ΑΠΟΤΕΛΕΣΜΑΤΑ ΑΣΚΗΣΗΣ ΚΜΕΑΝS ΑΛΓΟΡΙΘΜΟΥ

ΜΕΛΗ ΤΗΣ ΟΜΑΔΑΣ

Αγνή Παΐλα, ΑΜ: 4753

Νικολέτα Μπεράτη, ΑΜ: 4884

Μάθημα: Υπολογιστική Νοημοσύνη

Έτος: 2023-2024

Στην παρούσα άσκηση υλοποιούμε πρόγραμμα ομαδοποίησης (ΠΟ) με Μ ομάδες βασισμένο στον αλγόριθμο k-means.

Το πρόγραμμα ομαδοποίησης που κατασκευάσαμε δομείται από 3 κλάσεις:

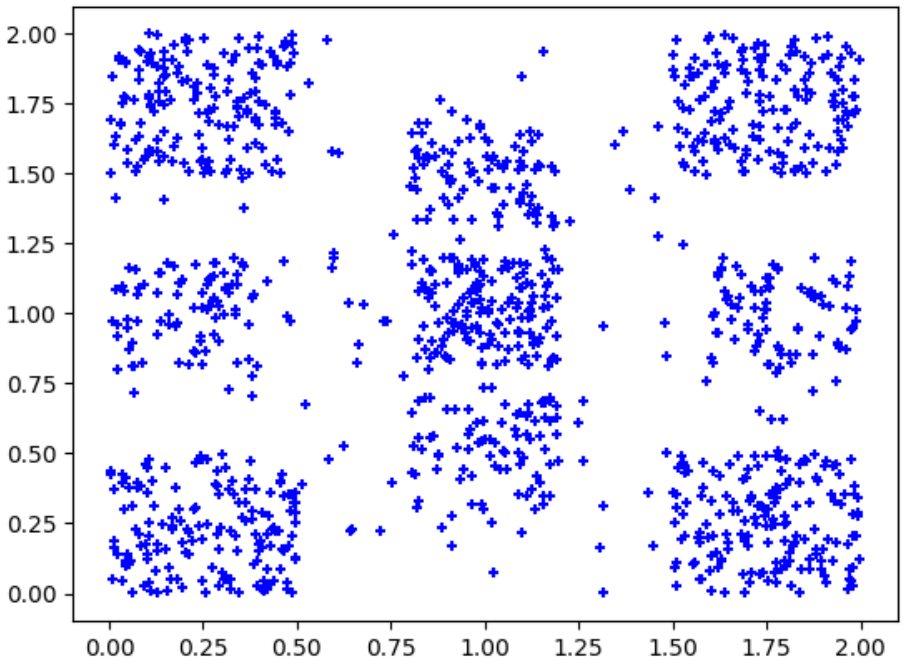
- Την κλάση ***KMeans*** η οποία περιέχει την main συνάρτηση του προγράμματος και υλοποιεί όλα τα βασικά βήματα του αλγορίθμου, όπως την αρχικοποίηση, την ανάθεση σημείων σε κάθε cluster, την ανανέωση των κέντρων κλπ.
- Την κλάση ***Cluster*** η οποία αναπαριστά τις ομάδες του προβλήματος. Κρατάει πληροφορίες σχετικά με τα σημεία που ανήκουν σε αυτήν, το κέντρο της και τον αριθμό της ομάδας που αντιπροσωπεύει.
- Την κλάση ***PointOfCluster*** η οποία αναπαριστά ένα σημείο-παράδειγμα της μορφής [x1, x2], όπως το εξάγουμε από το αρχείο με τα παραδείγματά μας (data.csv).

