Υπολογιστική Νοημοσύνη
Αναφορά Άσκησης:

Υλοποίηση προγράμματος ομαδοποίησης βασισμένο στον αλγόριθμο k-means

Μέλη ομάδας

Αντωνίου Χριστόδουλος AM: 2641
Τσιούρη Αγγελική AM: 3354

Εισαγωγή

Η άσκηση έχει χωριστεί στα τρία ακόλουθα μέρη:

1. Παραγωγή τυχαίων σημείων
2. Υλοποίηση του αλγορίθμου k-means
3. Πείραμα με πολλαπλές εκτελέσεις του αλγορίθμου k-means

Κάθε μέρος εκτελείται και μεταγλωττίζεται ξεχωριστά. Προς διευκόλυνση, έχει δημιουργηθεί και συμπεριληφθεί το shell script **compile.sh**, το οποίο κάνει compile και τα 3 προγράμματα μαζί. Για την εκτέλεση του script ή και οποιουδήποτε άλλου προγράμματος της άσκησης πρέπει να μεταβούμε στον **κατάλογο kmeans** (root directory της άσκησης). Επιπλέον, ίσως χρειαστεί να δώσουμε δικαιώματα εκτέλεσης στο compile script με την εξής εντολή:

chmod +x compile.sh

Τα προγράμματα έχουν δοκιμαστεί σε ubuntu 20.04 με java 14.0.2 καθώς και στα μηχανήματα της σχολής με java 11.0.13.

Περιεχόμενα

Η άσκηση περιέχει τους ακόλουθους καταλόγους:

1. **kmeans** (kmeans/kmeans) – kmeans package με τον πηγαίο κώδικα της άσκησης
2. **data** (kmeans/data) – κατάλογος με αρχεία .dat με δεδομένα που παράγουν τα προγράμματα της άσκησης
3. **plots** (kmeans/plots) – κατάλογος με γραφήματα που παράγουν τα προγράμματα της άσκησης
4. **scripts** (kmeans/scripts) – gnuplot scripts που παράγουν και εκτελούν τα προγράμματα της άσκησης

Ο πηγαίος κώδικας περιέχει τις ακόλουθες κλάσεις:

1. **Point:** Περιγράφει ένα σημείο στο 2-διάστατο χώρο.
2. **Cluster:** Περιγράφει μία ομάδα κατά της εκτέλεση του αλγορίθμου.
3. **Loader:** Κλάση υπεύθυνη για την φόρτωση των δεδομένων από αρχείο. Δεν πραγματοποιεί ελέγχους για λάθη στο αρχείο. Θεωρεί πως τα δεδομένα στο αρχείο είναι στη σωστή μορφή.
4. **Plotter:** Κλάση υπεύθυνη για τα γραφήματα. Γράφει δεδομένα σε .dat αρχεία. Γράφει gnuplot scripts τα οποία και εκτελεί για να παράγει γραφήματα. Παράγει γραφήματα για τα τυχαία σημεία, για τα αποτελέσματα του αλγορίθμου και για τα αποτελέσματα του πειράματος.
5. **PointsGenerator:** Κλάση υπεύθυνη για την παραγωγή τυχαίων σημείων.
6. **KMeans:** Κλάση υπεύθυνη για την υλοποίηση του αλγορίθμου k-means.
7. **Experiment:** Κλάση υπεύθυνη για την υλοποίηση του πειράματος με πολλαπλές εκτελέσεις του αλγορίθμου k-means για διαφορετικό αριθμό ομάδων.

Παραγωγή τυχαίων σημείων

Η παραγωγή των τυχαίων σημείων γίνεται από την κλάση **PointsGenerator**. Τα τυχαία σημεία παράγονται σε αρχείο δεδομένων και σε αντίστοιχο γράφημα.

Compile command:
(από τον κατάλογο kmeans (root directory της άσκησης))
javac kmeans/PointsGenerator.java

Run command:
(από τον κατάλογο kmeans (root directory της άσκησης))
java kmeans/PointsGenerator

Outputs:
"**data/examples.dat**": Αρχείο δεδομένων με πληροφορία για τα τυχαία σημεία που δημιουργήθηκαν.
"**plots/examples.png**": Γράφημα με τα τυχαία σημεία που δημιουργήθηκαν.
"**scripts/examples.p**": gnuplot script το οποίο παράγει το γράφημα των σημείων.

