Ανακάλυψη Γνώσης σε Βάσεις Δεδομένων (ISE709)

ΑΚΑΔ. ΕΤΟΣ 2024-25

Διδάσκουσα: Γ. Κολωνιάρη

HW3

Εξόρυξη Γνώσης σε Δεδομένα - Εργασία 3 (Clustering)

Φοιτητής: Ερρίκος Ματεβοσιάν
ΑΜ: iis23018

(A) Συσταδοποίηση με το χέρι, Hierarchical και KMeans

(i)

ΒΗΜΑ 1: a(0), b(3), c(5), d(8), e(10), f(12), g(16), h(17), i(19), j(20)
ΒΗΜΑ 2: a(0), b(3), c(5), d(8), e(10), f(12), g(16), h(17), ij(19,20)
ΒΗΜΑ 3: a(0), b(3), c(5), d(8), e(10), f(12), gh(16,17), ij(19,20)
ΒΗΜΑ 4: a(0), bc(3,5), d(8), e(10), f(12), gh(16,17), ij(19,20)
ΒΗΜΑ 5: a(0), bc(3,5), de(8,10), f(12), gh(16,17), ij(19,20)
ΒΗΜΑ 6: a(0), bc(3,5), def(8,10,12), gh(16,17), ij(19,20)
ΒΗΜΑ 7: abc(0,3,5), def(8,10,12), gh(16,17), ij(19,20)
ΒΗΜΑ 8: abcdef(0,3,5,8,10,12), gh(16,17), ij(19,20)
ΒΗΜΑ 9: abcdef(0,3,5,8,10,12), ghij(16,17,19,20)
ΒΗΜΑ 10: abcdef(0,3,5,8,10,12,16,17,19,20)

(ii)

ΒΗΜΑ 1: a(0), b(3), c(5), d(8), e(10), f(12), g(16), h(17), i(19), j(20)
ΒΗΜΑ 2: a(0), b(3), c(5), d(8), e(10), f(12), g(16), h(17), ij(19,20)
ΒΗΜΑ 3: a(0), b(3), c(5), d(8), e(10), f(12), gh(16,17), ij(19,20)
ΒΗΜΑ 4: a(0), bc(3,5), d(8), e(10), f(12), gh(16,17), ij(19,20)
ΒΗΜΑ 5: a(0), bc(3,5), de(8,10), f(12), gh(16,17), ij(19,20)
ΒΗΜΑ 6: a(0), bc(3,5), def(8,10,12), gh(16,17), ij(19,20)
ΒΗΜΑ 7: abc(0,3,5), def(8,10,12), gh(16,17), ij(19,20)
ΒΗΜΑ 8: abcdef(0,3,5,8,10,12), gh(16,17), ij(19,20)
ΒΗΜΑ 9: abcdef(0,3,5,8,10,12), ghij(16,17,19,20)
ΒΗΜΑ 10: abcdefgij(0,3,5,8,10,12,16,17,19,20)

(iii)

ΒΗΜΑ 1: c1-centroid=0, c2-centroid=3, c3-centroid=8
ανάθεση σημείων: c1(a) c2(b,c) c3(d,e,f,g,h,i,j)

ΒΗΜΑ 2: c1-centroid=0, c2-centroid=4, c3-centroid=14.57
ανάθεση σημείων: c1(a) c2(b,c,d) c3(e,f,g,h,i,j)

ΒΗΜΑ 3: c1-centroid=0, c2-centroid=5.33, c3-centroid=15.67
ανάθεση σημείων: c1(a) c2(b,c,d) c3(e,f,g,h,i,j)

ΒΗΜΑ 4: c1-centroid=0, c2-centroid=5.33, c3-centroid=15.67
ανάθεση σημείων: ΙΔΙΑ ΜΕ ΠΡΙΝ

Γ) Weka
K-Means:

- Για το K-Means, δοκιμάστηκαν 5 διαφορετικα random seeds (1, 10, 45, 100, 130), και τα αποτελέσματα για κάθε seed παρουσίασαν κάποιες διακυμάνσεις στις κατανομές των δεδομένων στους 4 κλάδους.
- Οι ομάδες 0, 1, 2 και 3 είχαν διάφορες κατανομές, με τις περισσότερες ομάδες να περιλαμβάνουν 13-23 άτομα, και με τις κατηγορίες να είναι κάπως ασαφείς, χωρίς να υπάρχει πλήρης αντιστοιχία μεταξύ των κατηγοριών και των ομάδων.
- Τα αποτελέσματα ήταν ικανοποιητικά, αν και απαιτήθηκαν πολλές δοκιμές με διάφορα seed για να επιτευχθούν πιο ισχυρές συστάσεις.

```
Clustered Instances

0       20 ( 32%)
1       23 ( 37%)
2        8 ( 13%)
3       12 ( 19%)
```

1

```
Clustered Instances

0       23 ( 37%)
1       20 ( 32%)
2       12 ( 19%)
3        8 ( 13%)
```

10

```
Clustered Instances

0       23 ( 37%)
1       12 ( 19%)
2        8 ( 13%)
3       20 ( 32%)
```

45

```
Clustered Instances

0       10 ( 16%)
1       13 ( 21%)
2       12 ( 19%)
3       28 ( 44%)
```

100

```
Clustered Instances

0       12 ( 19%)
1       23 ( 37%)
2        8 ( 13%)
3       20 ( 32%)
```

130

Hierarchical Clustering:

- Δοκιμάστηκαν τέσσερις διαφορετικοί τρόποι για την μέτρηση της απόστασης μεταξύ των ομάδων: SINGLE, COMPLETE, AVERAGE, και CENTROID.
- Τα αποτελέσματα ήταν παρόμοια για κάθε μέθοδο μέτρησης της απόστασης. Ο αλγόριθμος δημιούργησε 4 ομάδες, με κατανομές παρόμοιες με εκείνες του K-Means (όπως 37% στην ομάδα 0, 32% στην ομάδα 3 κλπ.).
- Ωστόσο με Hierarchical Clustering εμφανίστηκαν πιο "σφιχτές" ομάδες, χωρίς μεγάλες διακυμάνσεις στις κατανομές των δεδομένων, σε σχέση με το K-Means. Πέρα από τον COMPLETE, τα υπόλοιπα ήταν ίδια

```
Clustered Instances

0       23 ( 37%)
1        8 ( 13%)
2       12 ( 19%)
3       20 ( 32%)
```

ΟΛΑ

```
Clustered Instances

0       14 ( 22%)
1       21 ( 33%)
2        8 ( 13%)
3       20 ( 32%)
```

COMPLETE

Συγκρίνοντας τους δύο αλγόριθμους, η απόδοση του K-Means εξαρτάται από την αρχική επιλογή των seeds, με αποτέλεσμα να παρατηρούνται διαφορετικές κατανομές ανάλογα με την τυχαία αρχικοποίηση. Αυτό μπορεί να οδηγήσει σε λιγότερο σταθερά αποτελέσματα. Από την άλλη, ο Hierarchical παράγει πιο συνεκτικές και σταθερές ομάδες

Με βάση τα αποτελέσματα της άσκησης και την ανάλυση των δύο αλγορίθμων, **ο Hierachical φαίνεται να είναι η καλύτερη επιλογή** για το σύνολο δεδομένων "cancer.arff". Ο αλγόριθμος αυτός παρέχει πιο συνεκτικά και αξιόπιστα αποτελέσματα σε σύγκριση με το K-Means,