Επεκτάσεις του αλγορίθμου K-means (Spark)

Ιωάννης Αθανασιάδης
607
agioannis@csd.auth.gr

Αντιγόνη-Μαρία Φούντα
647
founanti@csd.auth.gr

ΠΕΡΙΛΗΨΗ

Στην αναφορά που ακολουθεί αναλύονται δύο υλοποιήσεις σχετικές με τον αλγόριθμο k-means, η υλοποίηση του Συντελεστή Σιλουέτας και της Αρχικοποίησης Κέντρων. Στο τέλος ακολουθούν συγκριτικά αποτελέσματα και συμπεράσματα ως προς αυτά, σε διαφορετικά σύνολα δεδομένων.

Λέξεις Κλειδιά

K-means, Συντελεστής Σιλουέτας, Συσταδοποίηση, Spark.

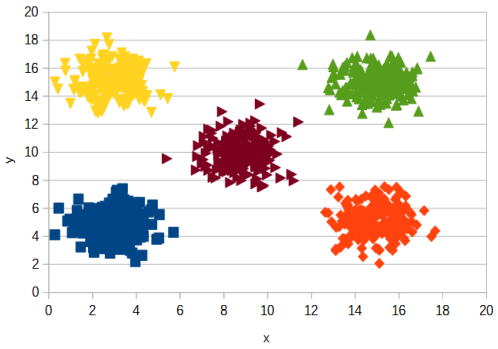
1. ΕΙΣΑΓΩΓΗ

Η παρούσα εργασία πραγματοποιείται στα πλαίσια του μαθήματος «Εξόρυξη από Μεγάλα Δεδομένα» του μεταπτυχιακού τμήματος του Αριστοτελείου Πανεπιστημίου. Στόχος της εργασίας είναι ο πειραματισμός με την υλοποίηση του αλγορίθμου k-means στην βιβλιοθήκη Spark. Πιο συγκεκριμένα κληθήκαμε να υλοποιήσουμε τον συντελεστή σιλουέτας προκειμένου να μπορούμε να αξιολογούμε την ποιότητα της ομαδοποίησης, καθώς επίσης και να υλοποιήσουμε την ανεύρεση αρχικών κέντρων που δίνονται ως όρισμα για να ξεκινήσει ο k-means και να συγκρίνουμε τα αποτελέσματα αυτής της μεθοδολογίας με άλλες μεθοδολογίες του (π.χ. k-means με τυχαία αρχικοποίηση ή παράλληλο k-means). Παρακάτω περιγράφουμε αναλυτικά τα δεδομένα που χρησιμοποιήθηκαν και τις υλοποιήσεις και τέλος σχολιάζουμε τα αποτελέσματα κάθε μέρους. Όλος ο κώδικας της εργασίας καθώς και τα τρία σύνολα δεδομένων βρίσκονται στο παρακάτω repository του GitHub: https://github.com/AndyFou/kmeans_contributions.

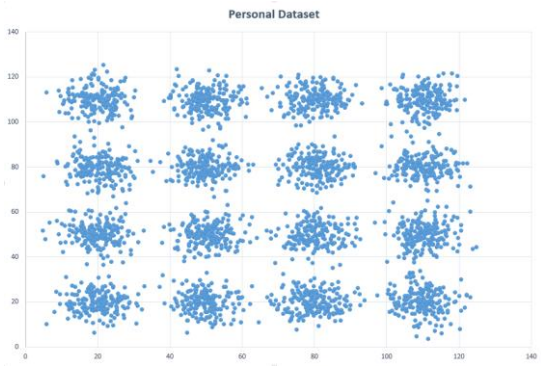
2. ΔΕΔΟΜΕΝΑ

Τα δεδομένα που χρησιμοποιήθηκαν ως δείγμα και στα δύο μέρη της εργασίας για την εκπόνηση των αρχικών αποτελεσμάτων αποτελούνται από 1000 σημεία στον δισδιάστατο χώρο, η αναπαράσταση των οποίων βρίσκεται στην παρακάτω εικόνα (Εικόνα 2.1). Θεωρήθηκε ως δεδομένο ότι οι συστάδες που δόθηκαν ως όρισμα στον αλγόριθμο είναι 5, όπως δόθηκε από την εκφώνηση. Το σύνολο δεδομένων αυτό από τώρα και για το υπόλοιπο της εργασίας θα το αποκαλούμε σύνολο εκπαίδευσης.