Υλοποίηση του αλγόριθμου k-means

Η υλοποίηση του αλγορίθμου k-means γίνεται από την κλάση **KMeans**. Συγκεκριμένα, η κύρια μέθοδος υλοποίησης του αλγορίθμου είναι η **runKMeans()**. Τα αποτελέσματα αποθηκεύονται σε αρχεία δεδομένων και αντίστοιχα γραφήματα. Έχουμε τα δεδομένα της κάθε ομάδας σε ξεχωριστό αρχείο δεδομένων και τα κέντρα ολών των ομάδων συγκεντρωμένα σε ένα ακόμα αρχείο δεδομένων. Τα αρχεία αυτά χρησιμοποιούνται για την δημιουργία του γραφήματος με το τελικό αποτέλεσμα της εκτέλεσης του αλγορίθμου. Στο γράφημα βλέπουμε τα κέντρα των ομάδων καθώς και τα μέλη της κάθε ομάδας με διαφορετικό χρώμα προς οπτική διευκόλυνση.

Compile command:
(από τον κατάλογο kmeans (root directory της άσκησης))
javac kmeans/KMeans.java

Run command:
(από τον κατάλογο kmeans (root directory της άσκησης))
java kmeans/KMeans <number of clusters>

Όπου <number of clusters>: ο αριθμός των ομάδων (M) που θα φτιάξει ο αλγόριθμος.

Outputs:
"**plots/clusters.png**": Γράφημα με τις ομάδες που δημιούργησε ο αλγόριθμος.
"**data/cluster_*_members.dat**" όπου * είναι ο αριθμός της ομάδας: Αρχεία δεδομένων με τα μέλη της κάθε ομάδας.
"**data/cluster_centers.dat**": Αρχείο δεδομένων με τα κέντρα των ομάδων.
"**scripts/plot.p**": gnuplot script το οποίο παράγει το γράφημα με το τελικό αποτέλεσμα.

Πείραμα με πολλαπλές εκτελέσεις του αλγορίθμου k-means

Η κλάση **Experiment** είναι υπεύθυνη για πολλαπλλές εκτελέσεις του αλγόριθμου k-means για διαφορετικούς αριθμούς ομάδων (3, 5, 7, 9, 11 και 13). Για κάθε αριθμό ομάδων κρατάει την εκτέλεση με το μικρότερο σφάλμα ομαδοποίησης και παράγει το γράφημα με τα αποτελέσματα της. Τέλος, αφού έχουν τελειώσει όλες οι εκτελέσεις του αλγορίθμου, παράγει το γράφημα μεταβολής του (ελάχιστου) σφάλματος ομαδοποίησης ανά αριθμό ομάδων.

Compile command:
(από τον κατάλογο kmeans (root directory της άσκησης))
javac kmeans/Experiment.java

Run command:
(από τον κατάλογο kmeans (root directory της άσκησης))
java kmeans/Experiment <reps>

Όπου <reps>: ο αριθμός που θα εκτελεστεί ο αλγόριθμος για κάθε αριθμό ομάδων.