Για την μεταγλώττιση και την εκτέλεση του προγράμματος, είναι απαραίτητο να εκτελεστούν οι παρακάτω εντολές:

- **javac *.java**
- **java KMeans**

***Σημείωση:** *Αφού πραγματοποιηθεί η μεταγλώττιση, στην σπάνια περίπτωση που δεν ολοκληρωθεί η εκτέλεση του προγράμματος, χρειάζεται να εκτελεστεί ξανά η εντολή java KMeans.*

Το αρχείο που περιέχει τα δεδομένα μας, δηλαδή τα 1200 παραδείγματα, είναι σε μορφή .csv (data.csv) και δημιουργήθηκε μέσα από μια java κλάση. Προκειμένου να το οπτικοποιήσουμε, δημιουργήσαμε ένα αρχείο Python, του οποίου η εκτέλεση απεικονίζεται ως εξής:



Μέσα από την main του προγράμματος εκτελούμε το πρόγραμμα ομαδοποίησης (ΠΟ) στο σύνολο δεδομένων (ΣΔΟ) για ***M = 3, 6, 9, 12*** ομάδες.

Για κάθε διαφορετικό αριθμό ομάδων Μ, εκτελούμε 15 τρεξίματα του προγράμματος και τυπώνουμε το σφάλμα ομαδοποίησης (*Clustering Error*) για κάθε ένα από αυτά. Βρίσκουμε και επιλέγουμε να κρατήσουμε ως λύση, αυτή που έχει το μικρότερο σφάλμα ομαδοποίησης κάθε φορά, το οποίο και τυπώνουμε στο τέλος της εκτέλεσης.

Επιπλέον, επιλέξαμε να τυπώνουμε την κατάσταση στην οποία καταλήγει το πρόγραμμα ομαδοποίησης, αφού εκτελέσει τον Kmeans αλγόριθμο μέχρι να τερματίσει. Έτσι φαίνεται ξεκάθαρα για κάθε ομάδα

- ο αριθμός της
- το σημείο (x1, x2) που έχει οριστεί τελικά ως κέντρο της
- το πλήθος των σημείων που της έχουν ανατεθεί

Αφού τερματίσουν οι εκτελέσεις για κάθε αριθμό ομάδων Μ, τυπώνουμε συγκεντρωτικά το σφάλμα ομαδοποίησης (*Clustering Error*) που προέκυψε για κάθε μία από αυτές.

ΑΛΓΟΡΙΘΜΟΣ K-MEANS

Ο **αλγόριθμος kmeans** που υλοποιήσαμε, αρχικοποιεί τα κέντρα των ομάδων σε τυχαία παραδείγματα από το σύνολο των παραδειγμάτων που έχουμε παράγει.

Έπειτα αναθέτει σημεία σε κάθε ομάδα, ανάλογα με την ευκλείδεια απόστασή τους από τα κέντρα.

Εφόσον έχουν ανατεθεί όλα τα παραδείγματα σε ομάδες, μετακινούμε τα κέντρα σε νέο σημείο, το οποίο προκύπτει υπολογίζοντας τον μέσο όρο των συντεταγμένων των σημείων της κάθε ομάδας.

Στη συνέχεια υπολογίζουμε πόσο έχουν μετακινηθεί τα κέντρα σε σχέση με την προηγούμενη θέση τους. Εάν η απόσταση αυτή είναι μικρότερη από ένα κατώφλι που ορίζουμε (*terminateThreshold*), τότε ο αλγόριθμος τερματίζει και επιστρέφει το συνολικό σφάλμα ομαδοποίησης.

ΕΚΤΕΛΕΣΗ

Παρακάτω παρουσιάζουμε την εκτέλεση του προγράμματός μας, για τον διαφορετικό αριθμό ομάδων.

