UNSUPERVISED CLUSTERING

ΕΡΓΑΣΙΑ 8Η

ΑΝΑΓΝΩΡΙΣΗ ΠΡΟΤΥΠΩΝ

ΚΑΡΑΠΕΠΕΡΑ ΕΛΠΙΔΑ
57423

K-MEANS

Ένας αλγόριθμος διαχωρισμού όπου κάθε ομάδα (cluster) συνδέεται με ένα centroid (κέντρο). Κάθε σημείο αποδίδεται στην ομάδα στης οποίας το κέντρο βρίσκεται πιο κοντά. Ο αριθμός των ομάδων K στις οποίες χωρίζονται τα σημεία πρέπει να καθοριστεί από πριν. Ο αλγόριθμος έχει ως εξής:

- Επιλογή τυχαία των αρχικών K points ως αρχικά κέντρα.
- Επανάληψη:

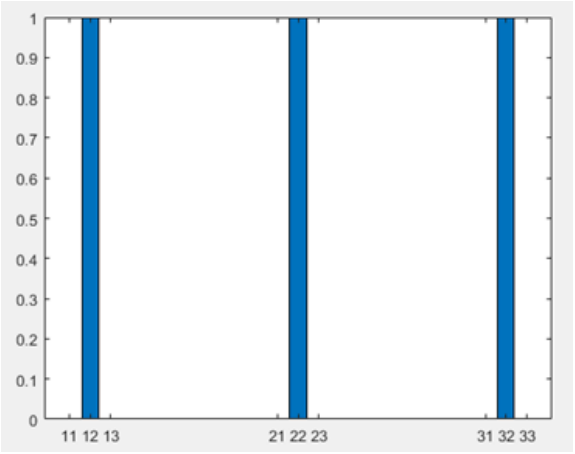
  Δημιουργία K clusters αναθέτοντας κάθε σημείο στο κοντινότερο σε αυτό centroid.

  Επαναϋπολογισμός των centroids κάθε cluster.
- Έως ότου τα centroids σταματήσουν να αλλάζουν για 2 συνεχόμενες επαναλήψεις.

Για κάθε σημείο, το σφάλμα του είναι η απόστασή του από το κοντινότερο σε αυτό κέντρο. Το άθροισμα των τετραγώνων όλων των σφαλμάτων ονομάζεται SSE.

ΕΡΩΤΗΜΑ 1

Τυπώνοντας τα αποτελέσματα των clusters στα οποία ταξινομήθηκαν τα δεδομένα παρατηρείται ότι όλα ταξινομούνται σε ένα μόνο cluster.
Αυτό λογικά συμβαίνει γιατί οι ομάδες δεν έχουν σφαιρικό σχήμα.
Το σφάλμα είναι επομένως 66.67%.



Στην εικόνα αυτή, όπως και στις επόμενες φαίνεται σε κάθε 3-στηλο η κατανομή των data points που ανήκουν σε μία συγκεκριμένη κλάση στις 3 κλάσεις. Έτσι, στις πρώτες 3 στήλες 11,12,13 φαίνεται ότι όλα τα data points που ανήκουν κανονικά στην κλάση 1 ταξινομήθηκαν στην κλάση 2. Το ίδιο συνέβη και στα δεδομένα της κλάσης 2 και της 3.

FUZZY C-MEANS

Μία παραλλαγή του K-means στην οποία κάθε στοιχείο μπορεί να ανήκει σε περισσότερα από ένα clusters. Εισάγεται μία νέα συνάρτηση membership uj.

Η διαδικασία που ακολουθείται έχει ως εξής:

- Ορίζεται ένας αριθμός clusters.
- Τυχαία ανατίθενται coefficients σε κάθε data point για το membership τους σε κάθε cluster.
- Επανάληψη ως ότου για 2 συνεχόμενα iterations η αλλαγή των coefficients δεν είναι μεγαλύτερη από μία ανεκτικότητα ε:

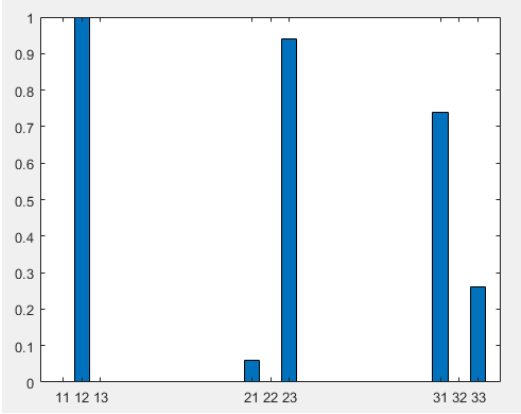
Υπολογισμός του centroid κάθε cluster σύμφωνα με τον διπλανό τύπο, όπου m η hyper-parameter που ορίζει πόσο fuzzy θα είναι το κάθε cluster. Όσο μεγαλύτερη η τιμή του m τόσο πιο fuzzy το cluster.