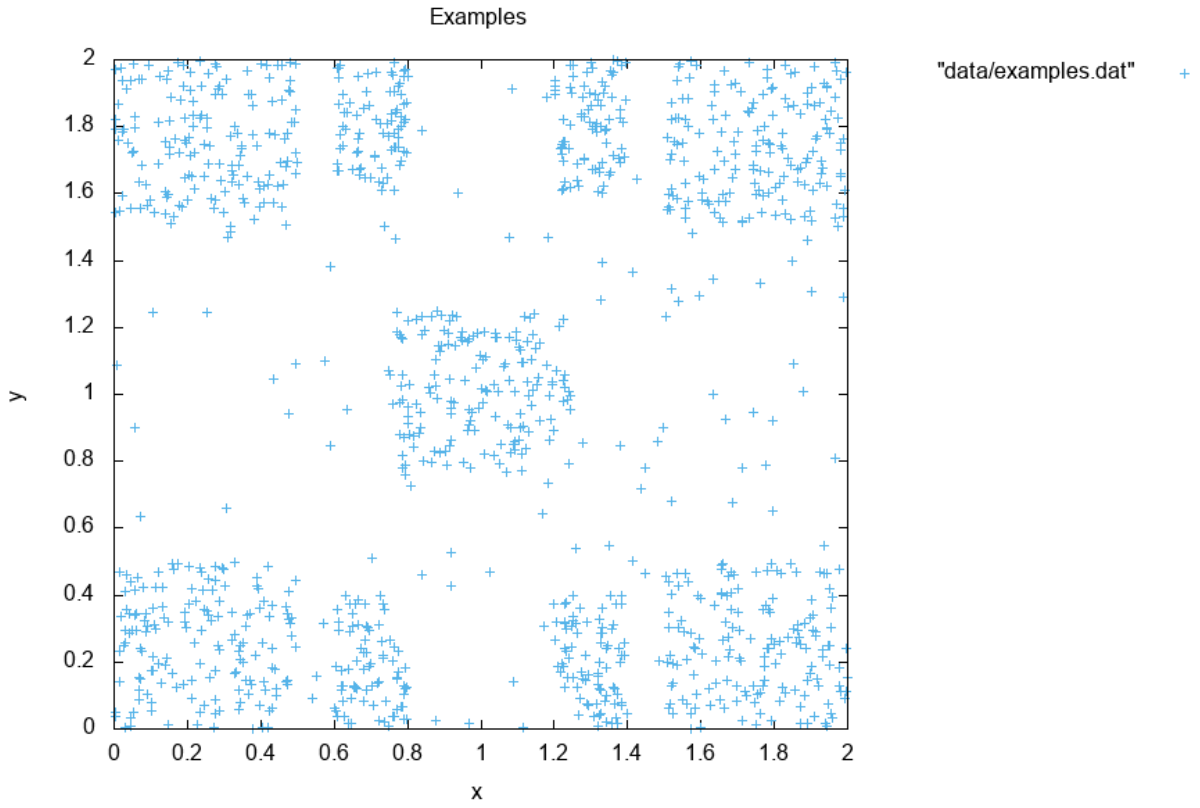
Outputs:
“plots/clusters_3.png”: Γράφημα για την εκτέλεση για 3 ομάδες με το ελάχιστο σφάλμα.
“plots/clusters_5.png”: Γράφημα για την εκτέλεση για 5 ομάδες με το ελάχιστο σφάλμα.
“plots/clusters_7.png”: Γράφημα για την εκτέλεση για 7 ομάδες με το ελάχιστο σφάλμα.
“plots/clusters_9.png”: Γράφημα για την εκτέλεση για 9 ομάδες με το ελάχιστο σφάλμα.
“plots/clusters_11.png”: Γράφημα για την εκτέλεση για 11 ομάδες με το ελάχιστο σφάλμα.
“plots/clusters_13.png”: Γράφημα για την εκτέλεση για 13 ομάδες με το ελάχιστο σφάλμα.
“plots/results.png”: Γράφημα μεταβολής ελάχιστου σφάλματος ανά αριθμό ομάδων.
“pltos/experiment_results.dat”: Αρχείο δεδομένων με το ελάχιστο σφάλμα ομαδοποίησης ανά αριθμό ομάδων.
“scripts/results.p”: gnuplot script το οποίο παράγει το γράφημα μεταβολής του ελάχοστπυ σφάλματος ανά αριθμό ομάδων.

Σημειώσεις:

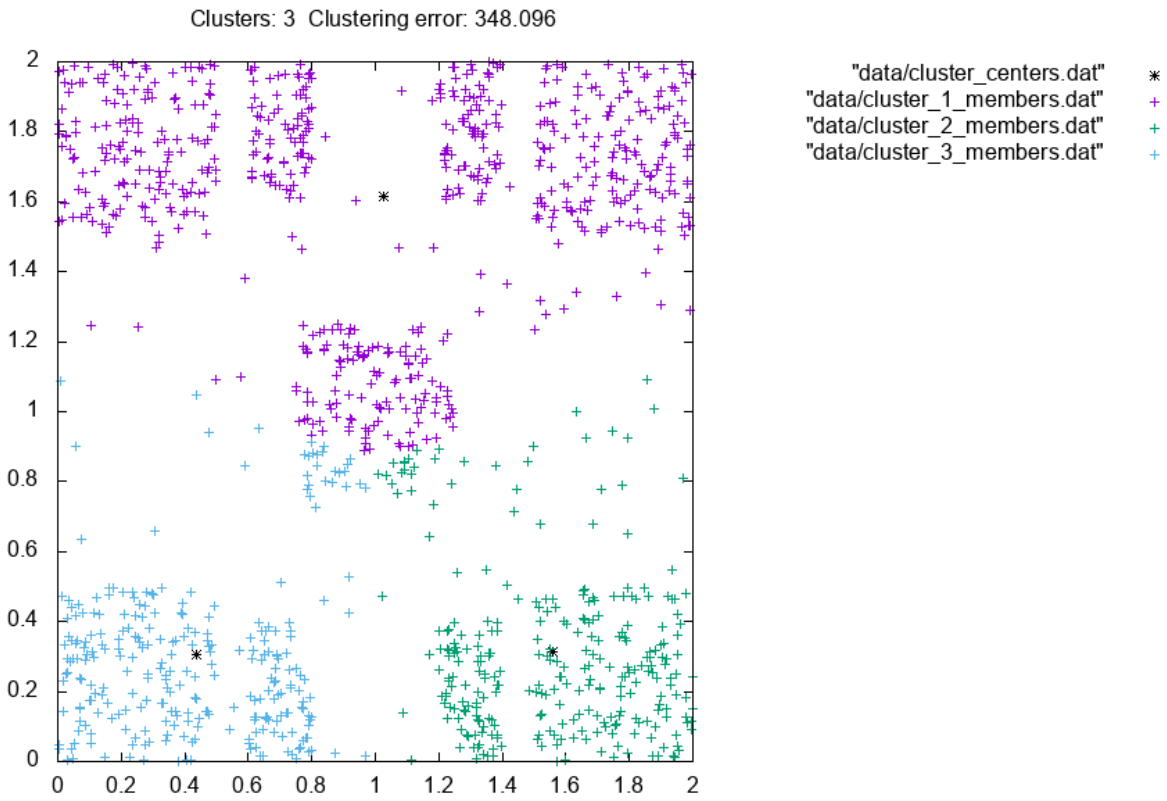
- Η κλάση φορτώνει τα δεδομένα απο το αρχείο **data/examples.dat**.
- Η κλάση χρησιμοποιεί τα ακόλουθα αρχεία για να παράγει τα γραφήματα για τους διαφορετικούς αριθμούς ομάδων: “scripts/plot.p”, “data/cluster_*_members.dat” όπου * είναι ο αριθμός ομάδας, “data/cluster_centers.dat”. Τα αρχεία αυτά γράφονται πολλές

φορές κατά την εκτέλεση του πειράματος (καθώς ο αλγόριθμος εκτελείται πολλές φορες). Επομένως, δεν είναι ιδιαίτερα χρήσιμα καθώς περιέχουν πληροφορία μόνο για το τελευταίο γράφημα.

