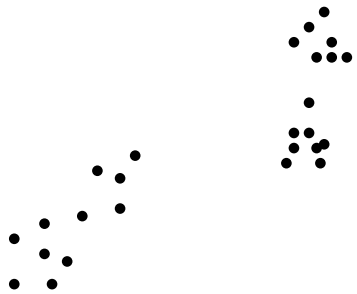
(η) Ομάδες μέσα σε ομάδες.



(θ) Ανυπαρξία ομάδων.

Σχήμα 5.1: *Ομάδες.*

Το μεγαλύτερο ρίσκο που προκύπτει κατά την εφαρμογή μεθόδων ομαδοποίησης είναι ότι μπορεί να έχει σαν επακόλουθο αντί της εύρεσης μιας "φυσικής" δομής των δεδομένων, την επιβολή μιας αυθαίρετης και τεχνητής δομής. Τα αποτελέσματα μιας μεθόδου ομαδοποίησης πρέπει να γίνονται πάντα δεκτά με σκεπτικισμό. Εάν τα αποτελέσματα μιας



Σχήμα 5.2: *Δύο ή τέσσερις ομάδες;*

τεχνικής ομαδοποίησης αυξάνουν την κατανόηση μας για το πρόβλημα, τότε υπάρχει μεγάλη πιθανότητα οι ομάδες να έχουν επιλεγεί σωστά. Εάν το αποτέλεσμα έρχεται σε αντίθεση με την διαίσθηση μας ή συσκοτίζει αντί να φωτίζει το πρόβλημα τότε δεν υπάρχουν πολλές πιθανότητες να επιτευχθούν οι στόχοι μας. Εάν δεν είναι δυνατόν να βρεθεί ένα αξιοπρεπές σχήμα ομαδοποίησης αυτό αποτελεί ένδειξη ότι δεν έχει γίνει πλήρως κατανοητό το πρόβλημα και πρέπει να γίνει νέα διερεύνηση του. Στην περίπτωση που δύο διαφορετικοί διαχωρισμοί φαίνονται εξίσου σωστοί, τότε τον τελευταίο λόγο έχει ένας εμπειρογνώμονας.

5.2. Εφαρμογές

Τεχνικές ομαδοποίησης έχουν εφαρμοστεί σε πολλά και διαφορετικά ερευνητικά προβλήματα. Ο Hartigon [Hart75] περιγράφει μια άριστη περίληψη πολλών δημοσιευμάτων που αφορούν αποτελέσματα ομαδοποίησης σε διάφορες ερευνητικές περιοχές. Για παράδειγμα, στην Ιατρική ομαδοποιήσεις ασθενειών και συμπτωμάτων έχουν σαν αποτέλεσμα την δημιουργία πολύ χρήσιμων ταξινομήσεων. Στην ψυχιατρική, η σωστή

διάγνωση ομάδων συμπτωμάτων όπως η παράνοια και η σχιζοφρένεια είναι βασική προϋπόθεση για μια επιτυχημένη θεραπεία. Στην Βιολογία χρησιμοποιούνται τεχνικές ανάλυσης ομάδων για την ταξινόμηση διαφόρων οργανισμών. Στην Αρχαιολογία ερευνητές έχουν εφαρμόσει τεχνικές ομαδοποίησης για την ταξινόμηση πέτρινων εργαλείων και ταφικών κτερισμάτων. Στην *ανάλυση εικόνας* (*image analysis*) χρησιμοποιείται για την εύρεση ομάδων εικονοκυττάρων με παρόμοια χαρακτηριστικά όπως χρώμα ή υφή (texture). Τα βιβλία των Everitt [Ever81], Gose [Gose96] και Theodoridis [Theo99], περιγράφουν διάφορες εφαρμογές ομαδοποίησης.

5.3. Ιδανικές ομάδες

Ο ορισμός των ομάδων έχει προταθεί στο προηγούμενο τμήμα αλλά κανένας ορισμός δεν είναι αρκετά ικανοποιητικός όπως φαίνεται και από τα παραδείγματα του Σχήματος 5.1. Γενικά, οι ιδιότητες που πρέπει να έχουν τα πρότυπα που ανήκουν σε μια ομάδα, και τι πρέπει να ισχύει για να μπορεί αυτή η ομάδα να θεωρηθεί ιδανική είναι:

(1) Μια ομάδα αποτελείται από ένα σύνολο παρόμοιων προτύπων. Πρότυπα από διαφορετικές ομάδες έχουν διαφορετικά χαρακτηριστικά.

(2) Γενικά η απόσταση μεταξύ των προτύπων που ανήκουν σε μια ομάδα είναι μικρότερη από την απόσταση μεταξύ προτύπων που ανήκουν σε διαφορετικές ομάδες.

(3) Οι ομάδες αποτελούν συνδεμένες περιοχές στον χώρο προτύπων με σχετικά μεγάλη πυκνότητα προτύπων, και διαχωρίζονται από τις άλλες ομάδες με περιοχές με χαμηλή πυκνότητα προτύπων.

Είναι προφανές ότι ο αρχικός σχεδιασμός ενός συστήματος αναγνώρισης προτύπων καθορίζει το τι σημαίνει ομάδα για την εφαρμογή και τις απαιτήσεις από την μέθοδο ομαδοποίησης που θα χρησιμοποιηθεί. Ένα από τα πρώτα βήματα για την κατασκευή ενός συστήματος δεδομένων

είναι η δημιουργία μιας αντίληψης για το πώς σχηματίζονται οι ομάδες. Για το σκοπό αυτό μπορούν να χρησιμοποιηθούν μαθηματικά μοντέλα που περιέχουν *a'priori* γνώση για το πρόβλημα ή πρόχειρες στατιστικές αναλύσεις.

Πολλές μεθοδολογίες ομαδοποίησης οι οποίες προτείνονται στην βιβλιογραφία βασίζονται σε ιδανικές δομές ομαδοποίησης και ουσιαστικά προσαρμόζουν ένα μείγμα στατιστικών κατανομών στα δεδομένα του προβλήματος. Ορισμένοι αλγόριθμοι ομαδοποίησης πάντοτε τοποθετούν τα δύο πλησιέστερα πρότυπα στην ίδια ομάδα.

5.4. Μεθοδολογίες ομαδοποίησης

Οι περισσότεροι αλγόριθμοι ομαδοποίησης μπορούν να ταξινομηθούν σε δύο μεγάλες κατηγορίες: σε *ιεραρχικούς* (*hierarchical*) και σε *διαχωριστικούς* (*partitional*) αλγόριθμους. Ένας ιεραρχικός αλγόριθμος ομαδοποίησης επιβάλει μια ιεραρχική δομή στα δεδομένα, δηλαδή τα δεδομένα χωρίζονται σε μεγάλες ομάδες, που με την σειρά τους χωρίζονται σε υποομάδες κλπ. Το τελικό αποτέλεσμα ενός ιεραρχικού αλγορίθμου είναι ένα *δενδρόγραμμα*.

Ο χωρισμός σε ομάδες μπορεί να ξεκινήσει από τις μικρότερες ομάδες προς τις μεγαλύτερες είτε αντίστροφα από τις μεγαλύτερες προς τις μικρότερες. Στην πρώτη περίπτωση ο αλγόριθμος ξεκινά με N ομάδες, μια για κάθε πρότυπο και σταδιακά οι ομάδες ενοποιούνται μέχρις ότου φτιαχτεί μια ομάδα που περιέχει όλα τα πρότυπα. Η προσέγγιση αυτή καλείται *συσσωρευτική* (*agglomerative*). Στην δεύτερη περίπτωση ο αλγόριθμος ξεκινά με μια ομάδα που περιέχει όλα τα πρότυπα και με μια διαδικασία βημάτων οι ομάδες διασπώνται μέχρις ότου οι ομάδες περιέχουν μόνο λίγα πρότυπα. Η προσέγγιση αυτή καλείται *διαμοραστική* (*divisive*). Οι ιεραρχικές μεθοδολογίες ομαδοποίησης είναι πολύ διαδεδομένες στις

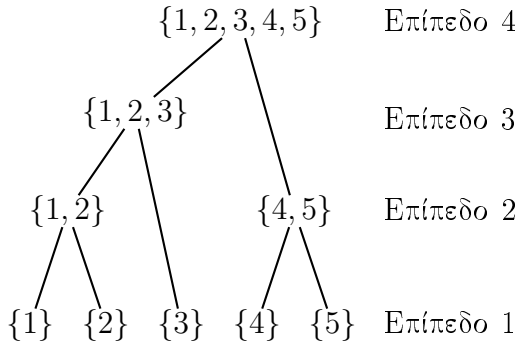
βιολογικές επιστήμες, όπου τα δεδομένα για παράδειγμα φυτά και ζώα συχνά αντιπροσωπεύουν μια ταξινομία.

Στην διαχωριστική ομαδοποίηση ο στόχος είναι η δημιουργία ενός συνόλου ομάδων οι οποίες διαχωρίζουν τα δεδομένα σε παρόμοιες ενότητες. Πρότυπα τα οποία βρίσκονται σε μικρή απόσταση θεωρούνται ότι είναι όμοια και ο στόχος των διαχωριστικών αλγορίθμων είναι να ομαδοποιηθούν τέτοια δεδομένα. Στους περισσότερους διαχωριστικούς αλγορίθμους το σύνολο των ομάδων που θα δημιουργηθούν προκαθορίζεται. Οι διαχωριστικές μέθοδοι χρησιμοποιούν *συναρτήσεις κριτηρίων* (*criterion functions*) όπως μέθοδοι ελαχιστοποίησης τετραγώνων, εκτιμητές πυκνότητας και πλησιέστερους γείτονες. Οι διαχωριστικές μεθοδολογίες ομαδοποίησης χρησιμοποιούνται σε περιπτώσεις που οι κατηγορίες του προβλήματος δεν σχηματίζουν τόσο ευδιάκριτες ομάδες, αλλά υπάρχει κάποια αλληλοεπικάλυψη μεταξύ τους. Σε αυτές τις περιπτώσεις κάθε πρότυπο μπορεί να ανήκει σε πολλές ομάδες ταυτόχρονα με ένα βαθμό σιγουριάς για κάθε ομάδα που κυμαίνεται από 0 έως 1.