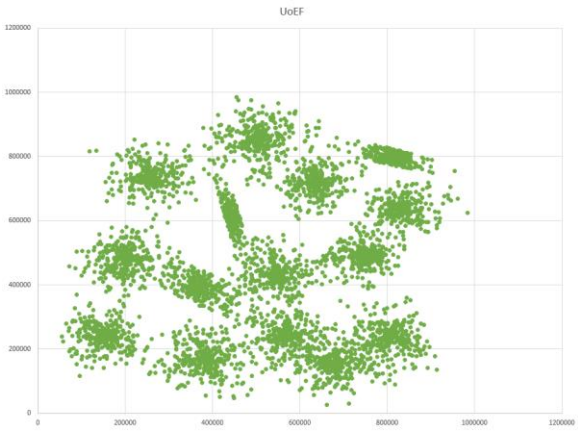
Αργότερα, για την σύγκριση των αποτελεσμάτων που παράγονται σε κάθε μέρος της εργασίας χρησιμοποιήθηκαν ακόμα δύο dataset. Το ένα σύνολο δημιουργήθηκε από εμάς, με περίπου 3000 σημεία και ακολουθώντας κανονική κατανομή, και οι ομάδες που δημιουργήθηκαν ήταν 16 κατόπιν εμπειρικής επιλογής (Εικόνα 2.2). Το δεύτερο σύνολο δεδομένων αντλήθηκε όπως θα αναφέρουμε και παρακάτω από το Πανεπιστήμιο της Ανατολικής Φινλανδίας και αποτελείται από 5000 στοιχεία και περίπου 15 συστάδες, κάποιες λιγότερα και άλλες περισσότερο διακριτές (Εικόνα 2.3). Αυτά τα δύο σύνολο από τώρα και για το υπόλοιπο της εργασίας θα τα αποκαλούμε πρώτο και δεύτερο σύνολο ελέγχου αντίστοιχα. Όπως βλέπουμε τα πειράματα πραγματοποιήθηκαν και σε απλές περιπτώσεις (π.χ. Εικ. 2.1) αλλά και σε πιο σύνθετες (π.χ. Εικ. 2.3).



Εικόνα 2.1. Οπτικοποίηση των αρχικών δεδομένων



Εικόνα 2.2. Οπτικοποίηση των δεδομένων που δημιουργήσαμε



Εικόνα 2.3. Οπτικοποίηση των δεδομένων του ΠΑΦ

3. ΣΧΕΤΙΚΕΣ ΕΡΓΑΣΙΕΣ – ΕΡΓΑΛΕΙΑ

Για την υλοποίηση της παρούσας εργασίας χρησιμοποιήθηκε το framework Spark[1] και η γλώσσα Scala μέσα από το IntelliJ IDE[2]. Το τελευταίο σύνολο δεδομένων που χρησιμοποιήθηκε στο δεύτερο μέρος της εργασίας (Ενότητα 5) για να πραγματοποιηθεί η σύγκριση των αλγορίθμων προέρχεται από την Μονάδα Επεξεργασίας Εικόνας και Ήχου του Τμήματος Πληροφορικής του Πανεπιστημίου της Ανατολικής Φινλανδίας [1]. Τέλος, όπως αναφέρθηκε και στην εισαγωγή, όλος ο κώδικας βρίσκεται ανεβασμένος στην πλατφόρμα GitHub[3].

4. ΣΥΝΤΕΛΕΣΤΗΣ ΣΙΛΟΥΕΤΑΣ

Πρόκειται για μία τεχνική η οποία χρησιμοποιείται για να αναδείξει την συνοχή ανάμεσα στις συστάδες που δημιουργούνται από τον k-means. Όσο πιο υψηλός ο συντελεστής, τόσο πιο συνεκτικές είναι οι συστάδες, επομένως και πιο καλά τα αποτελέσματα. Στην ουσία, όπως αναφέρεται και στην Wikipedia[4], ο συντελεστής σιλουέτας μετράει πόσο όμοιο είναι ένα αντικείμενο με τα υπόλοιπα σημεία της συστάδας του σε σχέση με το πόσο όμοιο είναι με τις υπόλοιπες συστάδες. Ο τύπος υπολογισμού του συντελεστή σιλουέτας είναι αυτός που φαίνεται στο παρακάτω σχήμα, όπου το a_i είναι ο μέσος όρος των αποστάσεων του σημείου i προς όλα τα σημεία της συστάδας στην οποία ανήκει, ενώ το b_i είναι ο μικρότερος από τους μέσους όρους των αποστάσεων προς τα σημεία των υπόλοιπων συστάδων, για κάθε συστάδα χωριστά. Το εύρος τιμών του S είναι [-1,1] και όσο μεγαλύτερη η τιμή του, τόσο μεγαλύτερη και η συνοχή των συστάδων.