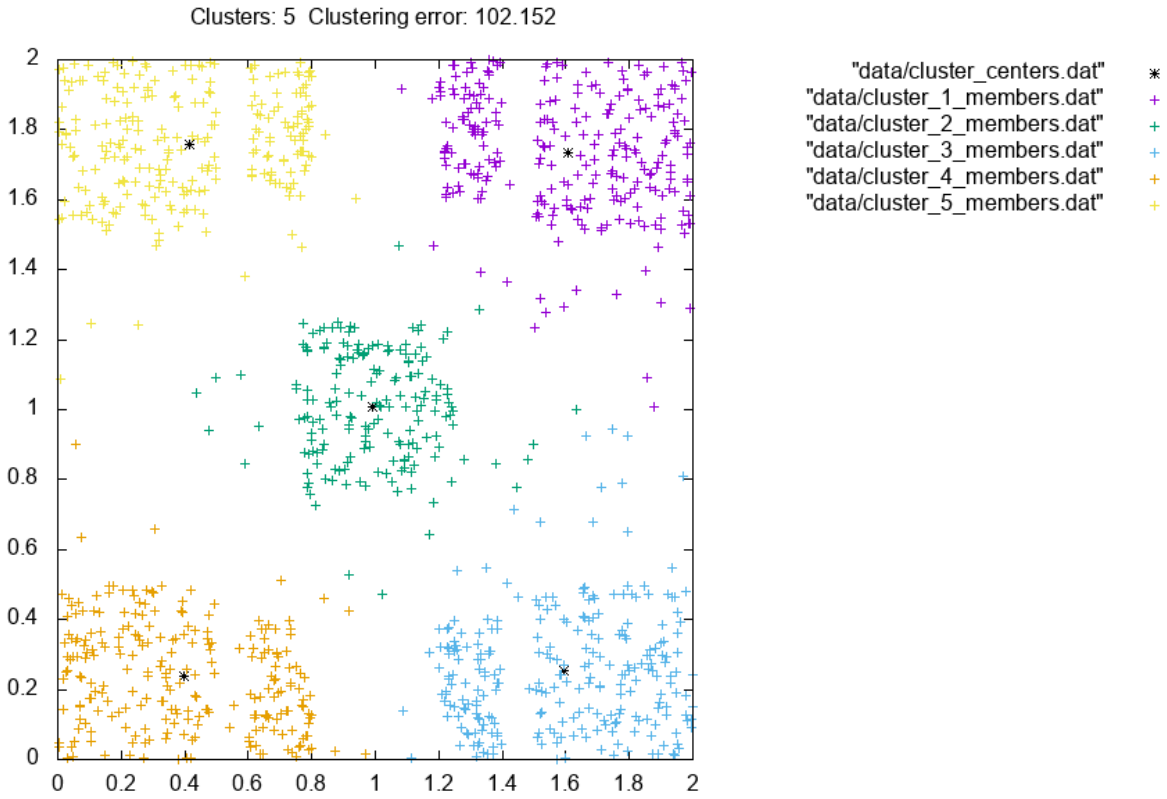
Ενδεικτικές εικόνες αποτελεσμάτων



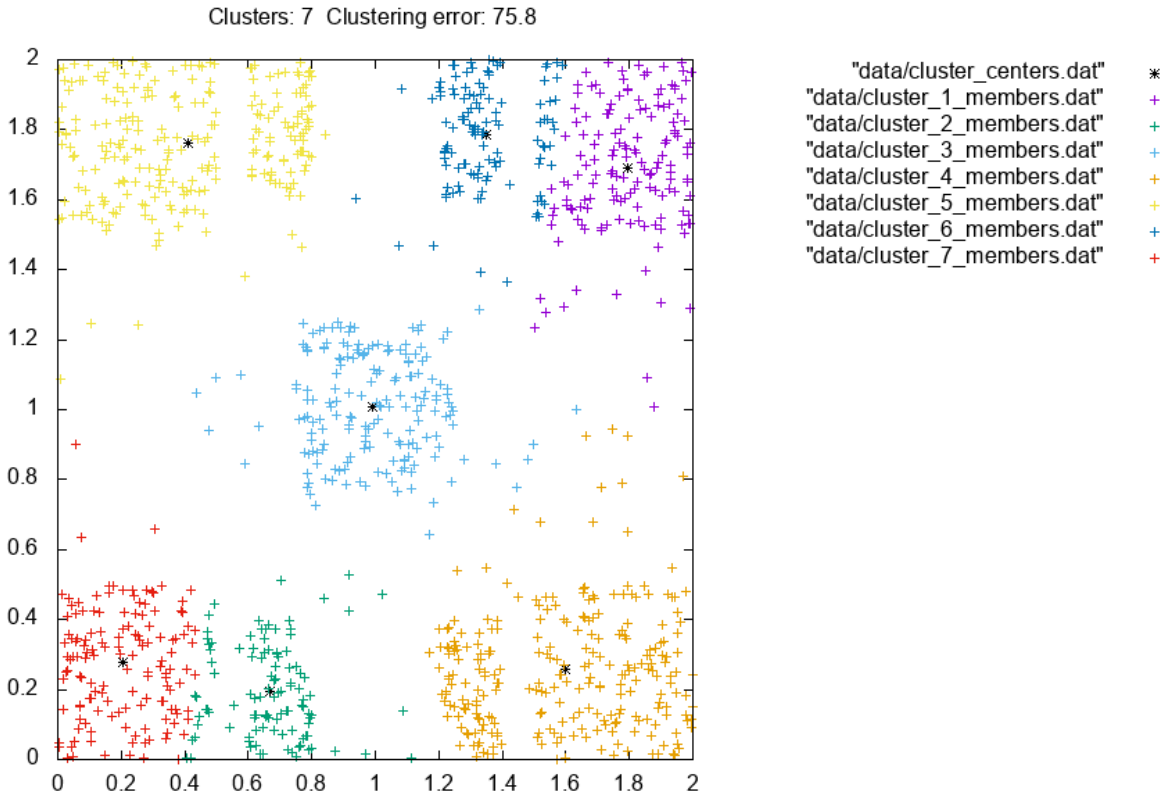
Τυχαία σημεία (1200 παραδείγματα)