Είναι σημαντικό να γίνει διάκριση μεταξύ μεθόδων ομαδοποίησης και αλγόριθμων ομαδοποίησης. Η ίδια μέθοδος ομαδοποίησης μπορεί να υλοποιηθεί διαφορετικά έχοντας σαν αποτέλεσμα την δημιουργία διαφορετικών αλγορίθμων ομαδοποίησης. Για παράδειγμα οι διαχωριστικοί αλγόριθμοι *Forgy's* και *Isodata*, είναι βασισμένοι σε μεθόδους οι οποίες ελαχιστοποιούν το τετραγωνικό σφάλμα.

5.5. Ιεραρχική ομαδοποίηση

Μια ιεραρχία μπορεί να αντιπροσωπευθεί από μια δενδρική δομή, όπως φαίνεται στο Σχήμα 5.3. Σε ένα ιατρείο μικρών ζώων τα ασθενή ζώα σχηματίζουν δύο μεγάλες ομάδες γάτες και σκύλους, κάθε μια από τις οποίες αποτελείται από μικρότερες υποομάδες. Κάθε άρρωστο ζώο αντιπροσωπεύεται από τους αριθμούς 1 έως 5 στα χαμηλότερο επίπεδο του



Σχήμα 5.3: *Παράδειγμα ιεραρχικής ανάλυσης ομάδων.*

δενδρογράμματος, όπως φαίνετε και στο Σχήμα 5.3. Κάθε ένα από αυτά τα πρότυπα αποτελεί μια ομάδα στο χαμηλότερο επίπεδο. Στο πάνω μέρος του σχήματος βρίσκεται η μεγαλύτερη ομάδα η οποία αποτελείται από όλα τα πρότυπα. Οι ομάδες που προκύπτουν σε κάθε επίπεδο είναι:

- **Επίπεδο** 1: $\{1\}, \{2\}$, $\{3\}$, $\{4\}$, $\{5\}$
- **Επίπεδο** 2: $\{1, 2\}$, $\{3\}$, $\{4, 5\}$
- **Επίπεδο** 3: $\{1, 2, 3\}$, $\{4, 5\}$
- **Επίπεδο** 4: $\{1, 2, 3, 4, 5\}$

Σε μια ιεραρχική ομαδοποίηση εάν δύο πρότυπα ανήκουν στην ίδια ομάδα σε ένα επίπεδο, τότε θα ανήκουν στην ίδια ομάδα σε κάθε υψηλότερο επίπεδο. Για παράδειγμα, όπως φαίνεται στο Σχήμα 5.3, τα πρότυπα 1 και 2 που ανήκουν στην ίδια ομάδα στο επίπεδο 2, ανήκουν στην ίδια ομάδα και στα επίπεδα 3 και 4.

Ένας συσσωρευτικός ιεραρχικός αλγόριθμος ομαδοποίησης ακολουθεί την παρακάτω μορφή:

(1) Καταχώρηση κάθε ένα από τα N πρότυπα σε μια μοναδική ομάδα, με αποτέλεσμα την δημιουργία N ομάδων.
(2) Να βρεθούν οι ομάδες με την μεγαλύτερη ομοιότητα μεταξύ τους και να συγχωνευτούν σε μια νέα ομάδα.
(3) Επανάληψη του βήματος 2 έως ότου όλα τα πρότυπα να ανήκουν στην ίδια ομάδα.

Χρησιμοποιώντας διαφορετικές μεθόδους για τον καθορισμό της ομοιότητας μεταξύ των ομάδων προκύπτουν διαφορετικοί αλγόριθμοι. Ένας τρόπος μέτρησης της ομοιότητας μεταξύ ομάδων είναι ο ορισμός μιας συνάρτησης μέτρησης της απόστασης μεταξύ των ομάδων. Τα μέτρα απόστασης μεταξύ ομάδων που χρησιμοποιούνται βασίζονται πάνω στα γνωστά μέτρα απόστασης όπως την Ευκλείδεια και την Ιπποδάμεια απόσταση.

5.6. Αλγόριθμος απλής σύνδεσης

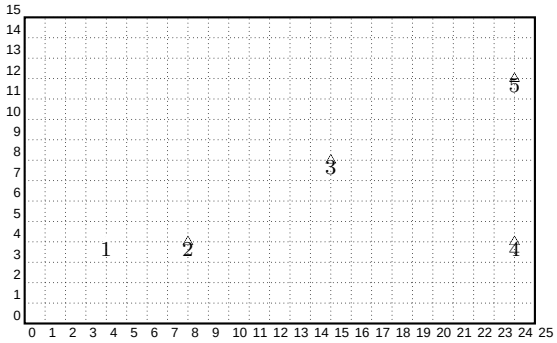
Ο αλγόριθμος απλής σύνδεσης είναι γνωστός στην βιβλιογραφία με πολλά ονόματα όπως μέθοδος πλησιέστερης γειτνίασης (nearest neighbor) ή ελάχιστη μέθοδος (minimum method). Στον αλγόριθμο *απλής σύνδεσης* (*single-linkage*) η απόσταση μεταξύ ομάδων ορίζεται σαν την ελάχιστη απόσταση μεταξύ δύο προτύπων διαφορετικών ομάδων. Με άλλα λόγια η απόσταση μεταξύ δύο ομάδων καθορίζεται από την απόσταση των πιο κοντινών προτύπων διαφορετικών ομάδων. Έτσι η απόσταση μεταξύ των ομάδων C_i και C_j ορίζεται από την σχέση

$$D_{SL}\left(C_i, C_j\right) = \min_{\boldsymbol{x}\in C_i, \boldsymbol{y}\in C_j} d\left(\boldsymbol{x}, \boldsymbol{y}\right), \tag{5.1}$$

όπου $d\left(x, y\right)$ είναι η συνάρτηση απόστασης μεταξύ των προτύπων $\boldsymbol{x}$, $\boldsymbol{y}$.

Για την πρακτική υλοποίηση του αλγορίθμου πρέπει να βρεθεί η απόσταση κάθε ομάδας με κάθε άλλη και τα αποτελέσματα δημιουργούν ένα συμμετρικό δισδιάστατο *πίνακα απόστασης* (*proximity matrix*). Οι

	x_1	x_2
1	4.0	4.0
2	8.0	4.0
3	15.0	8.0
4	24.0	4.0
5	24.0	12.0



Σχήμα 5.4: *Δεδομένα για προβλήματα ιεραρχικής ανάλυσης ομάδων.*

διαστάσεις του πίνακα απόστασης $\boldsymbol{P}$ είναι αρχικά $n \times n$, όπου n είναι ο αριθμός των προτύπων. Οι στήλες και οι γραμμές του πίνακα απόστασης αντιπροσωπεύουν τις ομάδες προτύπων και οι τιμές του πίνακα την απόσταση μεταξύ των ομάδων. Δηλαδή,

$$P_{ij} = D_{SL}\left(C_i, C_j\right), \tag{5.2}$$

Ο καλύτερος τρόπος επίδειξης της υλοποίησης του αλγορίθμου απλής σύνδεσης είναι η πρακτική εφαρμογή του σε ένα παράδειγμα. Το Σχήμα 5.4 παρουσιάζει 5 δισδιάστατα πρότυπα $\boldsymbol{x} = (x_1, x_2)$ σε γραφική μορφή, καθώς και τις ακριβείς τιμές τους. Αρχικά θα δημιουργηθούν 5 ομάδες και σε κάθε μια θα αντιστοιχιστεί ένα πρότυπο. Στην συνέχεια θα βρεθεί η απόσταση κάθε ομάδας με κάθε άλλη και τα αποτελέσματα δημιουργούν τον παρακάτω αρχικό πίνακα απόστασης:

	1	2	3	4	5
1	0.0	4.0	11.7	20.0	21.5
2	4.0	0.0	8.1	16.0	17.9
3	11.7	8.1	0.0	9.8	9.8
4	20.0	16.0	9.8	0.0	8.0
5	21.5	17.9	9.8	8.0	0.0

(5.3)

Η μικρότερη τιμή που εμφανίζεται στον πίνακα απόστασης είναι 4 μεταξύ των προτύπων 1 και 2. Τα πρότυπα 1 και 2 ενοποιούνται σε μια ομάδα και υπάρχει πλέον η κατάταξη:

$$\{1,2\},\{3\},\{4\},\{5\}$$

Στην συνέχεια υπολογίζεται ο παρακάτω νέος πίνακας απόστασης:

	$\{1,2\}$	3	4	5
$\{1,2\}$	0.0	8.1	16.0	17.9
3	8.1	0.0	9.8	9.8
4	16.0	9.8	0.0	8.0
5	17.9	9.8	8.0	0.0

Στον αλγόριθμος απλής σύνδεσης ο νέος πίνακας απόστασης, μετά την ένωση των δύο πλησιέστερων ομάδων, κρατάει τις ελάχιστος τιμές απόστασης στις στήλες που έχουν επηρεαστεί από την ένωση των δυο ομάδων. Στον αρχικό πίνακα απόστασης φαίνεται ότι $d(1,3) = 11.7$ και $d(2,3) = 8.1$. Οπότε η ελάχιστη απόσταση μεταξύ των ομάδων $\{1,2\}$ και $\{3\}$ είναι 8.1. Οι υπόλοιπες τιμές της πρώτης στήλης (ή της πρώτης σειράς) του δεύτερου πίνακα απόστασης υπολογίζονται παρόμοια. Οι υπόλοιπες στήλες και σειρές απλά διατηρούνται και αντιγράφονται από τον αρχικό πίνακα απόστασης.