Ορίσαμε **terminateThreshold = 0.1** .

```
****************************** FOR 3 CLUSTERS ******************************

For iteration 1
Clustering Error: 621.2875745557487

For iteration 2
Clustering Error: 622.2303056459383

For iteration 3
Clustering Error: 623.3004450990011

For iteration 4
Clustering Error: 624.0298865411229

For iteration 5
Clustering Error: 624.6156786630197

For iteration 6
Clustering Error: 625.1373755187162

For iteration 7
Clustering Error: 625.2319302298064

For iteration 8
Clustering Error: 625.2319302298064

For iteration 9
Clustering Error: 625.2319302298064

For iteration 10
Clustering Error: 625.2319302298064

For iteration 11
Clustering Error: 625.2319302298064

For iteration 12
Clustering Error: 625.2319302298064

For iteration 13
Clustering Error: 625.2319302298064

For iteration 14
Clustering Error: 625.2319302298064

For iteration 15
Clustering Error: 625.2319302298064


              Final State Of Clusters
_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _
Cluster: 0
Centroid: [0.4527954281994028,1.5119086216417912]
Points: 335
Cluster: 1
Centroid: [0.46767415467303364,0.3752113908018726]
Points: 267
Cluster: 2
Centroid: [1.5557640619899662,0.9866700665076912]
Points: 598
_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _

<------------------------------------------------------------------>
Minimum clustering error after 15 iterations is : 621.2875745557487
<------------------------------------------------------------------>
```

```
****************************** FOR 6 CLUSTERS ******************************

For iteration 1
Clustering Error: 346.46495311407637

For iteration 2
Clustering Error: 345.5330494769651

For iteration 3
Clustering Error: 344.0354296534215

For iteration 4
Clustering Error: 333.7286873376988

For iteration 5
Clustering Error: 333.70236080398416

For iteration 6
Clustering Error: 333.6886264737377

For iteration 7
Clustering Error: 333.6827422145001

For iteration 8
Clustering Error: 333.67492894379006

For iteration 9
Clustering Error: 333.6662262510719

For iteration 10
Clustering Error: 333.6662262510719

For iteration 11
Clustering Error: 333.6662262510719

For iteration 12
Clustering Error: 333.6662262510719

For iteration 13
Clustering Error: 333.6662262510719

For iteration 14
Clustering Error: 333.6662262510719

For iteration 15
Clustering Error: 333.6662262510719


              Final State Of Clusters
_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _
Cluster: 0
Centroid: [1.7496979014285723,0.3671104824695239]
Points: 210
Cluster: 1
Centroid: [1.7390307387096777,1.5853870400921661]
Points: 217
Cluster: 2
Centroid: [0.24437186103406114,1.5637986430131006]
Points: 229
Cluster: 3
Centroid: [1.016188022359551,0.7035263062359546]
Points: 178
Cluster: 4
Centroid: [0.26931936940582024,0.3121321068100528]
Points: 189
Cluster: 5
Centroid: [0.9850404646327678,1.3177785333333336]
Points: 177
_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _

<------------------------------------------------------------------>
Minimum clustering error after 15 iterations is : 333.6662262510719
<------------------------------------------------------------------>
```

********** FOR 9 CLUSTERS **********

For iteration 1
Clustering Error: 257.41427987658307

For iteration 2
Clustering Error: 257.1902624379905

For iteration 3
Clustering Error: 257.06087636557413

For iteration 4
Clustering Error: 257.0074563266205

For iteration 5
Clustering Error: 256.9788370785236

For iteration 6
Clustering Error: 256.8281736834717

For iteration 7
Clustering Error: 256.73264317956546

For iteration 8
Clustering Error: 256.73264317956546

For iteration 9
Clustering Error: 256.73264317956546

For iteration 10
Clustering Error: 256.73264317956546

For iteration 11
Clustering Error: 256.73264317956546

For iteration 12
Clustering Error: 256.73264317956546

For iteration 13
Clustering Error: 256.73264317956546

For iteration 14
Clustering Error: 256.73264317956546

For iteration 15
Clustering Error: 256.73264317956546

Final State Of Clusters

Cluster: 0
Centroid: [0.9914160578494627,1.5054010602150543]
Points: 93
Cluster: 1
Centroid: [1.025277296145833,0.5025561773958334]
Points: 96
Cluster: 2
Centroid: [1.7683247758241754,0.9971250334065938]
Points: 91
Cluster: 3
Centroid: [1.6171241462365582,1.7422099860215052]
Points: 93
Cluster: 4
Centroid: [0.9821263282840234,1.0094737591124263]
Points: 169
Cluster: 5
Centroid: [1.7455233619883048,0.2599464656643274]
Points: 171
Cluster: 6
Centroid: [1.8799226749999995,1.7395003263888895]
Points: 72
Cluster: 7
Centroid: [0.24381708349473685,1.566440827412281]
Points: 228
Cluster: 8
Centroid: [0.2650020897737969,0.3114625022839571]
Points: 187