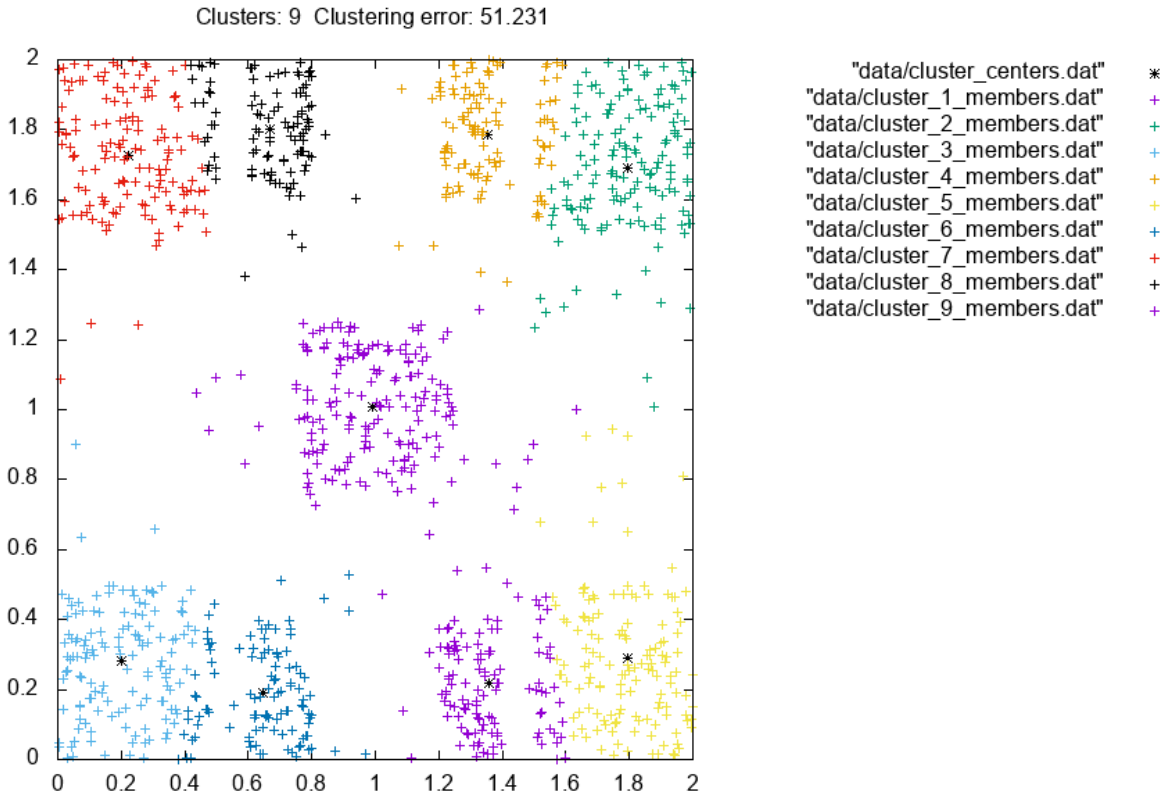
Αποτέλεσμα εκτέλεσης με το ελάχιστο σφάλμα ομαδοποίησης για 3 ομάδες



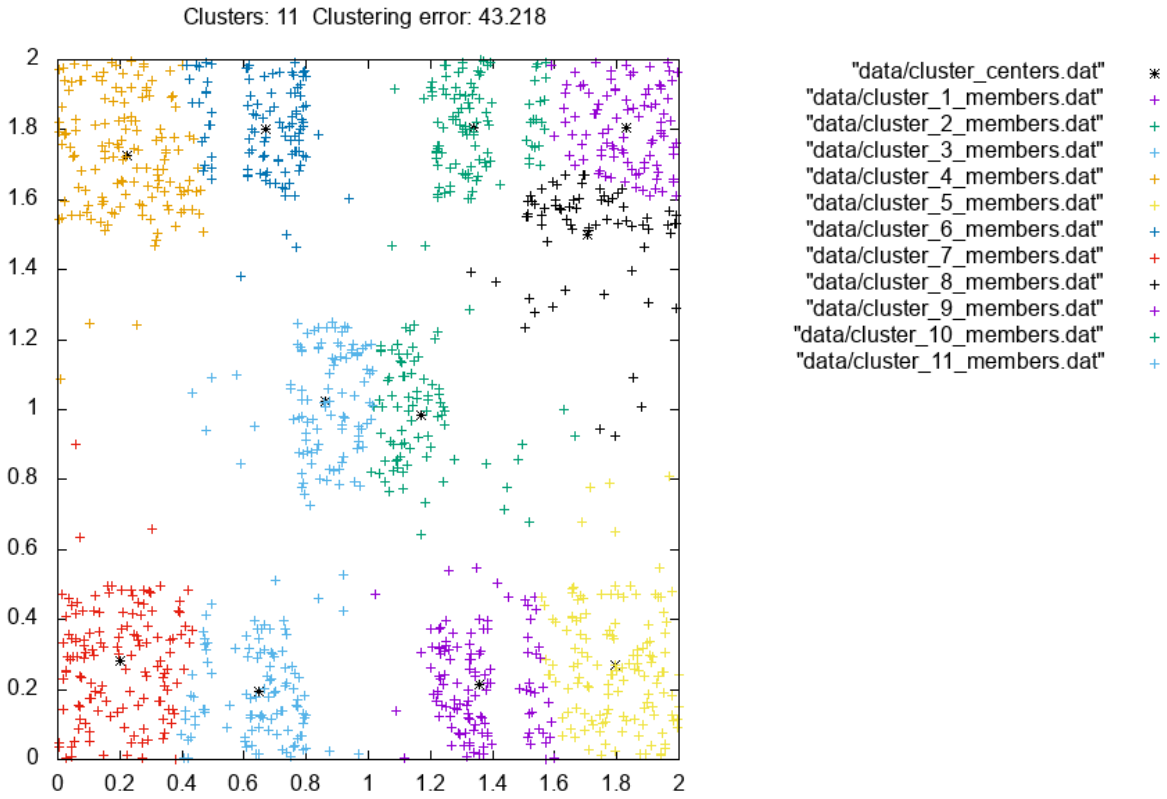
Αποτέλεσμα εκτέλεσης με το ελάχιστο σφάλμα ομαδοποίησης για 5 ομάδες