Στο επόμενο βήμα επαναλαμβάνεται η διαδικασία της εύρεσης της μικρότερης απόστασης, μεταξύ των ομάδων, από το δεύτερο πίνακα απόστασης, η οποία είναι 8.0 μεταξύ των ομάδων $\{4\}$ και $\{5\}$ οι οποίες ενώνονται. Σε αυτό το σημείο υπάρχουν οι τρεις παρακάτω ομάδες:

$$\{1,2\},\{3\},\{4,5\}$$

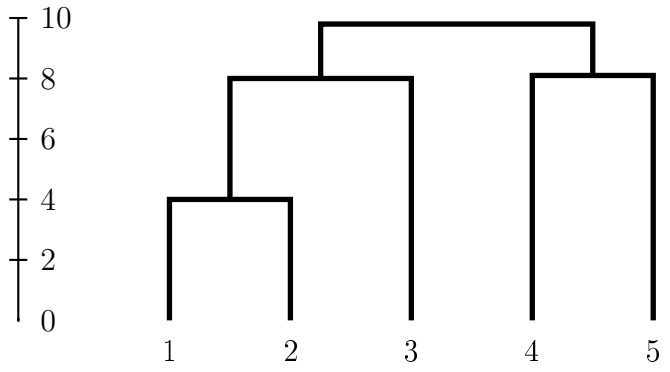
Στην συνέχεια υπολογίζεται ο παρακάτω νέος πίνακας απόστασης:

	$\{1,2\}$	3	$\{4,5\}$
$\{1,2\}$	0.0	8.1	16.0
3	8.1	0.0	9.8
$\{4,5\}$	16.0	9.8	0.0

Εφόσον η ελάχιστη τιμή του πίνακα είναι 8.1, οι ομάδες $\{1,2\}$ και $\{3\}$ ενώνονται. Σε αυτό το σημείο υπάρχουν οι παρακάτω δύο ομάδες:

$$\{1,2,3\}, \{4,5\}$$

Το τελικό βήμα είναι η ένωση των δύο ομάδων των οποίων η απόσταση είναι 9.8. Η ιεραρχική ομαδοποίηση έχει ολοκληρωθεί και το τελικό δενδρόγραμμα παρουσιάζεται στο Σχήμα 5.5. Από το δενδρόγραμμα μπορούν να πραγματοποιηθούν διάφοροι διαχωρισμοί κόβοντας το δενδρόγραμμα σε κάποιο επίπεδο ανομοιότητας. Στο Σχήμα 5.5 κόβοντας το δενδρόγραμμα στην απόσταση 9 έχει σαν αποτέλεσμα ένα διαχωρισμό που αποτελείται από δύο ομάδες. Η μια ομάδα αποτελείται από τα πρότυπα $\{1,2,3\}$ και η άλλη από τα πρότυπα $\{4,5\}$. Για την τελική δημιουργία τεσσάρων ομάδων θα πρέπει το δενδρόγραμμα να κοπεί στην απόσταση 5. Φυσικά, εκ των πρότερων δεν ξέρουμε πόσες ομάδες περιέχονται στο δεδομένα. Που λοιπόν πρέπει να κοπεί το δενδρόγραμμα; Μια ευρετική μέθοδος είναι η επιλογή της απόστασης όπου υπάρχει ένα μεγάλο κάθετο *κενό* στο δενδρόγραμμα. Με άλλα λόγια, σημαντικές ομάδες είναι αυτές που έχουν μεγάλο *χρόνο ζωής* οπού ο χρόνος ζωής μιας ομάδας ορίζεται ως η διαφορά μεταξύ της απόστασης, κατά την οποία η ομάδα ενώνεται με μια άλλη, από την απόσταση από την οποία η ομάδα δημιουργήθηκε. Βασισμένοι σε αυτή την ευρετική μέθοδο είναι πιο λογικό να κοπεί το δενδρόγραμμα του Σχήματος 5.5 στο 5 από ότι στο 9.



Σχήμα 5.5: *Ιεραρχική ομαδοποίηση χρησιμοποιώντας τον αλγόριθμο απλής σύνδεσης.*

5.7. Αλγόριθμος πλήρους σύνδεσης

Ο αλγόριθμος πλήρους σύνδεσης ονομάζεται επίσης η μέγιστη μέθοδος (maximum method) ή απομακρυσμένης γειτνίασης (farthest neighbor).

Στον αλγόριθμο *πλήρους σύνδεσης* (*complete-linkage*) η απόσταση μεταξύ δύο ομάδων ορίζεται ως η μέγιστη απόσταση μεταξύ δύο προτύπων οπού κάθε ένα από τα δύο πρότυπα ανήκει σε διαφορετική ομάδα. Με άλλα λόγια η απόσταση μεταξύ ομάδων καθορίζεται από την μέγιστη απόσταση μεταξύ δύο πρότυπων που ανήκουν σε διαφορετικές ομάδες. Έτσι η απόσταση μεταξύ των ομάδων C_i και C_j ορίζεται από την σχέση:

$$D_{CL}\left(C_i, C_j\right) = \max_{\boldsymbol{x}\in C_i, \boldsymbol{y}\in C_j} d\left(\boldsymbol{x}, \boldsymbol{y}\right), \tag{5.4}$$

όπου $d\left(x, y\right)$ είναι η συνάρτηση απόστασης μεταξύ των προτύπων $\boldsymbol{x}$, $\boldsymbol{y}$.

Ο αλγόριθμος πλήρης σύνδεσης θα παρουσιαστεί χρησιμοποιώντας το παράδειγμα του Σχήματος 5.4. Όπως και στον προηγούμενο αλγόριθμο αρχικά υπολογίζεται ο πίνακας απόστασης (5.3) και δημιουργούνται πέντε ομάδες κάθε μια από τις οποίες αποτελείται από ένα πρότυπο. Κατόπιν, παρόμοια με τον αλγόριθμο απλής σύνδεσης, οι πλησιέστερες ομάδες $\{1\}$

και $\{2\}$ ενώνονται. Έτσι το αποτέλεσμα είναι η δημιουργία των παρακάτω 4 ομάδων:

$$\{1,2\},\{3\},\{4\},\{5\}$$

Στην συνέχεια υπολογίζεται ο παρακάτω νέος πίνακας απόστασης:

	$\{1,2\}$	3	4	5
$\{1,2\}$	0.0	11.7	20.0	21.5
3	11.7	0.0	9.8	9.8
4	20.0	9.8	0.0	8.0
5	21.5	9.8	8.0	0.0

Στον αλγόριθμο πλήρους σύνδεσης ο νέος πίνακας απόστασης, μετά την ένωση των δύο πλησιέστερων ομάδων, κρατάει της μέγιστες τιμές απόστασης στις στήλες που έχουν επηρεαστεί από την ένωση των δύο ομάδων. Στον αρχικό πίνακα απόστασης (5.3) φαίνεται ότι $d(1,3) = 11.7$ και $d(2,3) = 8.1$. Οπότε η μέγιστη απόσταση μεταξύ των ομάδων $\{1,2\}$ και $\{3\}$ είναι 11.7. Οι υπόλοιπες τιμές της πρώτης στήλης του δεύτερου πίνακα απόστασης υπολογίζονται παρόμοια. Οι υπόλοιπες στήλες και σειρές απλά διατηρούνται και αντιγράφονται από τον αρχικό πίνακα απόστασης (5.3).

Επαναλαμβάνεται η διαδικασία της εύρεσης τις μικρότερης απόστασης μεταξύ των ομάδων η οποία είναι 8.0 και είναι η απόσταση μεταξύ μεταξύ των ομάδων $\{4\}$ και $\{5\}$ οι οποίες ενώνονται. Σε αυτό το σημείο υπάρχουν οι παρακάτω τρεις ομάδες:

$$\{1,2\},\{3\},\{4,5\}$$

Υπολογίζεται ο παρακάτω νέος πίνακας απόστασης των τριών ομάδων:

	$\{1,2\}$	3	$\{4,5\}$
$\{1,2\}$	0.0	11.7	21.5
3	11.7	0.0	9.8
$\{4,5\}$	21.5	9.8	0.0

Εφόσον η ελάχιστη τιμή του πίνακα είναι 9.8, οι ομάδες $\{3\}$ και $\{4,5\}$ ενώνονται. Σε αυτό το σημείο υπάρχουν οι παρακάτω δύο ομάδες:

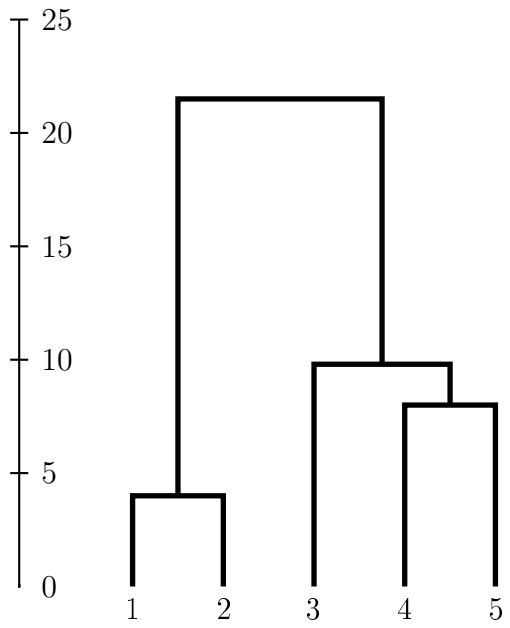
$$\{1,2\}, \{3,4,5\}$$

Πρέπει να σημειωθεί ότι σε αυτό το σημείο οι ομάδες που δημιουργήθηκαν είναι διαφορετικές από της αντίστοιχες του αλγορίθμου απλής σύνδεσης.

Το τελικό βήμα είναι η ένωση των δύο ομάδων. Η ιεραρχική ομαδοποίηση με τον αλγόριθμο πλήρης σύνδεσης έχει ολοκληρωθεί και το τελικό δενδρόγραμμα παρουσιάζεται στο Σχήμα 5.6. Η απόσταση D_{CL} μεταξύ των ομάδων που ενώθηκαν παρουσιάζεται στον κάθετο άξονα.

