Ανάλυση Δεδομένων

4η εργασία

Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
Τμήμα Μηχανολόγων Μηχανικών
31/05/2024

Δημήτριος Νεντίδης, 6821

Περιεχόμενα

1 Εισαγωγή

Η ομαδοποίηση (clustering) των δεδομένων διαφέρει αρκετά από την ταξινόμηση που παρουσιάστηκε στις προηγούμενες δύο εργασίες. Ενώ πριν στόχος ήταν η ταξινόμηση των παρατηρήσεων σε κάποια σετ με στόχο την πρόβλεψη της απόκρισης, εδώ γίνεται ομαδοποίηση των παρατηρήσεων ανάλογα με το πόσο όμοια είναι μεταξύ τους. Το αποτέλεσμα είναι η δημιουργία κάποιων clusters με τα κέντρα τους και ο υπολογισμός των αποστάσεων των

Στην ανάλυση αυτή εξετάζεται η εξάρτηση του προσδόκιμου ορίου ζωής από τον εμβολιασμό του πληθυσμού εναντίων κάποιων συγκεκριμένων ασθενειών, Ηπατίτιδα Β, Polio, Diptheria.

Αρχικά ομαδοποιούνται τα δεδομένα σε 3 και 4 clusters και παρουσιάζονται τα κεντροειδή της κάθε ομάδας καθώς και το πλήθος των παρατηρήσεων που την απαρτίζουν. Έπειτα με την χρήση της μεθόδου elbow επιχειρείται η εύρεση του βέλτιστου αριθμού clusters.

Στο δεύτερο μέρος γίνεται ιεραρχική ομαδοποίηση των δεδομένων και παράγονται τα αντίστοιχα δενδρογράμματα για με complete linkage και με single linkage.

2 K-means clustering

Ο αλγόριθμος K-means clustering είναι μια μέθοδος μη επιβλεπόμενης μάθησης για την ομαδοποίηση δεδομένων σε K ομάδες. Αρχικοποιεί K κεντροειδή και στη συνέχεια αναθέτει, με επαναλληπτική διαδικασία, σημεία δεδομένων στο πλησιέστερο κεντροειδές. Στην συνέχεια ενημερώνει τα κεντροειδή υπολογίζοντας την απόσταση του κάθε σημείου από το εκάστοτε κεντροειδές, μέχρι να σταθεροποιηθούν τα κεντροειδή.

2.1 K = 3

Για τρεις ομάδες τα αποτελέσματα φαίνονται παρακάτω.

Cluster	Hepatitis B	Polio	Diptheria
1	0.51317155	0.44674693	0.44303206
2	-1.2571265	-2.15908747	-2.09250953
3	-1.01764069	-0.27598103	-0.3015421

Πίνακας 1: Κεντροειδή ομάδων k=3.

Ομάδα	Παρατηρήσεις
1	999
2	295
3	169

Πίνακας 2: Αριθμός παρατηρήσεων ανά ομάδα, k=3.

2.2 K = 4

Για 4 ομάδες τα κεντροειδή και ο αριθμός παρατηρήσεων ανά ομάδα φαίνονται στους ακόλουθους πίνακες.

Cluster	Hepatitis B	Polio	Diptheria
1	0.50997413	0.45285415	0.4448392
2	-1.69078705	-1.63615923	-3.1096619
3	-2.25726238	0.18956665	0.17588103
4	-0.39306467	-1.14047574	-0.49559447

Πίνακας 3: Κεντροειδή ομάδων, k=4.

Είναι αξιοσημείωτο ότι και στις δύο περιπτώσεις, για k=3 και k=4, η πρώτη ομάδα, στην οποία βρίσκονται και οι περισσότερες παρατηρήσεις, διατηρεί σχεδόν το ίδιο κεντροειδές.

Ομάδα	Παρατηρήσεις
1	997
2	259
3	107
4	100

Πίνακας 4: Αριθμός παρατηρήσεων ανά ομάδα, k=4.

3 Βέλτιστο K

Ο βέλτιστος αριθμός clusters προκύπτει υπολογίζοντας το άθροισμα του αθροίσματος των αποστάσεων όλων των σημείων κάθε ομάδας από το κεντροειδές της, within-cluster sum of squares ή αδράνεια.