Minimum clustering error after 15 iterations is : 256.73264317956546

********** FOR 12 CLUSTERS **********

For iteration 1
Clustering Error: 202.0679491797501

For iteration 2
Clustering Error: 201.86594428941754

For iteration 3
Clustering Error: 201.7231877320224

For iteration 4
Clustering Error: 201.60691194352896

For iteration 5
Clustering Error: 201.58517348020678

For iteration 6
Clustering Error: 201.58695055987795

For iteration 7
Clustering Error: 201.58695055987795

For iteration 8
Clustering Error: 201.58695055987795

For iteration 9
Clustering Error: 201.58695055987795

For iteration 10
Clustering Error: 201.58695055987795

For iteration 11
Clustering Error: 201.58695055987795

For iteration 12
Clustering Error: 201.58695055987795

For iteration 13
Clustering Error: 201.58695055987795

For iteration 14
Clustering Error: 201.58695055987795

For iteration 15
Clustering Error: 201.58695055987795

Final State Of Clusters

Cluster: 0
Centroid: [0.2516978993177915,1.7526746251533745]
Points: 163
Cluster: 1
Centroid: [0.2674201421833333,1.1120024516666664]
Points: 60
Cluster: 2
Centroid: [1.7468339119047618,0.3969378985714286]
Points: 84
Cluster: 3
Centroid: [0.9873823887234047,1.506120811702128]
Points: 94
Cluster: 4
Centroid: [0.1866183578333333,0.8695838644444442]
Points: 36
Cluster: 5
Centroid: [1.7391563752808983,0.1339610078494382]
Points: 89
Cluster: 6
Centroid: [1.6780690791666661,1.5940162041666672]
Points: 72
Cluster: 7
Centroid: [0.27239336243913054,0.22742385936086945]
Points: 161
Cluster: 8
Centroid: [1.7721959629213482,0.9978770869662925]
Points: 89
Cluster: 9
Centroid: [1.7700889734042546,1.84867369893617]
Points: 94
Cluster: 10
Centroid: [1.0223358339999995,0.5061400061052631]
Points: 95
Cluster: 11
Centroid: [0.9991854342944783,1.0067094030061352]
Points: 163

Minimum clustering error after 15 iterations is : 201.58517348020678

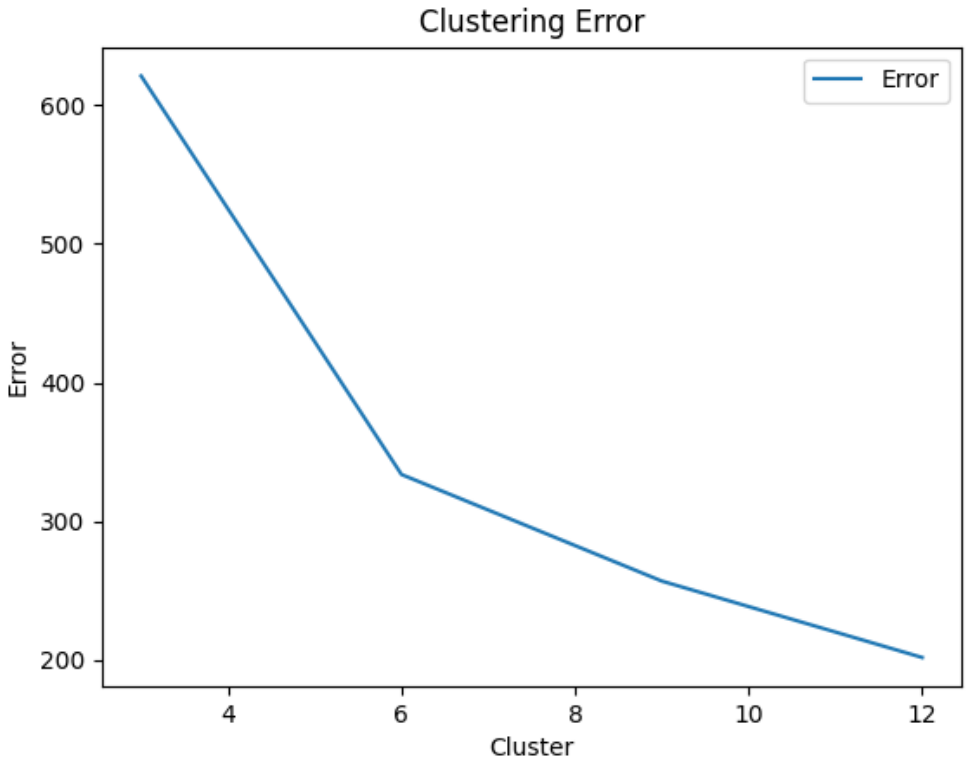
```
For 3 clusters : clustering error = 621.2875745557487
For 6 clusters : clustering error = 333.6662262510719
For 9 clusters : clustering error = 256.73264317956546
For 12 clusters : clustering error = 201.58517348020678
```