5.8. Σύγκριση απλής και πλήρους σύνδεσης

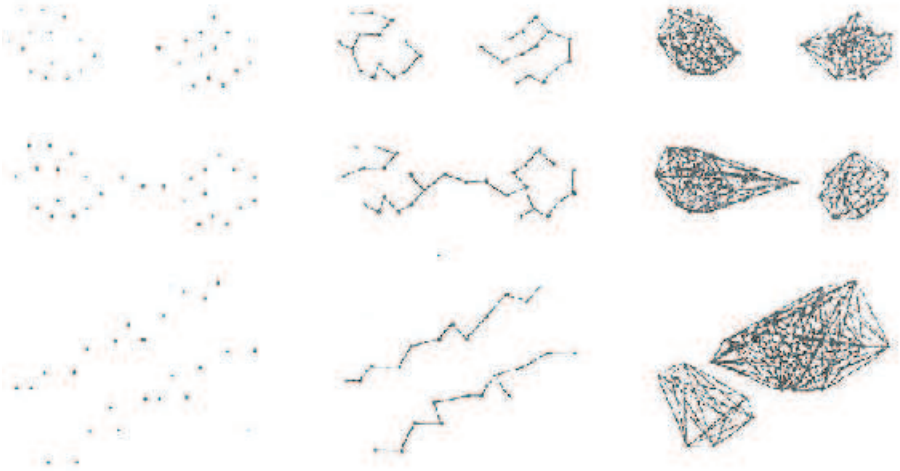
Ο αλγόριθμος απλής σύνδεσης και ο αλγόριθμος πλήρους σύνδεσης διαφέρουν μεταξύ τους στον τρόπο καθορισμού της ομοιότητας των προτύπων, τα οποία ανήκουν σε διαφορετικές κατηγόριες, για την συνένωση τους. Είναι φυσικό οι δύο αλγόριθμοι να έχουν σαν αποτέλεσμα διαφορετικές ομαδοποίησης για τα ίδια δεδομένα. Άρα ποια μέθοδος θα πρέπει να χρησιμοποιηθεί; Δυστυχώς, δεν υπάρχουν ξεκάθαρες οδηγίες για ένα χρήστη. Αρκετές προσπάθειες έχουν πραγματοποιηθεί για να αιτιολογήσουν την χρήση του αλγορίθμου απλής σύνδεσης σε μαθηματική βάση άλλα ο αλγόριθμος μειονεκτεί από το φαινόμενο της *αλυσίδας* όπου απομακρυσμένα πρότυπα τοποθετούνται στην ίδια ομάδα επειδή έχουν ένα



Σχήμα 5.6: *Ιεραρχική ομαδοποίηση χρησιμοποιώντας τον αλγόριθμο πλήρης σύνδεσης*

κοινό γειτονικό πρότυπο. Οι Duda και Hart [DuHa73] παρουσιάζουν διαγραμματικά τα αποτελέσματα ομαδοποίησης των δύο αλγορίθμων χρησιμοποιώντας τρία παραδείγματα. Τα παραδείγματα μαζί με τα αποτελέσματα των αλγορίθμων απεικονίζονται στον Σχήμα 5.7.

Η γραφική αναπαράσταση του αλγόριθμου απλής σύνδεσης είναι μια *ελάχιστη δενδρική εξάπλωση* (*minimum spanning tree*) η οποία δημιουργείται προσθέτοντας την πιο κοντινή ακμή μεταξύ των δύο ομάδων. Στο πρώτο παράδειγμα του Σχήματος 5.7(α) οι ομάδες είναι συμπαγείς και καλά διαχωρίσιμες οπότε ο αλγόριθμος απλής σύνδεσης βρίσκει εύκολα τις ομάδες. Στην δεύτερη περίπτωση, η παρουσίαση ορισμένων προτύπων των οποίων η θέση στον χώρο δημιουργεί μια γέφυρα μεταξύ των ομάδων έχει σαν αποτέλεσμα την δημιουργία απρόσμενων ομάδων, όπως φαίνεται στο Σχήμα 5.7(β). Έτσι, ο αλγόριθμος απλής σύνδεσης δημιουργεί μια μεγάλη επιμήκη ομάδα και μια μικρή και συμπαγή. Εδώ φαίνεται ξεκάθαρα



(α) (β) (γ)

Σχήμα 5.7: *Παραδείγματα ιεραρχικής ομαδοποίησης. (α) Τρία σύνολα δεδομένων. (β) Αποτελέσματα του αλγόριθμου απλής σύνδεσης. (γ) Αποτελέσματα του αλγόριθμου πλήρης σύνδεσης.*

το φαινόμενο της αλυσίδας το οποίο αποτελεί μειονέκτημα του αλγορίθμου απλής σύνδεσης. Εφόσον τα αποτελέσματα μιας ομαδοποίησης είναι ιδιαίτερα ευαίσθητα στον θόρυβο και σε μικρές αποκλίσεις των προτύπων στον χώρο, τότε πράγματι το φαινόμενο της αλυσίδας είναι ένα μειονέκτημα. Όμως, η τάση του αλγορίθμου της δημιουργίας αλυσίδας μπορεί να είναι και πλεονέκτημα εάν οι ομάδες είναι επιμήκεις, όπως φαίνεται στο αποτέλεσμα του τρίτου παραδείγματος στο Σχήμα 5.7(β).

Η γραφική αναπαράσταση του αλγορίθμου πλήρους σύνδεσης είναι η δημιουργία ενός σχεδιαγράμματος γράφου στο οποίο ακμές ενώνουν όλα τα πρότυπα ή κόμβους σε μία ομάδα. Στην ορολογία της θεωρίας γράφων, η κάθε ομάδα αποτελεί ένα πλήρες υπογράφο. Η απόσταση μεταξύ των ομάδων καθορίζεται από την απόσταση των πιο απομακρυσμένων προτύπων στις δύο ομάδες. Όταν γειτονικά πρότυπα συνενώνονται ο γράφος αλλάζει προσθέτοντας ακμές μεταξύ όλων των πρότυπων κάθε

ομάδας. Εάν οριστεί η *διάμετρος* μιας ομάδας ως η μεγίστη απόσταση μεταξύ προτύπων στην ομάδα, τότε η απόσταση μεταξύ δύο ομάδων είναι η διάμετρος της συνένωσης των δύο ομάδων. Κάθε επανάληψη των βημάτων του αλγορίθμου πλήρης σύνδεσης αυξάνει κατά το λιγότερο δυνατόν την διάμετρο της νέας ομάδας. Αυτό είναι πλεονέκτημα όταν οι πραγματικές ομάδες είναι συμπαγείς και παρόμοιες σε μέγεθος, όπως στα πρώτα δύο παραδείγματα του Σχήματος 5.7(γ). Αντίθετα, όταν οι πραγματικές ομάδες δεν έχουν αυτά τα χαρακτηριστικά, όπως το τρίτο παράδειγμα του Σχήματος 5.7(γ), τότε τα αποτελέσματα μπορεί να είναι αυθαίρετα. Το αποτέλεσμα του τρίτου παραδείγματος στο Σχήμα 5.7(γ) είναι ένα παράδειγμα επιβολής των δομών στα δεδομένα αντί της εύρεσης της δομής μέσα από τα δεδομένα.

Περιληπτικά, ο αλγόριθμος απλής σύνδεσης έχει καλά αποτελέσματα σε περίπτωσης οπού πραγματικές ομάδες είναι επιμήκης στον χώρο. Αντίθετα, ο αλγόριθμος πλήρους σύνδεσης έχει καλά αποτελέσματα όταν οι ομάδες είναι συμπαγείς και παρόμοιες σε μέγεθος.

Επιπρόσθετοι ιεραρχικοί αλγόριθμοι ομαδοποίησης μπορούν να χρησιμοποιήσουν άλλες συναρτήσεις όπως μέσες τιμές, και το κέντρο βάρους (centroid). Ο αλγόριθμος που χρησιμοποιεί συναρτήσεις μέσων τιμών ονομάζεται σύνδεσης μέσων τιμών (average linkage) όπως ο αλγόριθμος που χρησιμοποιεί συναρτήσεις κέντρων βάρους ονομάζεται σύνδεση κέντρων βάρους (centroid linkage). Επιπλέον, ένας άλλος ιεραρχικός αλγόριθμος ο οποίος χρησιμοποιεί τεχνικές ανάλυσης διασποράς (analysis of variance) για τον υπολογισμό των αποστάσεων μεταξύ ομάδων, ονομάζεται και μέθοδος Ward. Αναλυτικά θα αναπτυχθούν με παραδείγματα οι αλγόριθμοι σύνδεσης μέσων τιμών και η μέθοδος Ward.

5.9. Αλγόριθμος σύνδεσης μέσων τιμών

Ο αλγόριθμος απλής σύνδεσης δημιουργεί ομάδες επιμήκης ενώ ο αλγόριθμος πλήρης σύνδεσης παράγει πιο συμπαγείς ομάδες. Ο αλγόριθμος σύνδεσης μέσων τιμών είναι μια προσπάθεια του συμβιβασμού μεταξύ των άκρων των αλγόριθμων απλής και πλήρους σύνδεσης. Ο αλγόριθμος σύνδεσης μέσων τιμών ονομάζεται επίσης (Unweighted Pair Group method using Mathematic Averages – UPGMA) δηλαδή, αλγόριθμος μέσων τιμών χωρίς βάρη και είναι ένας από τους πιο δημοφιλείς αλγόριθμους ομαδοποίησης. Στον αλγόριθμο *σύνδεσης μέσων τιμών* (*average-linkage*) η απόσταση μεταξύ δύο ομάδων υπολογίζεται από την μέση απόσταση των πρότυπων κάθε διαφορετικής ομάδας. Εάν C_i είναι μια ομάδα με n_i μέλη και C_j είναι μια ομάδα με n_j μέλη, τότε η απόσταση μεταξύ δύο ομάδων ορίζεται ως:

$$D_{AL}\left(C_i, C_j\right) = \frac{1}{n_i n_j} \sum_{\boldsymbol{x} \in C_i, \boldsymbol{y} \in C_j} d\left(\boldsymbol{x}, \boldsymbol{y}\right), \tag{5.5}$$

Ο αλγόριθμος σύνδεσης μέσων τιμών θα παρουσιαστεί χρησιμοποιώντας το παράδειγμα του Σχήματος 5.4. Αρχικά υπολογίζεται ο πίνακας απόστασης (5.3) και δημιουργούνται πέντε ομάδες κάθε μια από τις οποίες αποτελείται από ένα πρότυπο. Οι πλησιέστερες ομάδες $\{1\}$ και $\{2\}$ ενώνονται. Έτσι το αποτέλεσμα είναι η δημιουργία των ομάδων:

$$\{1,2\}, \{3\}, \{4\}, \{5\}$$

Κατόπιν, υπολογίζεται ο παρακάτω νέος πίνακας απόστασης των τεσσάρων ομάδων:

	$\{1,2\}$	3	4	5
$\{1,2\}$	0.0	9.9	18.0	19.7
3	9.9	0.0	9.8	9.8
4	18.0	9.8	0.0	8.0
5	19.7	9.8	8.0	0.0

Στον αλγόριθμο σύνδεσης μέσων τιμών η απόσταση μεταξύ των ομάδων είναι ο μέσος όρος των αποστάσεων. Στον αρχικό πίνακα απόστασης (5.3) φαίνεται ότι $d(1,3) = 11.7$ και $d(2,3) = 8.1$. Ο μέσος όρος αυτών των τιμών είναι 9.9 και παρουσιάζεται στον παραπάνω πίνακα ως η απόσταση μεταξύ των ομάδων $\{1,2\}$ και $\{3\}$. Οι υπόλοιπες τιμές της πρώτης στήλης του παραπάνω πίνακα υπολογίζονται παρόμοια. Οι υπόλοιπες στήλες και σειρές απλά διατηρούνται και αντιγράφονται από τον αρχικό πίνακα απόστασης (5.3).