$$S_i = \frac{b_i - a_i}{\max\{ a_i, b_i \}}$$

Εικόνα 4.1. Ο τύπος υπολογισμού του συντελεστή σιλουέτας

4.1 ΑΛΓΟΡΙΘΜΟΣ

Η διαδικασία που ακολουθήσαμε σχετικά με την ανεύρεση του συντελεστή σιλουέτας περιγράφεται συνοπτικά στον ψευδοκώδικα της Εικόνας 4.2. Σε γενικές γραμμές αρχικά πραγματοποιούμε την ομαδοποίηση με τον k-means και προβλέπουμε την συστάδα στην οποία ανήκει κάθε σημείο, στην συνέχεια υπολογίζουμε τα στοιχεία που είναι απαραίτητα για να υπολογιστεί ο συντελεστής και τέλος παράγουμε τις τιμές για τα a και b τα οποία χρησιμοποιούνται για τον υπολογισμό της τελικής τιμής του συντελεστή.

4.2 ΑΠΟΤΕΛΕΣΜΑΤΑ

Όπως φαίνεται από την Εικόνα 2.1, τα δεδομένα εκπαίδευσης στα οποία δοκιμάστηκε ο αλγόριθμος ήταν πολύ καλά σχηματισμένα και δημιουργούν πέντε καθαρά διακριτές ομάδες. Αυτό όμως σαν

```
PSEUDOCODE FOR SILHOUETTE COEFFICIENT

//Clustering procedure
Load data
Run KMeans(data)
Assign clusters

//Data preprocessing to compute Silhouette coefficient
generate ids for points
Join: IDS, Coordinates, ClusterAssignments → rdd1
create combinations (rdd1)

//We have: idFrom,idTo,clusterFrom,clusterTo,coordsFrom,coordsTo
//Compute all a,b
.filter out rows where idFrom = idTo
.map(idFrom, clusterTo, clusterFrom, computeDistance)
.reduceByKey Key=(idFrom,clusterTo) to average distances → all a,b(rdd2)

//Filter out a,b
rdd2.filter out a(where clusterFrom = clusterTo) → final a/point
union
rdd2.filter out b(where clusterFrom != clusterTo)
    .reduceByKey key = id to average distances → final b/point
//Final
reduceByKey key = id to compute Si
.reduceByKey clusterFrom to find Silhouette for each cluster
```

Εικόνα 4.2 Ψευδοκώδικας που περιγράφει τον αλγόριθμο

αποτέλεσμα μπορεί να προκύψει και από τις τιμές του συντελεστή σιλουέτας όπως φαίνονται στην πρώτη στήλη του Πίνακα 3.1, όπου σχεδόν σε όλες τις συστάδες οι τιμές των S κυμαίνονται κοντά στο 0.75, γεγονός το οποίο αποδεικνύει ότι οι συστάδες είναι όντως πολύ καλά χωρισμένες. Ακόμα παρατηρούμε ότι η τελευταία στην σειρά τιμή του συντελεστή είναι σχετικά χαμηλότερη από τις υπόλοιπες, γεγονός που θα μπορούσε να συνδεθεί με την μεσαία συστάδα και τα σημεία που αποκλίνουν από εκείνη και πλησιάζουν στις άλλες. Επίσης η κεντρική συστάδα δεν διαθέτει ακρότατα σημεία εντός της ομάδας, τα οποία επηρεάζουν σημαντικά τις τιμές του S καθώς είναι πολύ απομακρυσμένα από όλα τα σημεία όλων των άλλων συστάδων οπότε αυξάνουν πολύ το b. Επομένως η έλλειψη τέτοιων σημείων ανεβάζει την δυσκολία στον διαχωρισμό και ταυτόχρονα μικραίνει και τον συντελεστή σιλουέτας.