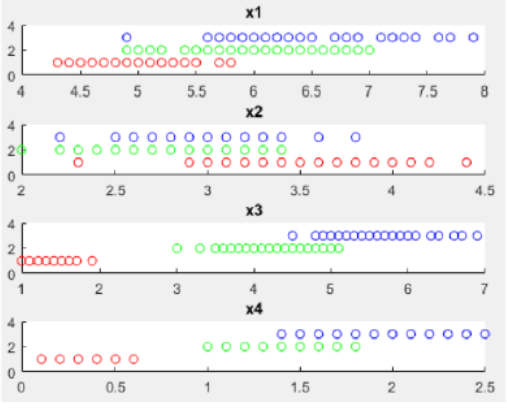
$$c_k = \frac{\sum_x u_k(x)^m x}{\sum u_k(x)^m}$$

Για κάθε σημείο υπολογισμός των coefficients των memberships για κάθε cluster.

ΕΡΩΤΗΜΑ 2

Παρατηρείται ότι το δεύτερο cluster αναγνωρίζεται πλέον με επιτυχία 100%. Τα άλλα 2 clusters, τα οποία όπως παρατηρήθηκε και από προηγούμενες εργασίες ήταν μπλεγμένα μεταξύ τους φαίνεται να είναι πιο δύσκολο να ταξινομηθούν.





Συγκεκριμένα, το δεύτερο cluster αναγνωρίζεται με ακρίβεια 94%, ενώ το τελευταίο 74%.

Αυτό συμβαίνει γιατί το πρώτο cluster είναι γραμμικά διαχωρίσιμο, ενώ τα άλλα δύο όχι. Στα στοιχεία, επομένως, τα οποία επικαλύπτονται από τη 1η και την 3η ομάδα είναι δύσκολος ο διαχωρισμός.

Έτσι, στις πρώτες 3 στήλες όλα τα data points που ανήκουν κανονικά στην κλάση 1 ταξινομήθηκαν σε μία ίδια κλάση, την κλάση 2. Το 94% των data points της 2ης κλάσης ταξινομήθηκαν όλα μαζί στην κλάση 3, ενώ το υπόλοιπο 6% ταξινομήθηκε στην κλάση 1. Τέλος, το 74% των στοιχείων που ανήκουν κανονικά στην κλάση 3 ταξινομήθηκαν όλα μαζί στην κλάση 1, ενώ το υπόλοιπο 26% στην

κλάση 3. Να σημειωθεί εδώ ότι δεν έχει σημασία ο αριθμός της κλάσης, αυτό είναι απλά η ονομασία που της δίνει τυχαία ο αλγόριθμος. Σημασία έχει το κατά πόσο τα στοιχεία της ίδιας κλάσης ταξινομούνται όλα μαζί σε μία ενιαία κλάση, και κατά πόσο κατανομούνται σε διαφορετικές κλάσεις στοιχεία που δεν ανήκουν στην ίδια κλάση.

Φαίνεται ότι, όντως, τα στοιχεία των 3 κλάσεων ταξινομούνται στην πλειοψηφία τους σωστά σε 3 διαφορετικές κλάσεις (2, 3 και 1). Το σφάλμα είναι επομένως 10,67% (μέσος όρος όλων των σφαλμάτων). Η ταξινόμηση είναι σχετικά επιτυχής.

ISODATA

Συντομογραφία του Iterative Self-Organizing Data Analysis Technique Algorithm. Είναι ουσιαστικά ο k-means αλγόριθμος με το επιπλέον χαρακτηριστικό ότι μπορεί αυτόματα να επιλέξει πλέον το πλήθος των κλάσεων.

Οι παράμετροι που επιλέγονται είναι οι εξής:

NMIN_EX -> ελάχιστο πλήθος δειγμάτων ανά cluster.

ND ->επιθυμητό πλήθος clusters.

σ_s^2 -> μέγιστη διασπορά για διαχωρισμό clusters.

DMERGE -> μέγιστη απόσταση για ένωση των clusters.

NMERGE -> μέγιστο πλήθος clusters που μπορούν να ενωθούν.

Η διαδικασία έχει ως εξής:

Επιλέγονται αυθαίρετα τα κέντρα και τα σημεία ανατίθενται στο κοντινότερο σε αυτά cluster.

Το standard deviation κάθε cluster και η απόσταση των clusters μεταξύ τους υπολογίζονται.

Τα clusters χωρίζονται αν ένα ή περισσότερα deviations είναι μεγαλύτερα από το προκαθορισμένο threshold.

Τα clusters ενώνονται αν η απόσταση μεταξύ τους είναι μικρότερη από το προκαθορισμένο threshold.