Για μια ομάδα C_i με κεντροειδές μ_i, η αδράνεια υπολογίζεται ως:

$$\text{Inertia}_i = \sum_{x \in C_i} \|x - \mu_i\|^2$$

Όπου: - x είναι ένα σημείο δεδομένων στη ομάδα C_i - μ_i είναι το κεντροειδές της ομάδας C_i - $\|\cdot\|$ δηλώνει την Ευκλείδεια απόσταση

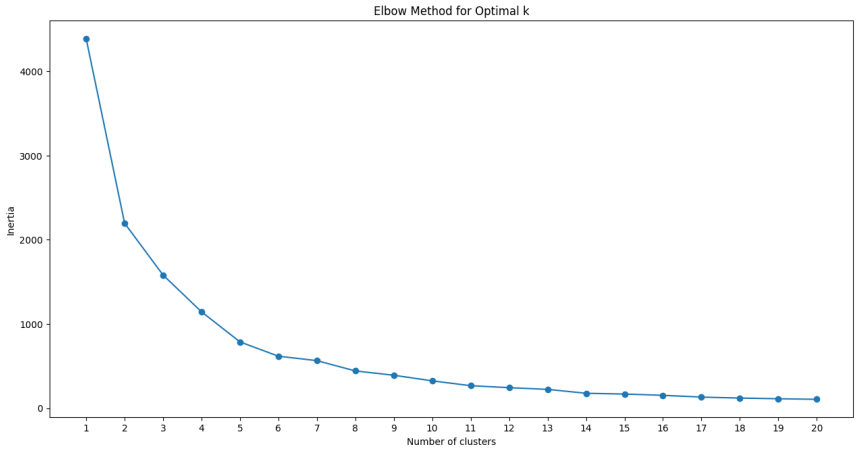
Η συνολική αδράνεια για όλη την ομαδοποίηση είναι το άθροισμα των τιμών αδράνειας για όλες τις ομάδες:

$$\text{Total Inertia} = \sum_{i=1}^{K} \text{Inertia}_i = \sum_{i=1}^{K} \sum_{x \in C_i} \|x - \mu_i\|^2$$

Η αδράνεια είναι ένα σημαντικό μέτρο για την αξιολόγηση της ποιότητας των ομάδων καθώς δίνει μια ποσοτική εκτίμηση του πόσο συμπαγείς είναι. Μικρή τιμή δείχνει ότι τα σημεία βρίσκονται αρκετά κοντά μεταξύ τους σε κάθε ομάδα, με τις ομάδες να έχουν οριστεί σωστά. Μεγάλη τιμή δείχνει ότι τα σημεία είναι αρκετά απλωμένα μέσα σε κάθε ομάδα.

Για την εφαρμογή της elbow method, υπολογίζεται η αδράνεια για διάφορες τιμές του K. Έπειτα βρίσκεται ο "αγκώνας" του γραφήματος, το σημείο που ο "ρυθμός βελτίωσης" (δεν είναι συνεχές μέγεθος άρα δεν υφίσταται ρυθμός), μειώνεται. Σύμφωνα με την "elbow method", ο βέλτιστος αριθμός clusters εδώ είναι 6.

Όπως και πριν έτσι και εδώ, το πρώτο cluster, με το μεγαλύτερο αριθμό παρατηρήσεων έχει σχεδόν το ίδιο κέντρο με πριν.



Σχήμα 1: Inertia με πλήθος K.

Cluster	Hepatitis B	Polio	Diptheria
1	0.5525116	0.49859809	0.49356559
2	-1.47496334	-0.72489302	-2.91141582
3	-2.38359224	0.18094352	0.17547732
4	-0.25931646	-0.36351979	-0.38278384
5	-0.43179295	-3.31526521	-0.21428379
6	-2.02946244	-3.23728624	-3.3798574

Πίνακας 5: Κεντροειδή ομάδων, k=6.