Επαναλαμβάνεται η διαδικασία της εύρεσης τις μικρότερης απόστασης μεταξύ των ομάδων η οποία είναι 8.0 και είναι η απόσταση μεταξύ των ομάδων $\{4\}$ και $\{5\}$ οι οποίες ενώνονται. Σε αυτό το σημείο υπάρχουν οι παρακάτω τρεις ομάδες:

$$\{1,2\}, \{3\}, \{4,5\}$$

Υπολογίζεται ο παρακάτω νέος πίνακας απόστασης των τριών ομάδων:

	$\{1,2\}$	3	$\{4,5\}$
$\{1,2\}$	0.0	9.9	18.9
3	9.9	0.0	9.8
$\{4,5\}$	18.9	9.8	0.0

Εφόσον η ελάχιστη τιμή του πίνακα είναι 9.8, οι ομάδες $\{3\}$ και $\{4,5\}$ ενώνονται. Σε αυτό το σημείο υπάρχουν οι παρακάτω δύο ομάδες:

$$\{1,2\}, \{3,4,5\}$$

Το τελευταίο βήμα είναι η ένωση των δύο ομάδων. Η ιεραρχική ομαδοποίηση του αλγόριθμου σύνδεσης μεσών τιμών έχει ολοκληρωθεί. Το αποτέλεσμα είναι ακριβώς το ίδιο με τον αλγόριθμο πλήρης σύνδεσης.

Ο αλγόριθμος σύνδεσης μεσών τιμών είναι πολύ αποτελεσματικός όταν τα πρότυπα στο χώρο είναι συμπαγή. Επιπλέον, μπορεί να χρησιμοποιηθεί και σε επιμήκεις ομάδες.

5.10. Η μέθοδος του Ward

Η μέθοδος του Ward ονομάζεται επίσης και μέθοδος ελάχιστης διακύμανσης (minimum-variance). Η μέθοδος του Ward είναι διαφορετική από τους υπόλοιπους αλγορίθμους διότι χρησιμοποιεί τεχνικές ανάλυσης διακύμανσης για τον υπολογισμό των αποστάσεων μεταξύ των ομάδων. Όπως και στους άλλους αλγόριθμους, στη μέθοδο του Ward αρχικά όλα τα πρότυπα γίνονται ατομικές ομάδες. Σε κάθε επανάληψη του αλγόριθμου, μεταξύ όλων των ζευγαριών των ομάδων, συνενώνεται το ζευγάρι το οποίο παράγει το μικρότερο *τετραγωνικό σφάλμα*. Εάν μια ομάδα αποτελείται από m πρότυπα $\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_m}$, όπου $\boldsymbol{x_i}$ είναι το διάνυσμα χαρακτηριστικών γνωρισμάτων $[x_{i1}, x_{i2}, \ldots, x_{in}]$. Το *τετραγωνικό σφάλμα* του προτύπου $\boldsymbol{x_i}$ είναι η τετραγωνική Ευκλείδεια απόσταση του από τον μέσο όρο της ομάδας και ορίζεται ως:

$$\sum_{j=1}^{n} (x_{ij} - \mu_j)^2,$$

όπου μ_j είναι ο μέσος όρος του χαρακτηριστικού γνωρίσματος j για κάθε πρότυπο που ανήκει στην ομάδα και ορίζεται ως:

$$\mu_j = \frac{1}{m} \sum_{i=1}^{m} x_{ij},$$

Το τετραγωνικό σφάλμα E όλης της ομάδας είναι το σύνολο των τετραγωνικών σφαλμάτων όλων των προτύπων που ανήκουν στην ομάδα και δίνεται από την εξίσωση:

$$E = \sum_{i=1}^{m} \sum_{j=1}^{n} (x_{ij} - \mu_j)^2 = m\sigma^2.$$

Το διάνυσμα του οποίου τα στοιχεία είναι ο μέσος όρος κάθε προτύπου της ομάδας, $[\mu_1, \mu_2, \ldots, \mu_n] = \boldsymbol{\mu}$, ονομάζεται το *μέσο διάνυσμα* ή το

Πίνακας 5.1: *Τετραγωνικά σφάλματα E για κάθε τρόπο δημιουργίας τεσσάρων ομάδων.*

Ομάδες	E
$\{1,2\},\{3\},\{4\},\{5\}$	8.0
$\{1,3\},\{2\},\{4\},\{5\}$	68.5
$\{1,4\},\{2\},\{3\},\{5\}$	200.0
$\{1,5\},\{2\},\{3\},\{4\}$	232.0
$\{2,3\},\{1\},\{4\},\{5\}$	32.5
$\{2,4\},\{1\},\{3\},\{5\}$	128.0
$\{2,5\},\{1\},\{3\},\{4\}$	160.0
$\{3,4\},\{1\},\{2\},\{5\}$	48.5
$\{3,5\},\{1\},\{2\},\{4\}$	48.5
$\{4,5\},\{1\},\{2\},\{3\}$	32.0

κέντρο βάρους της ομάδας. Το τετραγωνικό σφάλμα μιας ομάδας είναι η *ολική διακύμανση* της ομάδας σ^2 επί του αριθμού των προτύπων στην ομάδα m. Η ολική διακύμανση ορίζεται ως: $\sigma^2 = \sigma_1^2 + \ldots + \sigma_m^2$, δηλαδή το σύνολο των διακυμάνσεων κάθε προτύπου. Παρόμοια το τετραγωνικό σφάλμα για ένα σύνολο ομάδων ορίζεται ως το σύνολο των τετραγωνικών σφαλμάτων κάθε ομάδας.

Η μέθοδος του Ward θα παρουσιαστεί χρησιμοποιώντας το παράδειγμα του Σχήματος 5.4. Αρχικά δημιουργούνται 5 ομάδες κάθε μια από τις οποίες αποτελείται από ένα πρότυπο. Σε αυτό το σημείο το τετραγωνικό σφάλμα είναι μηδέν. Υπάρχουν 10 διαφορετική τρόποι για την συνένωση δυο ομάδων, όταν το σύνολο των ομάδων είναι 5. Ο Πίνακας 5.1 παρουσιάζει το σφάλμα για κάθε ένα από τους 10 διαφορετικούς τρόπους. Για παράδειγμα, έστω ότι εξετάζεται η συνένωση των ομάδων $\{1\}$ και $\{2\}$. Εφόσον το διάνυσμα του προτύπου $\{1\}$ είναι $[4,4]$

Πίνακας 5.2: *Τετραγωνικά σφάλματα E για τρεις ομάδες.*

Ομάδες	E
$\{1,2,3\},\{4\},\{5\}$	72.7
$\{1,2,4\},\{3\},\{5\}$	224.0
$\{1,2,5\},\{3\},\{4\}$	266.7
$\{1,2\},\{3,4\},\{5\}$	56.5
$\{1,2\},\{3,5\},\{4\}$	56.5
$\{1,2\},\{4,5\},\{3\}$	40.0

και το διάνυσμα του πρότυπου $\{2\}$ είναι $[8,4]$, ο μέσος όρος των χαρακτηριστικών γνωρισμάτων είναι: $\mu_1 = 6$ και $\mu_2 = 4$ και το μέσο διάνυσμα είναι $[6,4]$. Το τετραγωνικό σφάλμα για την ομάδα $\{1,2\}$ είναι:

$$(4-6)^2 + (8-6)^2 + (4-4)^2 + (4-4)^2 = 8.$$

Το τετραγωνικό σφάλμα για κάθε μια από τις υπόλοιπες ομάδες $\{3\}$, $\{4\}$ και $\{5\}$ είναι 0. Έτσι, το ολικό τετραγωνικό σφάλμα για τις ομάδες $\{1,2\}$, $\{3\}$, $\{4\}$ και $\{5\}$ είναι:

$$8 + 0 + 0 + 0 = 8.$$

Παρόμοια υπολογίζεται το ολικό τετραγωνικό σφάλμα για τις υπόλοιπες 9 περιπτώσεις και τα αποτελέσματα παρουσιάζονται στον Πίνακα 5.1. Εφόσον το μικρότερο ολικό τετραγωνικό σφάλμα στον Πίνακα 5.1 είναι 8, οι ομάδες $\{1\}$ και $\{2\}$ ενώνονται. Έτσι το αποτέλεσμα είναι η δημιουργία των ομάδων:

$$\{1,2\},\{3\},\{4\},\{5\}$$

Πίνακας 5.3: *Τετραγωνικά σφάλματα E για δύο ομάδες.*

Ομάδες	E
$\{1,2,3\},\{4,5\}$	104.7
$\{1,2,4,5\},\{3\}$	380.0
$\{1,2\},\{3,4,5\}$	94.0