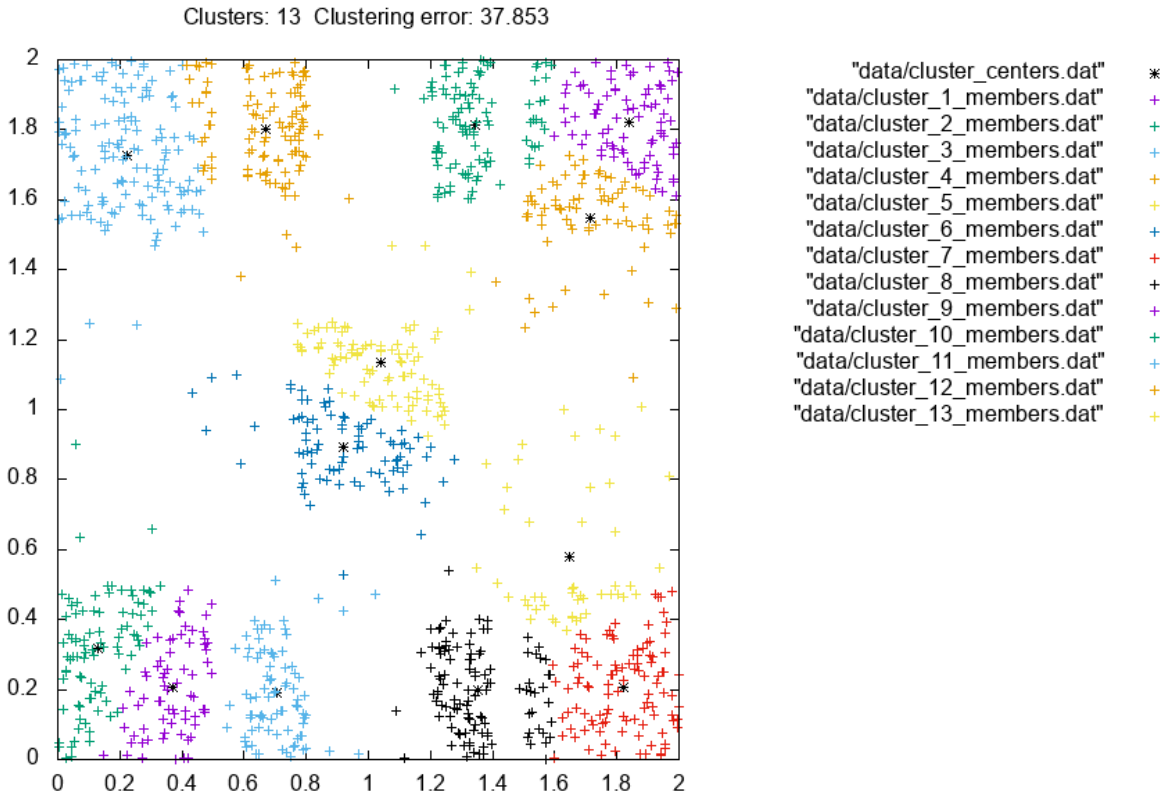
Αποτέλεσμα της εκτέλεσης με το ελάχιστο σφάλμα ομαδοποίησης για 7 ομάδες