Cluster	Observations
1	908
2	298
3	91
4	72
5	56
6	38

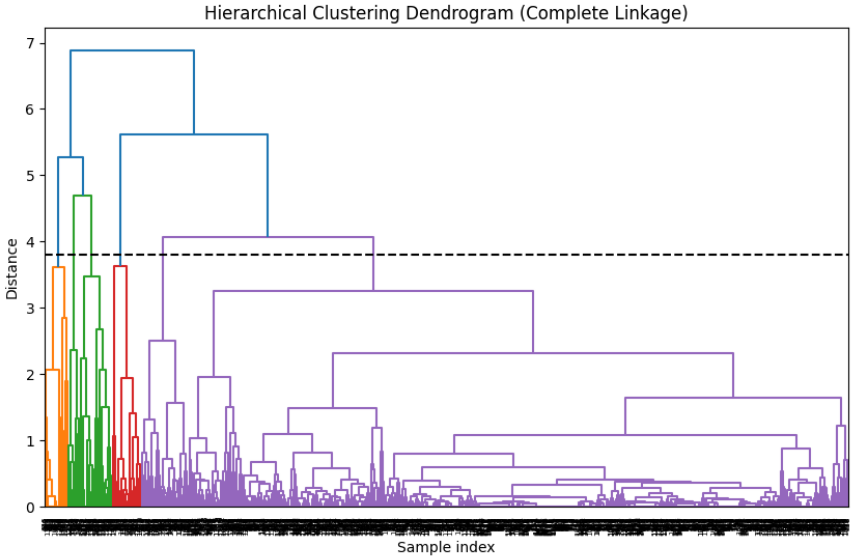
Πίνακας 6: Αριθμός παρατηρήσεων ανά ομάδα, k=6.

4 Hierarchical clustering

Ενώ πριν είχαμε προκαθορισμένο αριθμό clusters για την ομαδοποίηση των δεδομένων, τώρα η ομαδοποίηση φαίνεται σε ένα δενδρόγραμμα. Τέρμα κάτω φαίνονται όλες οι η παρατηρήσεις, ενώ κάθε branch που δημιουργείται είναι και μια ομαδοποίηση κάποιων παρατηρήσεων.

4.1 Complete linkage

Η μέθοδος Complete linkage υπολογίζει όλα τα ζεύγη dissimilarity των παρατηρήσεων μεταξύ ενός cluster A και B και παίρνει την μεγαλύτερη από αυτές.



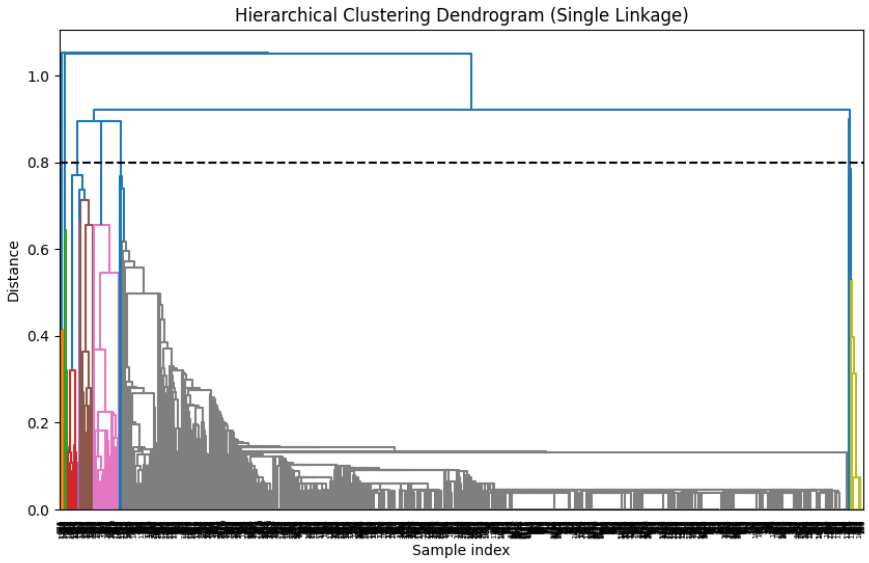
Σχήμα 2: Complete Linkage.

Όπως φαίνεται και στο δενδρόγραμμα επιλέγοντας μια απόσταση 3.8, προκύπτουν συνολικά 6 ομάδες, όσες ήταν και ο βέλτιστος αριθμός προηγουμένως με την K-cluster όπως προέκυψε από το elbow. Θα μπορούσαμε βέβαια να επιλέξουμε έναν αριθμό σχεδόν μεταξύ του 2

και 4 και να υπήρχε αρκετός χώρος για να μην υπάρξει τομή με κάποιο σημείο διακλάδωσης.

4.2 Single linkage

Αντίθετα η Single linkage παίρνει το μικρότερο dissimilarity. Το μειονέκτημα αυτής της μέθοδου μπορεί να φανεί και στο δενδρόγραμμα που ακολουθεί με τις παρατηρήσεις να ενώνονται μια μια κάθε φορά.



Σχήμα 3: Single Linkage.

Ομοίως και εδώ επιλέγεται ένας αριθμός για την απόσταση στο 0.8 με τις ομάδες να χωρίζονται στις 7.