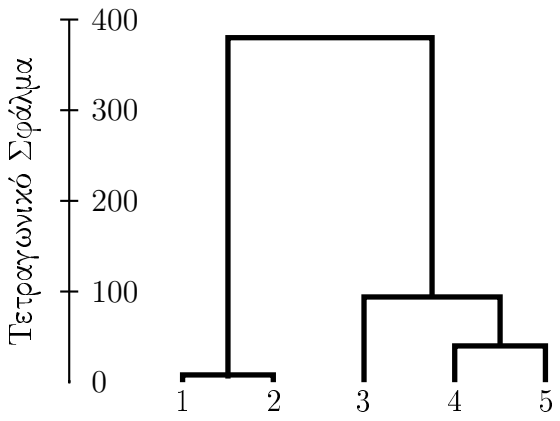
Ο Πίνακας 5.2 παρουσιάζει το τετραγωνικό σφάλμα για όλες τις πιθανές περιπτώσεις συνενώσεις ομάδων για την δημιουργία 3 ομάδων. Εφόσον το μικρότερο τετραγωνικό σφάλμα στον Πίνακα 5.2 είναι 40, οι ομάδες $\{4\}$ και $\{5\}$ συνενώνονται για να δημιουργήσουν τις ομάδες:

$$\{1,2\},\{3\},\{4,5\}.$$

Ο Πίνακας 5.3 παρουσιάζει το τετραγωνικό σφάλμα για όλες τις πιθανές περιπτώσεις συνενώσεις των 3 ομάδων για την δημιουργία 2 ομάδων. Εφόσον το μικρότερο τετραγωνικό σφάλμα στον Πίνακα 5.3 είναι 94, οι ομάδες $\{3\}$ και $\{4,5\}$ συνενώνονται για να δημιουργηθούν οι δύο ομάδες:

$$\{1,2\},\{3,4,5\}.$$

Το τελευταίο βήμα είναι η ένωση των δύο ομάδων. Η ιεραρχική ομαδοποίηση με τη μέθοδο του Ward έχει ολοκληρωθεί και το τελικό δενδρόγραμμα παρουσιάζεται στο Σχήμα 5.8. Γενικά, η μέθοδος του Ward θεωρείται πολύ αποτελεσματική, όμως έχει την τάση δημιουργίας ομάδων μικρού μεγέθους. Ένα από τα μειονεκτήματα όλων των ιεραρχικών αλγόριθμων είναι ότι αποτελεσματικά μπορούν να χρησιμοποιηθούν μόνο για σχετικά μικρό αριθμό προτύπων (< 20) και ειδικά η μέθοδος του Ward.



Σχήμα 5.8: *Ιεραρχική ομαδοποίηση χρησιμοποιώντας τον αλγόριθμο απλής σύνδεσης.*

5.11. Ασκήσεις

5.1 Για τα πρότυπα:

$$[2,1],[2,2],[3,2],[3,1],[-2,-1],[-2,-2],[-3,-2]$$

να εφαρμοστεί ο αλγόριθμος απλής σύνδεσης χρησιμοποιώντας:

α) Ιπποδάμεια μετρική.

β) Ευκλείδια μετρική.

Να υπολογιστούν όλοι οι Πίνακες απόστασης καθώς και να σχεδιαστεί το τελικό δενδρόγραμμα με τις αποστάσεις των ομάδων στον κάθετο άξονα. Πόσες "φυσικές" ομάδες δημιουργούνται;

5.2 Για τα πρότυπα:

$$[1,-1],[1,-2],[2,-2],[-2,2],[-3,2]$$

να εφαρμοστεί ο αλγόριθμος απλής σύνδεσης χρησιμοποιώντας:

α) Ιπποδάμεια μετρική.

β) Ευκλείδια μετρική.

Να υπολογιστούν όλοι οι Πίνακες απόστασης καθώς και να σχεδιαστεί το τελικό δενδρόγραμμα με τις αποστάσεις των ομάδων στον κάθετο άξονα. Πόσες "φυσικές" ομάδες δημιουργούνται;

5.3 Για τα πρότυπα:

$$[2,1],[2,2],[3,2],[3,1],[-2,-1],[-2,-2],[-3,-2]$$

να εφαρμοστεί ο αλγόριθμος πλήρους σύνδεσης χρησιμοποιώντας:

α) Ιπποδάμεια μετρική.

β) Ευκλείδια μετρική.

Να υπολογιστούν όλοι οι Πίνακες απόστασης καθώς και να σχεδιαστεί το τελικό δενδρόγραμμα με τις αποστάσεις των ομάδων στον κάθετο άξονα. Πόσες "φυσικές" ομάδες δημιουργούνται;

5.4 Για τα πρότυπα:

$$[1,-1],[1,-2],[2,-2],[-2,2],[-3,2]$$

να εφαρμοστεί ο αλγόριθμος πλήρους σύνδεσης χρησιμοποιώντας:

α) Ιπποδάμεια μετρική.

β) Ευκλείδια μετρική.

Να υπολογιστούν όλοι οι Πίνακες απόστασης καθώς και να σχεδιαστεί το τελικό δενδρόγραμμα με τις αποστάσεις των ομάδων στον κάθετο άξονα. Πόσες "φυσικές" ομάδες δημιουργούνται;

5.5 Για τα πρότυπα:

$$[2,1],[2,2],[3,2],[3,1],[-2,-1],[-2,-2],[-3,-2]$$

να εφαρμοστεί ο αλγόριθμος σύνδεσης μέσων τιμών χρησιμοποιώντας:

α) Ιπποδάμεια μετρική.

β) Ευκλείδια μετρική.

Να υπολογιστούν όλοι οι Πίνακες απόστασης καθώς και να σχεδιαστεί το τελικό δενδρόγραμμα με τις αποστάσεις των ομάδων στον κάθετο άξονα. Πόσες "φυσικές" ομάδες δημιουργούνται;

5.6 Για τα πρότυπα:

$$[1,-1],[1,-2],[2,-2],[-2,2],[-3,2]$$

να εφαρμοστεί ο αλγόριθμος σύνδεσης μέσων τιμών χρησιμοποιώντας:

α) Ιπποδάμεια μετρική.

β) Ευκλείδια μετρική.

Να υπολογιστούν όλοι οι Πίνακες απόστασης καθώς και να σχεδιαστεί το τελικό δενδρόγραμμα με τις αποστάσεις των ομάδων στον κάθετο άξονα. Πόσες "φυσικές" ομάδες δημιουργούνται;

5.7 Για τα πρότυπα:

$$[2,1],[2,2],[3,2],[3,1],[-2,-1],[-2,-2],[-3,-2]$$

να εφαρμοστεί η μέθοδος του Ward χρησιμοποιώντας:

α) Ιπποδάμεια μετρική.

β) Ευκλείδια μετρική.

Να υπολογιστούν όλοι οι Πίνακες απόστασης καθώς και να σχεδιαστεί το τελικό δενδρόγραμμα με τις αποστάσεις των ομάδων στον κάθετο άξονα. Πόσες "φυσικές" ομάδες δημιουργούνται;

5.8 Για τα πρότυπα:

$$[1,-1],[1,-2],[2,-2],[-2,2],[-3,2]$$

να εφαρμοστεί η μέθοδος του Ward χρησιμοποιώντας:

α) Ιπποδάμεια μετρική.

β) Ευκλείδια μετρική.

Να υπολογιστούν όλοι οι Πίνακες απόστασης καθώς και να σχεδιαστεί το τελικό δενδρόγραμμα με τις αποστάσεις των ομάδων στον κάθετο άξονα. Πόσες "φυσικές" ομάδες δημιουργούνται;

Κεφάλαιο 6

Διαχωριστικη Ομαδοποιηση

6.1. Εισαγωγή

Οι τεχνικές διαχωριστικής ομαδοποίησης διαχωρίζουν ένα σύνολο N προτύπων και δημιουργούν k ομάδες, όπου $k << N$. Ο επιθυμητός αριθμός των ομάδων k συνήθως προκαθορίζεται από τον χρήστη. Αντίθετα με τις ιεραρχικές τεχνικές, οι διαχωριστικές τεχνικές επιτρέπουν τα πρότυπα να μετακινούνται από μια ομάδα σε άλλη. Με αυτό τον τρόπο, ένας κακός αρχικός διαχωρισμός μπορεί να διορθωθεί αργότερα. Ο τρόπος με τον οποίο γίνεται ο διαχωρισμός στις περισσότερες τεχνικές είναι με την μεγιστοποίηση κάποιας συνάρτησης κριτηρίων. Όμως η εύρεση του

βέλτιστου διαχωρισμού δεν είναι μαθηματικά εφικτή. Για παράδειγμα, υπάρχουν 1.93×10^8 διαφορετικοί διαχωρισμοί 19 προτύπων σε 3 ομάδες. Έτσι, οι περισσότερες διαχωριστικές τεχνικές χρησιμοποιούν ορισμένες *ευριστικές* (*heuristic*) μεθόδους ώστε να βρεθεί μια ικανοποιητική λύση. Το αποτέλεσμα όμως είναι η δημιουργία υποβελτιστοποιημένων διαχωρισμών. Μεταξύ όλων των κριτηρίων τα οποία χρησιμοποιούνται για την δημιουργία ενός διαχωρισμού, το κριτήριο του τετραγωνικού σφάλματος είναι το πιο δημοφιλές. Βέβαια, η χρησιμοποίηση αυτού του κριτηρίου έχει σαν αποτέλεσμα την δημιουργία ομάδων υπερελλειψοειδούς μορφής. Επίσης, στις τεχνικές διαχωριστικής ομαδοποίησης χρησιμοποιείται συνήθως η Ευκλείδεια μετρική για τον υπολογισμό αποστάσεων.

Αναλυτικά θα παρουσιαστούν τρεις ευρέως διαδομένες τεχνικές διαχωριστικής ομαδοποίησης: ο αλγόριθμος Forgy, ο αλγόριθμος k-means και ο αλγόριθμος Isodata.