ΑΠΟΤΕΛΕΣΜΑΤΑ

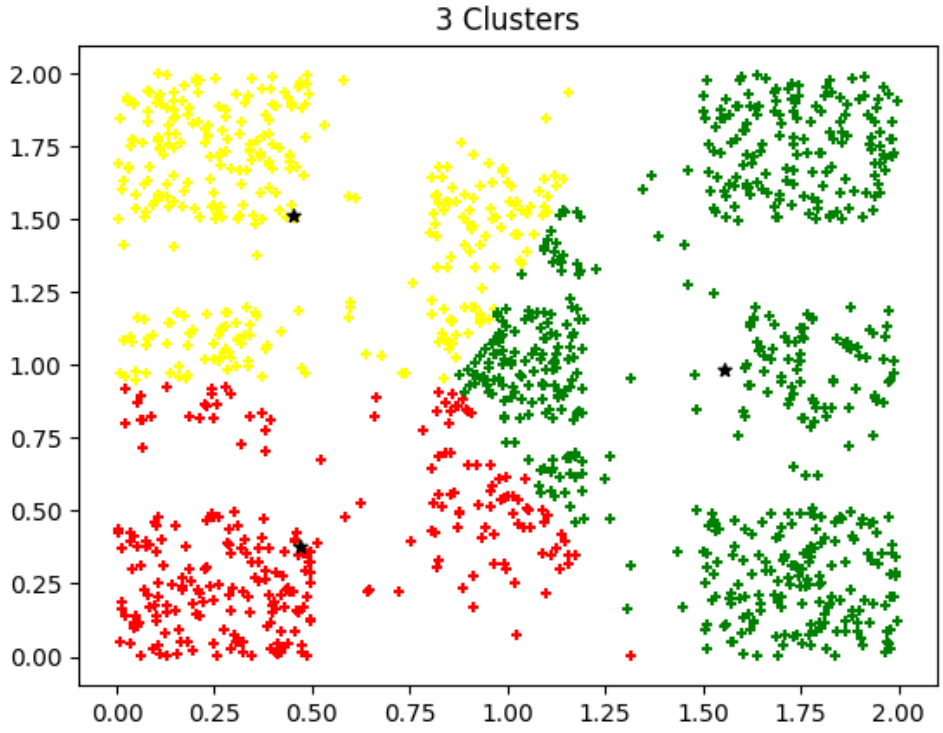
Για τη λύση με το μικρότερο σφάλμα ομαδοποίησης που έχουμε βρει, φτιάχνουμε αρχεία .csv για κάθε ομάδα, τα οποία περιέχουν τα σημεία που της ανατέθηκαν. Επιπλέον, φτιάχνουμε αρχείο .csv το οποίο περιέχει τα κέντρα για κάθε ομάδα.