Στα δεδομένα ελέγχου από την άλλη, όσο γίνονται λιγότερο διακριτές οι ομάδες, τόσο πέφτουν και οι τιμές των συντελεστών. Επομένως στο πρώτο σύνολο ελέγχου για παράδειγμα οι περισσότεροι συντελεστές κυμαίνονται κοντά στο 0.65, το οποίο σαν τιμή είναι αρκετά καλό αν σκεφτεί κανείς το εύρος τιμών του συντελεστή, αλλά και πάλι είναι χαμηλότερες οι τιμές σε σχέση με αυτές του συνόλου εκπαίδευσης, το οποίο ήταν πιο διακριτά χωρισμένο. Εξαίρεση σε αυτήν την περίπτωση αποτελούν οι τέσσερις τελευταίοι συντελεστές, των οποίων οι τιμές μειώνονται σημαντικά. Εάν συγκρίνουμε αυτούς τους συντελεστές με την οπτικοποίηση των δεδομένων που υπάρχει στο πρώτο κεφάλαιο (Εικ. 2.2), μπορούμε με ασφάλεια να υποθέσουμε ότι οι τέσσερις χαμηλές τιμές της δεύτερης στήλης του Πίνακα 4.1 αντιστοιχούν στις τέσσερις ομάδες που βρίσκονται κεντρικά στο γράφημα. Επομένως η πτώση των τιμών αυτών είναι φυσιολογική για τους λόγους τους οποίους περιγράψαμε στην προηγούμενη παράγραφο. Τέλος, στα αποτελέσματα του τρίτου συνόλου δεδομένων οι τιμές των συντελεστών ποικίλλουν, από ~0.5 μέχρι ~0.75, καθώς κάποιες συστάδες είναι ευδιάκριτες και άλλες είναι αρκετά πιο δυσδιάκριτες (Εικ. 2.3).

Αξίζει επίσης να σημειωθεί ότι έγιναν πολλές δοκιμές, για το σύνολο εκπαίδευσης, σε διαφορετικά μηχανήματα και τα

[1] Spark Framework - http://spark.apache.org/

[2] IntelliJ IDE - https://www.jetbrains.com/idea/

[3] Github - https://github.com/

[4] Πηγή: https://en.wikipedia.org/wiki/Silhouette_(clustering)

αποτελέσματα ήταν πάντα τα ίδια. Τέλος, μία παρατήρηση που θα μπορούσαμε να κάνουμε είναι ότι, εξαιτίας της χρήσης του Spark framework η διαδικασία μπορεί εύκολα να παραλληλοποιηθεί.

Σύνολο Εκπαίδ.	1ο Σύνολο Ελέγχου	2ο Σύνολο Ελέγχου
0.77920811	0.683719499	0.757356871
0.77790082	0.670013729	0.742288077
0.7754347	0.669651453	0.707339403
0.76387935	0.665956086	0.685168212
0.73221311	0.665564846	0.647019165
	0.65741723	0.643173934
	0.652660004	0.640823368
	0.651263795	0.635573563
	0.639619591	0.592006317
	0.637398214	0.58553009
	0.636738418	0.580262557
	0.632402625	0.570600296
	0.56069027	0.551235779
	0.360041372	0.52179336
	0.316729391	0.516705252
	0.30386551	

Πίνακας 4.1 Τα αποτελέσματα του Συντελεστή Σιλουέτας S στα 3 σύνολα δεδομένων.