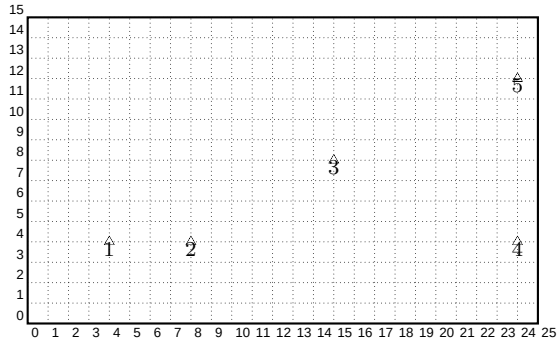
6.2. Ο αλγόριθμος Forgy

Ένας από τους πιο απλούς διαχωριστικούς αλγόριθμους ομαδοποίησης είναι ο αλγόριθμος του Forgy [Forgy65]. Εκτός από τα πρότυπα για ομαδοποίηση, επιπλέον είσοδοι στον αλγόριθμο είναι ο αριθμός των επιθυμητών ομάδων k που θα δημιουργηθούν και τα k αντιπροσωπευτικά πρότυπα των ομάδων. Η αρχική επιλογή των αντιπροσωπευτικών προτύπων μπορεί να γίνει τυχαία από τα δεδομένα ή εφόσον είναι γνωστή η επιθυμητή δομή των ομάδων τότε η γνώση αυτή μπορεί να χρησιμοποιηθεί για την καθοδήγηση της αρχικής επιλογής. Ο ψευδοκώδικας του αλγορίθμου Forgy είναι:

(1) Αρχικός καθορισμός των k ομάδων και των k αντιπροσωπευτικών προτύπων.

(2) Για κάθε πρότυπο να βρεθεί το πλησιέστερο αντιπροσωπευτικό πρότυπο και να καταχωρηθεί στην αντίστοιχη ομάδα.

	x_1	x_2
1	4.0	4.0
2	8.0	4.0
3	15.0	8.0
4	24.0	4.0
5	24.0	12.0



Σχήμα 6.1: *Δεδομένα για ανάλυση διαχωριστικής ομαδοποίησης.*

(3) Να υπολογιστεί το νέο αντιπροσωπευτικό πρότυπο για κάθε ομάδα, το οποίο είναι ο μέσος όρος των προτύπων της ομάδας.

(4) Εάν υπάρχουν αλλαγές στις καταχωρήσεις των προτύπων σε ομάδες, δηλαδή τα νέα αντιπροσωπευτικά πρότυπα δεν είναι ίδια με τα προηγούμενα, επιστροφή στο 2.

Ο αλγόριθμος Forgy θα παρουσιαστεί χρησιμοποιώντας το παράδειγμα του Σχήματος 6.1. Αρχικά, καθορίζεται $k = 2$ ώστε να δημιουργηθούν δύο ομάδες και θα χρησιμοποιηθούν τα δύο πρώτα πρότυπα $[4, 4]$ και $[8, 4]$ σαν τα αρχικά αντιπροσωπευτικά πρότυπα των ομάδων. Στην εκτέλεση του δεύτερου βήματος υπολογίζεται η πλησιέστερη ομάδα για κάθε πρότυπο.

Ο Πίνακας 6.1 παρουσιάζει την αρχική καταχώρηση του αλγόριθμος Forgy. Δημιουργούνται οι ομάδες: {[4,4]} και {[8,4], [15,8], [24,4], [24,12]} Στην εκτέλεση του τρίτου βήματος υπολογίζονται τα νέα αντιπροσωπευτικά πρότυπα των ομάδων. Το αντιπροσωπευτικό πρότυπο της πρώτης ομάδας είναι: $[4, 4]$. Το αντιπροσωπευτικό πρότυπο της δεύτερης ομάδας υπολογίζεται από τον μέσο όρο των προτύπων τα οποία

Πίνακας 6.1: *Αρχική καταχώρηση του αλγόριθμου Forgy.*

Πρότυπο	Πλησιέστερο Αντιπροσωπευτικό Πρότυπο
$[4, 4]$	[4,4]
$[8, 4]$	[8,4]
$[15, 8]$	[8,4]
$[24, 4]$	[8,4]
$[24, 12]$	[8,4]

ανήκουν στην ομάδα, δηλαδή:

$$\frac{8 + 15 + 24 + 24}{4} = 17.75$$
$$\frac{4 + 8 + 4 + 12}{4} = 7$$

Έτσι το νέο αντιπροσωπευτικό πρότυπο είναι: $[17.75, 7]$. Εφόσον υπάρχει αλλαγή στην καταχώρηση των προτύπων δηλαδή τα νέα αντιπροσωπευτικά πρότυπα δεν είναι ίδια με τα αρχικά, επιστροφή στο δεύτερο βήμα του αλγορίθμού.

Στην δεύτερη επανάληψη του αλγορίθμου για κάθε πρότυπο βρίσκεται το πλησιέστερο αντιπροσωπευτικό πρότυπο. Ο Πίνακας 6.2 παρουσιάζει τα αποτελέσματα. Δημιουργούνται οι ομάδες:

$$C_1 = \{[4, 4], [8, 4]\}$$
$$C_2 = \{[15, 8], [24, 4], [24, 12]\}$$

Στην εκτέλεση του τρίτου βήματος της δεύτερης επανάληψης υπολογίζονται τα νέα αντιπροσωπευτικά πρότυπα των ομάδων τα οποία είναι: $[6, 4]$ και $[21, 8]$. Εφόσον υπάρχει αλλαγή στην καταχώρηση των προτύπων επιστροφή στο δεύτερο βήμα του αλγορίθμού όπου για κάθε πρότυπο βρίσκεται το πλησιέστερο αντιπροσωπευτικό πρότυπο.

Πίνακας 6.2: *Δεύτερη καταχώρηση του αλγόριθμου Forgy.*

Πρότυπο	Πλησιέστερο Αντιπροσωπευτικό Πρότυπο
[4, 4]	[4,4]
[8, 4]	[4,4]
[15, 8]	[17.75,7]
[24, 4]	[17.75,7]
[24, 12]	[17.75,7]

Πίνακας 6.3: *Τρίτη καταχώρηση του αλγόριθμου Forgy.*

Πρότυπο	Πλησιέστερο Αντιπροσωπευτικό Πρότυπο
[4, 4]	[6,4]
[8, 4]	[6,4]
[15, 8]	[21,8]
[24, 4]	[21,8]
[24, 12]	[21,8]

Ο Πίνακας 6.3 παρουσιάζει τα αποτελέσματα και όπως φαίνεται κανένα πρότυπο δεν αλλάζει ομάδα. Στην εκτέλεση του τρίτου βήματος υπολογίζονται τα νέα αντιπροσωπευτικά πρότυπα των ομάδων τα οποία παραμένουν σταθερά. Εφόσον κανένα πρότυπο δεν αλλάζει ομάδα ο αλγόριθμος τερματίζεται.

Σε αυτή την έκδοση του αλγόριθμου Forgy η επιλογή των αντιπροσωπευτικών προτύπων είναι τυχαία. Έχουν όμως προταθεί εναλλακτικές

μεθοδολογίες επιλογής των αντιπροσωπευτικών προτύπων. Μια εναλλακτική πρόταση είναι η χρησιμοποίηση ιεραρχικών τεχνικών ομαδοποίησης για την παραγωγή των k αρχικών αντιπροσωπευτικών προτύπων.

Ποιό πολύπλοκες εκδόσεις του αλγόριθμου Forgy επιτρέπουν στο χρήστη να εισάγει παραμέτρους οι οποίοι επηρεάζουν δυναμικά την δημιουργία ομάδων. Σε μια τέτοια έκδοση [Forgy65], όταν έχουν δημιουργηθεί και σταθεροποιηθεί οι ομάδες, μια νέα ομάδα μπορεί να δημιουργηθεί ή μια υπάρχουσα ομάδα να διαγραφτεί. Αρχικά υπολογίζεται ή μέση απόσταση ενός προτύπου από όλα τα αντιπροσωπευτικά πρότυπα των ομάδων. Όταν μια παράμετρος T_1, η οποία δίνεται από τον χρήστη και παίρνει τιμές: $0 \leq T_1 \leq 1$, αυξάνεται τότε δημιουργούνται περισσότερες ομάδες. Επίσης μια υπάρχουσα ομάδα μπορεί να διαγραφτεί εάν υπάρχουν λιγότερα από T_2 πρότυπα μετά την σύγκλιση του αρχικού αλγορίθμου. Τα πρότυπα αυτά θεωρούνται εξαιρέσεις ή πρότυπα με θόρυβο και δεν λαμβάνονται υπόψη στην επεξεργασία.

Έχει αποδειχθεί [SeIs84] ότι ο αλγόριθμος του Forgy συγκλίνει και τερματίζεται, δηλαδή τελικά δεν υπάρχουν πρότυπα τα οποία να αλλάζουν ομάδα. Όμως, εάν ο αριθμός των προτύπων είναι μεγάλος μπορεί η εκτέλεση του αλγορίθμου να είναι χρονοβόρα διότι απαιτούνται πολλές επαναλήψεις για την δημιουργία σταθερών ομάδων. Για αυτό τον λόγο, ορισμένες εκδόσεις του αλγόριθμου Forgy δίνουν την δυνατότητα στον χρήστη να καθορίζει τον αριθμό των επαναλήψεων. Κατόπιν, ο αλγόριθμος τερματίζεται.

Βιβλιογραφία

[Aris96] Aristotle, Waterfield, R. and Bostock, D., *Physics*, Oxford University Press, Oxford, UK, (1996).

[Bellm56] Bellman, R.E., *Dynamic Programming*, Princeton University Press, Princeton, NJ, (1957).

[Bloom91] Bloom, A., *The Republic of Plato*, Basic Books, New York, NY, (1991).

[Diday74] Diday, E., "Recent progreess in distance and similarity measures in pattern recognition", *Second International Joint Conferenxe on Pattern Recobnition*, pp. 534–539, Copenhagen, (1964).

[DuHa73] Duda, R.O. and Hart, P.E., *Classification and Scene Analysis*,Wiley, New York, NY, (1973).

[Ever81] Everitt B., *Cluster Analysis*, Halsted Press, John Wiley and Sons, New York, NY, (1981).

[Fisch87] Fischler M. and Firschein O., *Reading in Computer Vision: Issues, Problems, Principles and Paradigms*, Morgan Kaufmann, San Mateo, CA, (1987).

[Fisher36] Fisher, R.A., "The use of multiple measurements in taxonomy problems", *Annals of Eugenics*, 7 Part II:179–188, (1936).

[Forgy65] Forgy, E., "Cluster analysis of multivariate data: efficiency versus interpretability of classifications", *Biometrics*,(1965).

[Fuk90] Fukunaga, K., *Introduction to Stastical Pattern Recognition*, Academic Press, New York, NY, (1990).