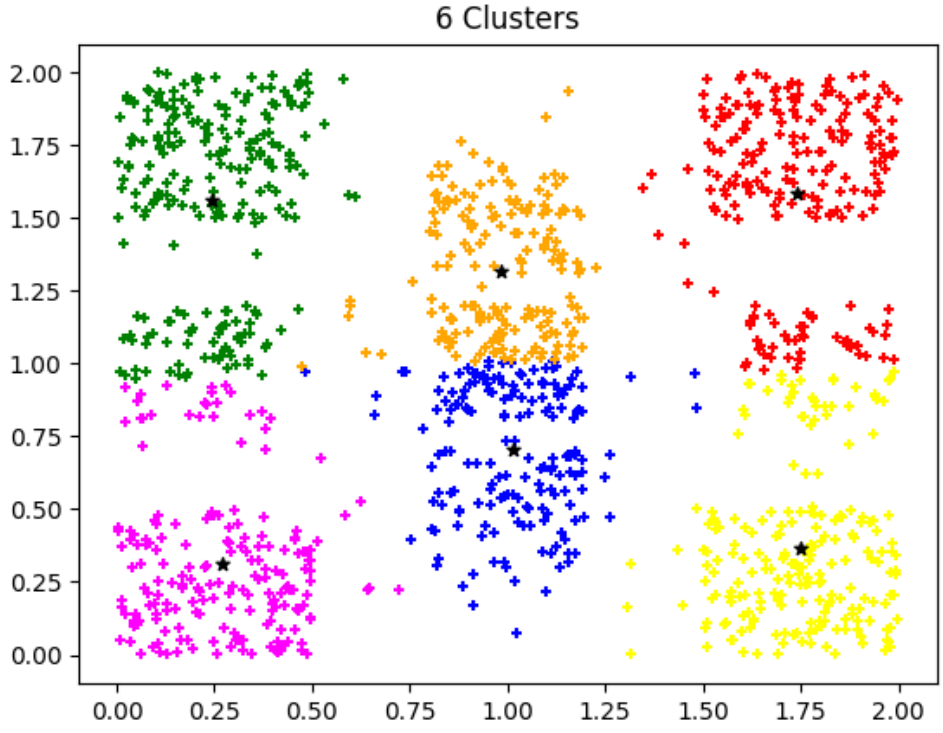
Έτσι, είμαστε σε θέση να εμφανίσουμε με plot στο ίδιο σχήμα τόσο τα παραδείγματα (+) όσο και τις θέσεις των κέντρων που βρήκαμε (*). Όλα τα σημεία της κάθε ομάδας εμφανίζονται με το ίδιο χρώμα για να τα ξεχωρίζουμε, ενώ τα κέντρα των ομάδων με μαύρο χρώμα.

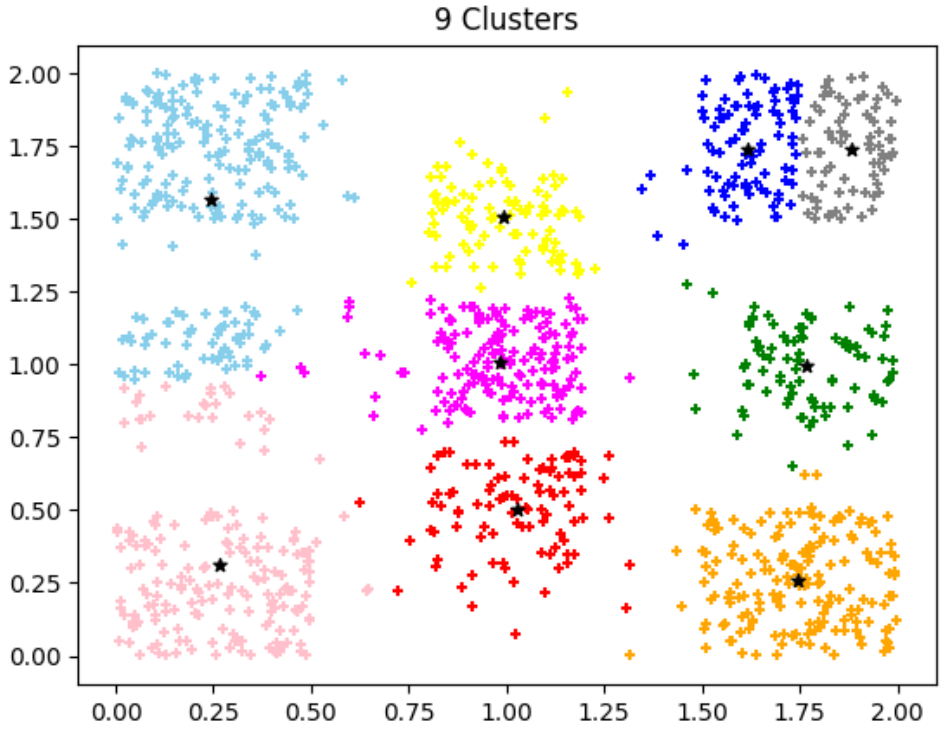
Απεικονίζουμε παρακάτω το διάγραμμα που παρουσιάζει πως μεταβάλλεται το σφάλμα ομαδοποίησης όσο αυξάνεται ο αριθμός των ομάδων:

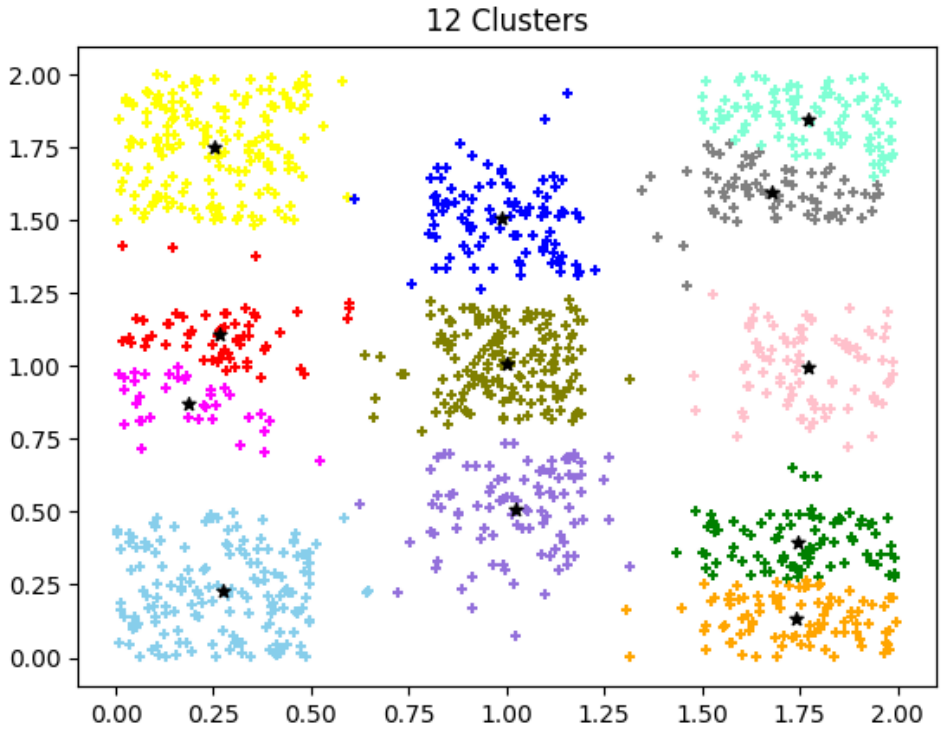


Επίσης σε ένα σχήμα, απεικονίζουμε την ανάθεση των παραδειγμάτων σε ομάδες (+), και τα κέντρα για αυτές (*) :






9 Clusters
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00


12 Clusters
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00

ΣΥΜΠΕΡΑΣΜΑΤΑ

Με βάση τα αποτελέσματα του προγράμματος και λαμβάνοντας υπόψη το διάγραμμα που απεικονίζει την μεταβολή του σφάλματος ομαδοποίησης συναρτήσει του αριθμού των ομάδων, συμπεραίνουμε ότι το σφάλμα ομαδοποίησης δεν μπορεί να χρησιμοποιηθεί για να εκτιμήσουμε τον πραγματικό αριθμό των ομάδων που έχει το πρόβλημα ομαδοποίησης.

Στο γράφημα που απεικονίζει το σφάλμα, δηλαδή την διασπορά, παρατηρούμε ότι όσο αυξάνεται ο αριθμός των ομάδων Μ, το σφάλμα-διασπορά μειώνεται. Έτσι παρόλο που οι ομάδες του προβλήματός μας στην πραγματικότητα είναι 9, για Μ=12 το σφάλμα-διασπορά είναι ακόμη μικρότερο. Άρα δεν είναι εφικτό να προβλέψουμε τον ακριβή αριθμό ομάδων, μόνο από το συγκεκριμένο γράφημα.