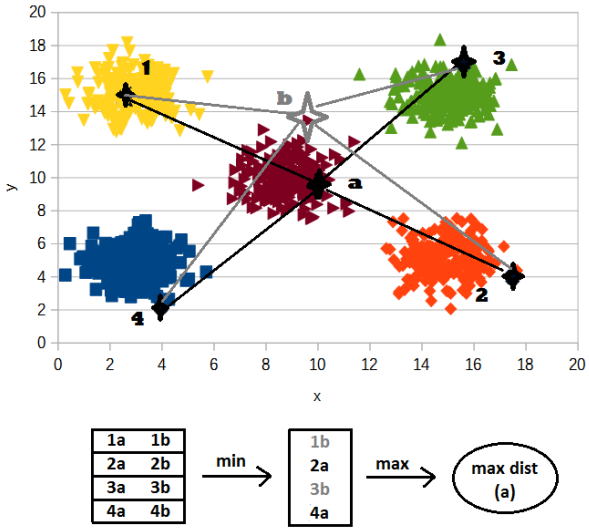
5. ΑΡΧΙΚΟΠΟΙΗΣΗ ΚΕΝΤΡΩΝ

Ο αλγόριθμος ομαδοποίησης που εξετάζεται στην παρούσα αναφορά είναι ο k-means, ο οποίος παίρνοντας ως είσοδο το πλήθος των κέντρων ξεκινάει μία επαναληπτική διαδικασία στην οποία ψάχνει να βρει τα κέντρα των ομάδων και αναθέτει σε κάθε ομάδα τα σημεία του συνόλου δεδομένων. Κατά την αρχικοποίηση του αλγορίθμου οι τιμές των κέντρων μπορούν να δοθούν είτε τυχαία, είτε σαν όρισμα από τον χρήστη. Σε αυτό το μέρος της αναφοράς εξετάζουμε την υλοποίηση μιας μεθοδολογίας κατά την οποία υπολογίζονται τα αρχικά κέντρα με βάση τις αποστάσεις μεταξύ των σημείων, όπως περιγράφεται στο βιβλίο των Rajaraman και Ullman [2]. Στην συνέχεια τα κέντρα αυτά δίνονται ως είσοδο στον αλγόριθμο και πραγματοποιείται κανονικά η εκτέλεση του αλγορίθμου έως ότου οι τιμές στις επαναλήψεις συγκλίνουν. Τέλος, γίνεται μία σύγκριση των αποτελεσμάτων αυτής της μεθοδολογίας με δύο άλλες, τον τυχαίο k-means και τον παράλληλο k-means όπου παρουσιάζονται οι διαφορές στις εκτελέσεις με βάση τον χρόνο εκτέλεσης και τις επαναλήψεις που γίνονται.

5.1 ΜΕΘΟΔΟΛΟΓΙΑ

Η λογική πίσω από την μεθοδολογία της αρχικοποίησης των κέντρων βασίζεται στην ανεύρεση των μεγαλύτερων αποστάσεων μεταξύ των σημείων που έχουν ανατεθεί ως κέντρα. Η σύγκριση γίνεται από κάθε σημείο προς όλα τα άλλα.

Πιο συγκεκριμένα αρχικά ανατίθεται το πρώτο κέντρο τυχαία και στην συνέχεια από κάθε κέντρο που έχει ανατεθεί υπολογίζονται όλες οι αποστάσεις από τα όλα τα σημεία. Κατόπιν από όλες τις αποστάσεις κρατάμε για κάθε σημείο την μικρότερη (min) και καταλήγουμε σε μία λίστα από αποστάσεις σημείων από τις οποίες κρατάμε την μεγαλύτερη. Με αυτόν τον τρόπο η αρχικοποίηση επιτυγχάνει να υπολογίσει ένα σύνολο από κέντρα τα οποία απέχουν την μεγαλύτερη δυνατή απόσταση μεταξύ τους. Σκοπός αυτής της μεθοδολογίας είναι να δοθούν αυτά τα απομακρυσμένα κέντρα στον αλγόριθμο ώστε κατά τον επαναπροσδιορισμό των κέντρων να μην χρειαστούν πολλές επαναλήψεις και κυρίως η διάκριση των δεδομένων να γίνει όσο τον δυνατόν καλύτερα.