[Gose96] Gose, E., Johnsonbaugh, R., and Jost, S., *Pattern Recognition and Image Analysis*, Prentice Hall, Upper Saddle River, NJ, (1996).

[Hart75] Hartigan, J.A., *Clustering Algorithms*, Wiley, New York, NY, (1975).

[Hertz91] Hertz, J., Krogh A. and Palmer R., *Introduction to the Theory of Neural Computation*, Addison-Wesley, Redwood City, CA, (1991).

[Luger94] Luger, G.F., *Cognitive Science: The Science of Intelligent Systems*, Academic Press, New York, NY, (1294).

[NaSm93] Nadler, M. and Smith, E.P., (Editors), *Pattern Recognition Engineering*, John Wiley, New York, NY, (1993).

[Rab93] Rabiner, L. and Juang, B.H., *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, NJ, (1393).

[Sebes62] Sebestyen, G., *Decision Making Process in Pattern Recognition*, Macmillan, Nlw York, NY,(1962).

[SeIs84] Selim, S.Z. and Ismail , M.A., "K-means type algorithms: a generalized convergence theorem and characterization of local optimality", *IEEE Transactions on Pattern Analysis and Machine Intelligence*,Vol.6, pp. 81–87,(1984).

[Shav90] Shavlik, J.W. and Dietterich, K., (Editors), *Readings in Machine Learning*, Morgan Kaufmann, San Mateo, CA, (1990).

[Tan58] Tanimoto, T., "An elementary mathematical theory of classification and prediction", *Int'l Rpt.*, IBM Corp., (1958).

[Theo99] Theodoridis, S. and Koutroumbas K., *Pattern Recognition*, Academic Press, London, UK, (1999).

[TouGo74] Tou J.T. and Gonzalez R.C., *Pattern Recognition Principles* , Addison-Wesley, Redwood City, CA, (1974).

[Uttal73] Uttal, W.R., *The Psychology of Sensory Coding*, Harper Collins, New York, NY, (3973).

[YoFu86] Young T.Y. and Fu K.S., (Editors), *Handbook of Pattern Recognition and Image Processing*, Academic Press, London, UK,(1986).

Περιεχόμενα

Κεφάλαιο 1. Βασικές Αρχές της Αναγνώρισης Προτύπων 1
1.1. Ορισμός 1
1.2. Εφαρμογές 4
1.3. Μεθοδολογίες αναγνώρισης προτύπων 6
1.4. Ιστορική αναδρομή 9

Κεφάλαιο 2. Εισαγωγή στη Στατιστική Αναγνώριση Προτύπων 11
2.1. Σύστημα Αναγνώρισης Προτύπων 11
2.2. Εφαρμογή: Βιομηχανικό Robot 21
2.3. Ερευνητικά Θέματα Αναγνώρισης Προτύπων 26

Κεφάλαιο 3. Ταξινόμηση προτύπων με συναρτήσεις απόφασης 33
3.1. Απλές Γραμμικές Συναρτήσεις Απόφασης 33
3.2. Γραμμικές Συναρτήσεις Απόφασης 35
3.3. Γενικευμένες Συναρτήσεις Απόφασης 40
3.4. Τμηματικός Γραμμικός Διαχωρισμός. 43
3.5. Γεωμετρικές ιδιότητες υπερεπιπέδων. 44

Κεφάλαιο 4. Ταξινόμηση προτύπων με συναρτήσεις απόστασης 49
4.1. Απλοί ταξινομητές 49
4.2. Ταξινομητές Ελάχιστης Απόστασης 51
4.3. Μέτρα απόστασης 55
4.4. Μέτρα ομοιότητας 61
4.5. Ταίριασμα με υποδείγματα 64
4.6. Σύστημα Αναγνώρισης Δορυφορικής Εικόνας 66
4.7. Ασκήσεις 70

Κεφάλαιο 5. Ομάδες 73
5.1. Δημιουργία ομάδων 73
5.2. Εφαρμογές 76
5.3. Ιδανικές ομάδες 77
5.4. Μεθοδολογίες ομαδοποίησης 78
5.5. Ιεραρχική ομαδοποίηση 79
5.6. Αλγόριθμος απλής σύνδεσης 81
5.7. Αλγόριθμος πλήρους σύνδεσης 85
5.8. Σύγκριση απλής και πλήρους σύνδεσης 87
5.9. Αλγόριθμος σύνδεσης μέσων τιμών 91
5.10. Η μέθοδος του Ward 93
5.11. Ασκήσεις 97

Κεφάλαιο 6. Διαχωριστική Ομαδοποίηση 101
6.1. Εισαγωγή 101
6.2. Ο αλγόριθμος Forgy 102

Βιβλιογραφία 107

Κατάλογος Σχημάτων 111

Κατάλογος Πινάκων 115

Κατάλογος Σχημάτων

1.1	Το πρόβλημα της αναγνώρισης προτύπων.	3
1.2	Το πρόβλημα της συντακτικής αναγνώρισης προτύπων. Πως περιγράφεται ένα δένδρο με συντακτικούς κανόνες;	8
2.1	Αυτόματο σύστημα αναγνώρισης προτύπων.	13
2.2	Δειγματοληψία διανύσματος προτύπου.	13
2.3	Διαχωρίσιμες κατηγορίες.	14
2.4	Μη διαχωρίσιμες κατηγορίες.	15
2.5	Διαχωρίσιμες κατηγορίες στον τρισδιάστατο χώρο.	16
2.6	Δύο στατιστικές κατανομές προτύπων.	19
2.7	Δύο στατιστικές κατανομές σε τρισδιάστατη μορφή.	19
2.8	Λειτουργικό διάγραμμα βιομηχανικού robot.	23
2.9	Περιβάλλον της αναγνώρισης ενός εξαρτήματος.	23
2.10	Εξαρτήματα μηχανής σε διάφορες γωνίες.	25
2.11	Μερικά χαρακτηριστικά γνωρίσματα για βιομηχανικές εφαρμογές.	25
3.1	Απλή συνάρτηση απόφασης για δύο κατηγορίες.	34
3.2	Γραμμικές συναρτήσεις απόφασης με την πρώτη μεθοδολογία.	37

3.3	Γραμμικές συναρτήσεις απόφασης με την δεύτερη μεθοδολογία.	38
3.4	Γραμμικές συναρτήσεις απόφασης με την τρίτη μεθοδολογία.	39
3.5	Το πρόβλημα του Sebestyen.	40
3.6	Μη γραμμικά διαχωρίσιμες κατηγορίες.	41
3.7	Υπερβολική παραβολή στο πρόβλημα του Sebestyen.	42
3.8	Τμηματικός γραμμικός διαχωρισμός στο πρόβλημα του Sebestyen	44
3.9	Γεωμετρικές ιδιότητες ενός υπερεπιπέδου.	45
4.1	Κατάταξη δια μέσου της εγγύτητας.	50
4.2	Κατηγορίες μη-κατάταξιμες δια μέσου της έννοιας της εγγύτητας.	51
4.3	Υλοποίηση ενός ταξινομητή ελάχιστης απόστασης.	53
4.4	Όρια απόφασης και ισομετρικές για την Ευκλείδεια απόσταση.	54
4.5	Όρια απόφασης για την Ευκλείδεια απόσταση με επικαλυπτόμενες κατηγορίες.	54
4.6	Όρια απόφασης για την Ευκλείδεια απόσταση και τρεις κατηγορίες.	55
4.7	Σύγκριση Ευκλείδειας και Ιπποδάμειας απόστασης.	57
4.8	Ισομετρικές στην Ιπποδάμεια απόσταση.	58
4.9	Ισομετρικές αποστάσεις στην απόσταση Mahalanobis.	60
4.10	Δείκτης απόστασης συνημίτονου.	63
4.11	Ταίριασμα με υποδείγματα.	65
4.12	Ταίριασμα με υποδείγματα ενός χαρακτήρα.	65

4.13	Δισδιάστατο διάγραμμα προτύπων για την αναγνώριση χρήσεις γης από δορυφορικές εικόνες.	67
4.14	Ταξινομητής ελάχιστης απόστασης.	68
4.15	Στατιστική κατανομή των κατηγοριών σε τρισδιάστη μορφή.	69
4.16	Ισομετρική κατανομή των κατηγοριών.	70
4.17	Υποδείγματα και αλλοιωμένοι χαρακτήρες.	72
5.1	Ομάδες.	75
5.2	Δύο ή τέσσερις ομάδες;	76
5.3	Παράδειγμα ιεραρχικής ανάλυσης ομάδων.	80
5.4	Δεδομένα για προβλήματα ιεραρχικής ανάλυσης ομάδων.	82
5.5	Ιεραρχική ομαδοποίηση χρησιμοποιώντας τον αλγόριθμο απλής σύνδεσης.	85
5.6	Ιεραρχική ομαδοποίηση χρησιμοποιώντας τον αλγόριθμο πλήρης σύνδεσης	88
5.7	Παραδείγματα ιεραρχικής ομαδοποίησης. (α) Τρία σύνολα δεδομένων. (β) Αποτελέσματα του αλγόριθμου απλής σύνδεσης. (γ) Αποτελέσματα του αλγόριθμου πλήρης σύνδεσης.	89
5.8	Ιεραρχική ομαδοποίηση χρησιμοποιώντας τον αλγόριθμο απλής σύνδεσης.	97
6.1	Δεδομένα για ανάλυση διαχωριστικής ομαδοποίησης.	103

Κατάλογος Πινάκων

1.1	Εφαρμογές αναγνώρισης προτύπων.	3
4.1	Συνοπτικός πίνακας μέτρων απόστασης.	61
5.1	Τετραγωνικά σφάλματα E για κάθε τρόπο δημιουργίας τεσσάρων ομάδων.	94
5.2	Τετραγωνικά σφάλματα E για τρεις ομάδες.	95
5.3	Τετραγωνικά σφάλματα E για δύο ομάδες.	96
6.1	Αρχική καταχώρηση του αλγόριθμου Forgy.	104
6.2	Δεύτερη καταχώρηση του αλγόριθμου Forgy.	105
6.3	Τρίτη καταχώρηση του αλγόριθμου Forgy.	105