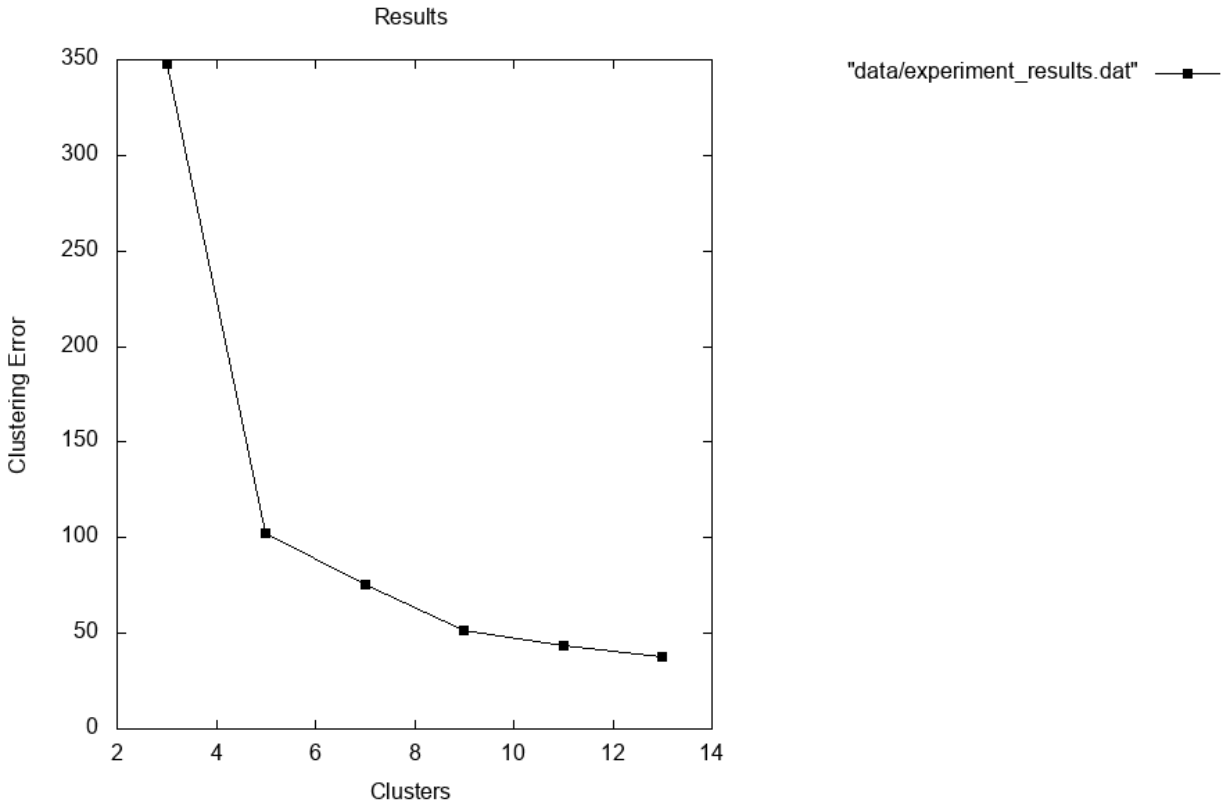
Αποτέλεσμα της εκτέλεσης με το ελάχιστο σφάλμα ομαδοποίησης για 9 ομάδες



Αποτέλεσμα της εκτέλεσης με το ελάχιστο σφάλμα ομαδοποίησης για 11 ομάδες



Αποτέλεσμα της εκτέλεσης με το ελάχιστο σφάλμα ομαδοποίησης για 13 ομάδες



Διάγραμμα μεταβολής σφάλματος ομαδοποίησης ανά αριθμό ομάδων

Συμπέρασμα

Στο διάγραμμα μεταβολής του σφάλματος ομαδοποίησης ανά αριθμό ομάδων παρατηρούμε πως μετά το σφάλμα ομαδοποίησης μειώνεται όσο αυξάνεται ο αριθμός ομάδων. Για αριθμό ομάδων μικρότερο του 9 ο ρυθμός μείωσης είναι μεγάλος, ενώ για αριθμό ομάδων μεγαλύτερο του 9 ο ρυθμός μείωσης είναι πολύ πιο μικρός. Άρα, το διάγραμα μεταβολής του σφάλματος ομαδοποίησης μπορεί να χρησιμοποιηθεί για να προσεγγίσουμε τον πραγματικό αριθμό ομάδων. Αρκεί να ψάξουμε το σημείο (αριθμό ομάδων) στο διάγραμμα όπου ο ρυθμός μείωσης του σφάλματος ελαττώνεται σημαντικά, στην περίπτωση μας το 9.