Εικόνα 5.1 Παράδειγμα υπολογισμού αρχικών κέντρων

5.2 ΑΠΟΤΕΛΕΣΜΑΤΑ

Ύστερα από αρκετά πειράματα που έγιναν στο σύνολο εκπαίδευσης, διαπιστώσαμε ότι πράγματι η ανάθεση των κέντρων γίνεται με βάση την λογική που περιγράψαμε νωρίτερα, καθώς σε κάθε εκτέλεση της μεθοδολογίας που υλοποιήσαμε τα αποτελέσματα ήταν πάντα αντίστοιχα: το πρώτο κέντρο που ανατίθεται, σε σχέση με το δεύτερο, είναι πάντα διαμετρικά αντίθετα. Επομένως είναι εκείνα τα οποία απέχουν την μεγαλύτερη απόσταση, και η αναζήτηση των υπολοίπων συνεχίζει με τον ίδιο τρόπο. Η μόνη περίπτωση στην οποία αυτό δεν ισχύει είναι όταν το πρώτο κέντρο που δημιουργείται ανήκει στην μεσαία συστάδα, όπου η σειρά με την οποία υπολογίζονται τα υπόλοιπα κέντρα δεν έχει σημασία καθώς οι αποστάσεις είναι λίγο-πολύ αντίστοιχες. Τέλος, παρατηρήσαμε ότι εάν το πρώτο κέντρο που υπολογίζεται ανήκει σε κάποια ακριανή συστάδα, τότε το κέντρο της κεντρικής συστάδας εμφανίζεται τελευταίο, πράγμα το οποίο είναι λογικό καθώς όλες οι ακριανές συστάδες δεν απέχουν πολύ από την κεντρική.

5.3 ΣΥΓΚΡΙΣΗ ΑΛΓΟΡΙΘΜΩΝ

Παρακάτω υπάρχουν αναλυτικά τα συγκριτικά αποτελέσματα των τριών αλγορίθμων ως προς τους χρόνους εκτέλεσης και τις επαναλήψεις, στα τρία σύνολα δεδομένων που χρησιμοποιήσαμε (Πίνακες 5.1, 5.2, 5.3 αντίστοιχα). Σε γενικές γραμμές τα αποτελέσματα είναι παρόμοια και δείχνουν ότι υπάρχει μεγάλη εξάρτηση από το σύνολο δεδομένων που χρησιμοποιείται κάθε φορά και το υλικό του μηχανήματος στο οποίο εκτελούνται.

k=5	**Χρόνος Εκτέλεσης**	**Επαναλήψεις**
Random k-means	0.852	3
Parallel k-means	2.328	2
Init. k-means	5.538 + 1.396	3

Πίνακας 5.1 Συγκριτικά αποτελέσματα των τριών αλγορίθμων για το σύνολο εκπαίδευσης.

k=16	**Χρόνος Εκτέλεσης**	**Επαναλήψεις**
Random k-means	2.13	24
Parallel k-means	2.682	8
Init. k-means	21.283 + 0.992	5

Πίνακας 5.2 Συγκριτικά αποτελέσματα των τριών αλγορίθμων για το πρώτο σύνολο ελέγχου.

k=15	**Χρόνος Εκτέλεσης**	**Επαναλήψεις**
Random k-means	2.206	18
Parallel k-means	2.624	5
Init. k-means	22.352 + 4.641	24

Πίνακας 5.3 Συγκριτικά αποτελέσματα των τριών αλγορίθμων για το δεύτερο σύνολο ελέγχου.

Ως προς τους χρόνους εκτέλεσης, σε κάθε περίπτωση ο πιο γρήγορος από τους αλγορίθμους είναι ο random k-means, με δεύτερο τον parallel. Η αρχικοποίηση των κέντρων πάντα διαρκεί αρκετά περισσότερο απ' όσο διαρκεί η εκτέλεση στις άλλες περιπτώσεις (ο χρόνος παρατηρείται ότι ανεβαίνει σχεδόν γραμμικά ανάλογα με το μέγεθος του dataset), γεγονός το οποίο μοιάζει να απορρίπτει την μέθοδο αυτή. Παρ' όλα αυτά, αν συγκρίνουμε μόνο τους χρόνους εκτέλεσης χωρίς την προετοιμασία θα δούμε ότι ο k-means με αρχικοποιημένα κέντρα είναι αρκετά γρήγορος. Επομένως, πριν βιαστούμε να απορρίψουμε αυτήν την μέθοδο πρέπει να έχουμε υπόψιν ότι τα νούμερα αυτά σε πολύ δύσκολα και πολύπλοκα σύνολα δεδομένων αλλάζουν ριζικά.

