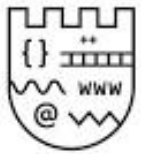
Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης



SCHOOL OF INFORMATICS AUTH

Αποθήκες δεδομένων και εξόρυξη δεδομένων

Παρουσίαση εργασίας
Διαμαντής Νικόλαος Α.Ε.Μ.: 3396
Κιζιρίδης Κωνσταντίνος Α.Ε.Μ.: 3566

Εισαγωγή

Η συγκεκριμένη εργασία αφορά στην αξιοποίηση του εργαλείου Apache Spark για την ανάλυση ενός συνόλου δεδομένων και εξαγωγή πληροφοριών απ’ αυτά.

Αναλυτικότερα, ζητούμενο αποτελεί η ανάγνωση τους, ο καθαρισμός τους, η μορφοποίησή τους και η εφαρμογή του αλγορίθμου K-means για τον διαχωρισμό σε συστάδες και τον εντοπισμό ανωμαλιών (outliers).

Σημείωση: Για την υλοποίηση της εργασίας χρησιμοποιήθηκε η γλώσσα προγραμματισμού Scala στο προγραμματιστικό περιβάλλον Intellij IDEA.

Βήμα 1ο - Διάβασμα δεδομένων

Πρωτίστως, διαβάζουμε το αρχείο των δεδομένων με .csv format, καθώς έτσι τα δεδομένα καταχωρούνται σε δύο στήλες _c0 και _c1 και γίνεται ευκολότερος ο χειρίσμος τους. Τα δεδομένα αποθηκεύονται σε ένα αντικείμενο DataFrame.

Βήμα 2ο – Καθαρισμός δεδομένων

Στην συνέχεια, πρέπει τα δεδομένα που έχουν διαβαστεί απ’ το αρχείο εισόδου να καθαριστούν, διότι υπάρχουν εγγραφές σημείων απ’ τις οποίες απουσιάζουν είτε τιμές x είτε τιμές y.

Ο καθαρισμός επιτυγχάνεται χρησιμοποιώντας την εντολή na.drop() που εφαρμόζεται στο DataFrame που δημιουργήθηκε στο βήμα 1. Η εντολή αυτή διαγράφει τις γραμμές στις οποίες εντοπίζει null τιμή στις στήλες _c0, _c1.

Βήμα 3ο – Μετασχηματισμός 0 – 1

Συνεχίζοντας, χρησιμοποιούμε τον τύπο :

$$z_i = \frac{x_i - min}{max - min}$$

προκειμένου να μετασχηματίσουμε όλα τα δεδομένα μας σε τιμές μεταξύ του 0 και του 1.

Βήμα 4ο - Εφαρμογή K-means

Έχοντας φτάσει στο σημείο αυτό, εφαρμόζουμε τον αλγόριθμο K-means για τον σχηματισμό συστάδων στα δεδομένα. Ως βάση για τον K-means χρησιμοποιήθηκε το δοθέν παράδειγμα. Ο K – means εκτελέστηκε με δεδομένο, ότι ψάχνουμε για 5 ομάδες.

Βήμα 5ο – Εντοπισμός outliers.

Για τον εντοπισμό των outliers εντός των συστάδων χρησιμοποιούμε την ευκλείδεια απόσταση μεταξύ των κέντρων που βρήκε ο K – means και των σημείων που καταχώρησε σε κάθε συστάδα.

Έπειτα θέτοντας ένα φίλτρο για το ποιες τιμές θεωρούμε αποδεκτές αποκλείουμε σημεία που το παραβιάζουν ως outliers.

Αποτελέσματα εκτέλεσης αρχείου εισόδου

Με την εκτέλεση του K – means εντοπίζονται τα παρακάτω κέντρα:

```
Cluster Centers:
[0.7650176315287435,0.13573743587265333]
[0.17283704493797516,0.7387380578633888]
[0.500396764294586,0.2706373501334986]
[0.17362909127121856,0.14271574296719783]
[0.725078964190676,0.738305089680052]
```

Εντοπισμός ανωμαλιών - outliers

Για τον εντοπισμό των ανωμαλιών, υπολογίστηκε η ευκλείδεια απόσταση όλων των σημείων από το κέντρο της συστάδας της οποίας είναι μέλη.

Στην συνέχεια, για κάθε συστάδα υπολογίστηκε η μέση απόσταση που έχουν τα μέλη της απ’ το κέντρο της και σχηματίστηκε ο λόγος

$$\frac{απόσταση\ απ'το\ κεντρο\ της\ συσταδας}{μεση\ αποσταση\ απ'το\ κεντρο\ της\ συσταδας}$$

για κάθε σημείο του συνόλου δεδομένων.

Όσο μεγαλύτερος είναι ο συγκεκριμένος λόγος, τόσο πιο απομακρυσμένο είναι το σημείο. Παρατηρήσαμε για το δοθέν σύνολο δεδομένων ότι όταν θέσαμε ως φίλτρο τον λόγο να είναι μεγαλύτερος από 2,4 τότε εμφανίστηκαν 6 ανώμαλες τιμές.

Οι ανώμαλες τιμές που εντοπίστηκαν

```
Outliers:
+------------------+------------------+--------------------+----------+--------------------+------------------+
|               _c0|               _c1|            features|prediction|            distance|       divDistance|
+------------------+------------------+--------------------+----------+--------------------+------------------+
|0.4433307368319466|0.9048165137614679|[0.44333073683194...|         1| 0.3174096578050532| 2.806962696960487|
|0.7324594842383174|0.4461009174311927|[0.73245948423831...|         2|0.29093052338404635|2.5727986104561684|
|0.9982652613390135|0.7385321100917431|[0.99826526133901...|         4| 0.2731863914763821| 2.415881153378191|
|0.8974556584199025| 0.948394495412844|[0.89745565841990...|         4|  0.271755925629146| 2.403231052243166|
|               1.0|0.7362385321100917|[1.0,0.7362385321...|         4| 0.2749288027665737| 2.431289895280224|
|0.9016961817642882|0.5321100917431193|[0.90169618176428...|         4|0.27149589079331643|2.4009314748165327|
+------------------+------------------+--------------------+----------+--------------------+------------------+
```

Χρόνοι εκτέλεσης και κλιμάκωση

Για το ενδεικτικό αρχείο εισόδου η υλοποίηση χρειάστηκε περίπου 9 δευτερόλεπτα για να ολοκληρωθεί. Δοκιμάζοντας ένα νέο αρχείο εισόδου με περίπου δεκαπλάσιο αριθμό εγγραφών χρειάστηκαν περίπου 18 δευτερόλεπτα για την ολοκλήρωση της εκτέλεσης.

Εξάγεται, λοιπόν, το συμπέρασμα ότι η υλοποίηση έχει καλή χρονική κλιμάκωση για μεγαλύτερο όγκο δεδομένων.

ΤΕΛΟΣ
ΕΡΓΑΣΙΑΣ