Εφαρμόζεται περεταίρω iterations των παραπάνω με τα καινούρια πλέον cluster centers έως ότου:

Ο μέσος όρος των inner-center distances είναι μικρότερος από το προκαθορισμένο threshold.

Ο μέσος όρος των αλλαγών στα inner-center distances μεταξύ των iterations είναι μικρότερος από το προκαθορισμένο threshold.

Έχει πραγματοποιηθεί ο μέγιστος επιτρεπόμενος αριθμός iterations.

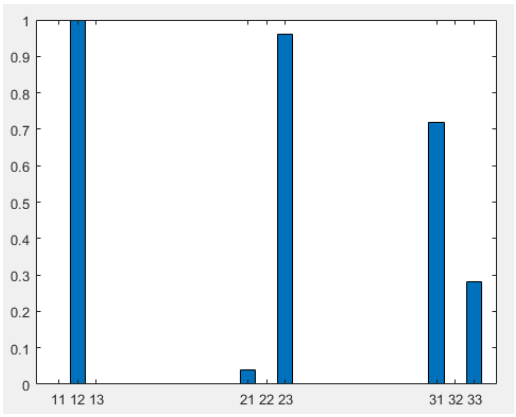
Το ISODATA δίνει στον αλγόριθμο τη δυνατότητα να διαιρεί clusters με ανομοιότητες και να ενώνει clusters με ομοιότητες.

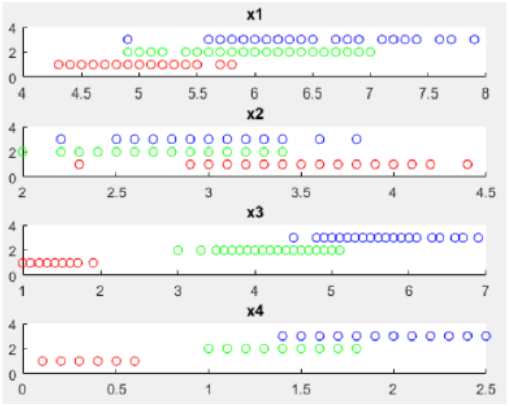
Επιπλέον, μπορεί να καταργεί clusters με ελάχιστα δείγματα και έχει δυνατότητες αυτό-οργάνωσης.

Ωστόσο, εξακολουθεί να ισχύει ο περιορισμός ότι τα δεδομένα πρέπει να είναι γραμμικά διαχωρίσιμα και ο προκαθορισμός των thresholds και των υπόλοιπων παραμέτρων είναι δύσκολος, αλλά ταυτόχρονα καθοριστικός

ΕΡΩΤΗΜΑ 3

Παρατηρείται ότι το πρώτο cluster αναγνωρίζεται πλέον με επιτυχία 100%. Τα άλλα 2 clusters, τα οποία όπως παρατηρήθηκε και από προηγούμενες εργασίες ήταν μπλεγμένα μεταξύ τους φαίνεται να είναι πιο δύσκολο να ταξινομηθούν.





Συγκεκριμένα, το δεύτερο cluster αναγνωρίζεται με ακρίβεια 96%, ενώ το τελευταίο 72%.

Αυτό συμβαίνει γιατί το πρώτο cluster είναι γραμμικά διαχωρίσιμο, ενώ τα άλλα δύο όχι. Στα στοιχεία, επομένως, τα οποία επικαλύπτονται από τη 2η και την 3η ομάδα είναι δύσκολος ο διαχωρισμός. Το συνολικό σφάλμα είναι επομένως 10.66%.

Το αποτέλεσμα αυτό προέκυψε μετά από πολλούς πειραματισμούς όσον αφορά τις παραμέτρους του ISODATA. Συγκεκριμένα, αυτές που έδωσαν το τελικό βέλτιστο αποτέλεσμα ήταν οι παρακάτω:

Οριακός αριθμός datapoints για την εξάλειψη μιας ομάδας: ON = 28

Οριακή απόσταση για την ένωση ομάδων: OC = 0

Τυπικό όριο ακόκλισης για τη διαίρεση μιας ομάδας σε δύο μικρότερες: OS = 0.1

Μέγιστος αριθμός ομάδων: k = 4

Μέγιστος αριθμός ομάδων που μπορούν να ενωθούν σε μία επανάληψη: L = 5

Μέγιστος αριθμός επαναλήψεων: I = 100

Επιπλέον παράμετρος για αυτόματη απάντηση όχι στο αίτημα του cambial οποιασδήποτε παραμέτρου: NO = 1

Ελάχιστη απόσταση που πρέπει να έχει κάθε σημείο από το κοντινότερο κέντρο: min_dist = 3

Φαίνεται ότι η καλύτερη ταξινόμηση γίνεται με τη χρήση του fuzzy c-means.