Σχετικά με τις επαναλήψεις, παρατηρούμε ότι στις περισσότερες περιπτώσεις η καλύτερη απόδοση επιτυγχάνεται από τον παράλληλο k-means. Επιπλέον, βλέπουμε ότι οι επαναλήψεις δεν είναι ανάλογες του χρόνου εκτέλεσης, καθώς οι χρόνοι εκτέλεσης παραδείγματος χάριν στα δύο πρώτα σύνολα δεδομένων είναι παρόμοιοι, ενώ οι επαναλήψεις που χρειάζονται για να βρεθούν οι συστάδες αυξάνονται σημαντικά. Σε γενικές γραμμές, προκειμένου να καταλήξουμε σε ένα ασφαλές συμπέρασμα σχετικά με αυτό το αποτέλεσμα θα ήταν σημαντικό να πραγματοποιηθούν περισσότερα πειράματα και σε περισσότερα μηχανήματα.

Τέλος, πραγματοποιήθηκαν πειράματα σχετικά με το πλήθος των σημείων που ανατίθενται σε κάθε ομάδα και την συσχέτιση αυτών με την απόδοση των αλγορίθμων. Παρόλο που στα δύο πρώτα σύνολα δεδομένων οι συστάδες έχουν πράγματι παρόμοιο αριθμό σημείων σε κάθε ομάδα, αυτό δεν έδειξε να σχετίζεται με την εκτέλεση των αλγορίθμων, καθώς όλα τα αποτελέσματα ήταν πολύ κοντινά. Επομένως, μπορούμε να βγάλουμε συμπεράσματα σχετικά με αυτό ως προς την φύση των δεδομένων μας αλλά όχι ως προς την αξιολόγηση των αλγορίθμων.

6. ΣΥΝΟΨΗ

Στην παρούσα εργασία πραγματοποιήθηκαν δύο υλοποιήσεις σχετικές με τον αλγόριθμο k-means του Spark. Πιο συγκεκριμένα, στο πρώτο μέρος της εργασίας υλοποιήθηκε ο Συντελεστής Σιλουέτας, ο οποίος χρησιμοποιείται ως μετρική για το πόσο καλά είναι διαχωρισμένες οι συστάδες που παράχθηκαν. Ως αποτέλεσμα από αυτό το μέρος της εργασίας θα μπορούσαμε να πούμε ότι ο συντελεστής σιλουέτας αποτελεί σημαντικό εργαλείο για την εξακρίβωση των αποτελεσμάτων της ομαδοποίησης, και είναι γρήγορος στην εκτέλεση των αποτελεσμάτων του.

Στο δεύτερο μέρος της εργασίας υλοποιήθηκε η Αρχικοποίηση Κέντρων, μέσα από την οποία υπολογίζονται τα αρχικά κέντρα τα οποία στην συνέχεια δίνονται ως όρισμα στον k-means για να εκτελεστεί κατόπιν κανονικά. Για να βγάλουμε συμπεράσματα ως προς αυτήν την τεχνική, βγάλαμε συγκριτικά αποτελέσματα ως προς δύο ακόμα παραλλαγές του αλγορίθμου, τον τυχαίο k-means και τον παράλληλο k-means. Από τα αποτελέσματα που προέκυψαν μπορούμε σε γενικές γραμμές να αναφέρουμε ότι και οι τρεις τεχνικές λειτουργούν καλά, έχοντας διαφορετικά προτερήματα η κάθε μία. Επομένως η επιλογή του εκάστοτε κατάλληλου αλγορίθμου εξαρτάται πάντα από το σύνολο δεδομένων στο οποίο θα εφαρμοστεί το μοντέλο και τις αντίστοιχες συνθήκες.

Συγκεκριμένα, ως προς τον χρόνο εκτέλεσης ο αλγόριθμος που είχε πάντα τα καλύτερα αποτελέσματα ήταν ο τυχαίος k-means, γεγονός το οποίο δείχνει ότι παρόλο που η μέθοδος υπολογισμού των κέντρων και των συστάδων δεν είναι η βέλτιστη, ο αλγόριθμος εξακολουθεί να είναι ανταγωνιστικός σε σχέση με άλλες παραλλαγές. Παρ' όλα αυτά σε πιο δύσκολα σύνολα δεδομένων είναι σχεδόν βέβαιο ότι τα αποτελέσματα θα διέφεραν αρκετά.

7. ΑΝΑΦΟΡΕΣ

[1] Clustering Datasets. Speech and Image Processing Unit, School of Computing, University of Eastern Finland. http://cs.joensuu.fi/sipu/datasets/

[2] Rajaraman, A., Ullman, J. 2007. Mining of Massive Datasets. Cambridge University Press.