ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ

Εργασία 3

ΠΑΝΤΕΛΕΗΜΩΝ ΠΡΩΙΟΣ
ice18390023
7ο Εξάμηνο
ice18390023@uniwa.gr

Τμήμα Τρίτης 9:00-13:00

ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ
UNIVERSITY OF WEST ATTICA

Υπεύθυνοι καθηγητές

ΠΑΡΙΣ ΜΑΣΤΟΡΟΚΩΣΤΑΣ

Τμήμα Μηχανικών και Πληροφορικής Υπολογιστών
22 Ιανουαρίου 2022

Περιεχόμενα

Κώδικες

Κατάλογος σχημάτων

1 Σύνολο δεδομένων του Enron

Στην ακόλουθη ενότητα, θα γίνει συσταδοποίηση στο σύνολο δεδομένων των email του Enron.

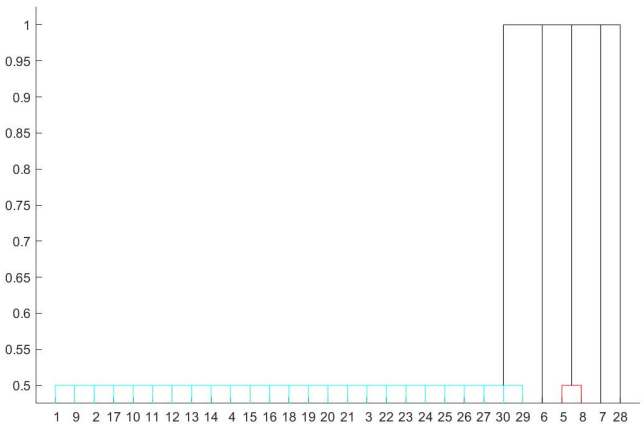
1.1 Δενδόγραμμα για διαφορετικούς δεσμούς και δείκτες ομοιότητας

Το script 1.1, κάνει συσταδοποίηση της δεύτερης και τρίτης στήλης των δεδομένων με:

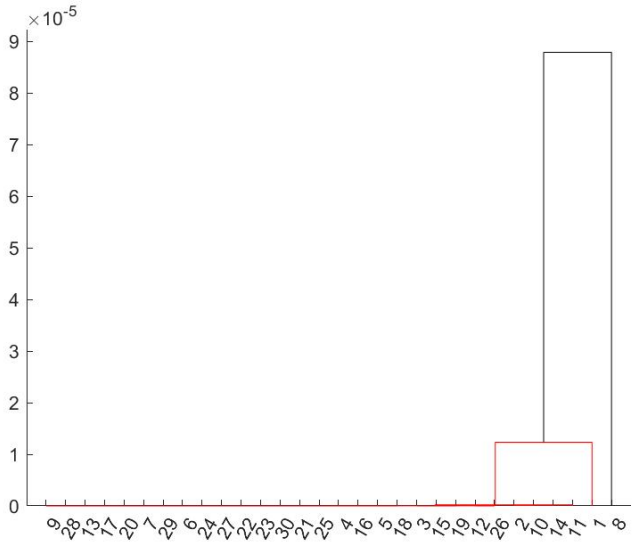
- απλό δεσμό και ομοιότητα jaccard
- απλό δεσμό και ομοιότητα cosine
- μέσο δεσμό και ομοιότητα cosine

και εμφανίζει το δενδρόγραμμα για κάθε ένα από αυτά.

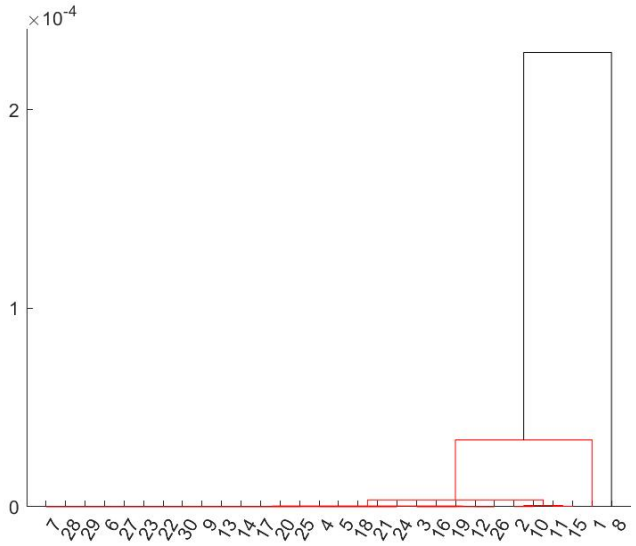
Είναι εμφανές σε ποίο επίπεδο θα γίνει η διαφοροποίηση σε κάθε ένα από τα τρία δενδρογράμματα. Ωστόσο, στο τρίτο δενδρόγραμμα (εικ. 1.3), αντί για τρεις συστάδες θα μπορούσαν να υπάρχουν τέσσερις. Επίσης, η μέθοδος που φαίνεται πως χωρίζει καλύτερα τα δεδομένα που έχουν περισσότερα κοινά, είναι με τον μέσο δεσμό και την ομοιότητα cosine.



Σχήμα 1.1: Απλός δεσμός με ομοιότητα jaccard



Σχήμα 1.2: Απλός δεσμός με ομοιότητα cosine



Σχήμα 1.3: Μέσος δεσμός με ομοιότητα cosine

```
load('enron100.mat')

d = pdist(en2(1:100,2:3),'jaccard');
Z = linkage(d); % single
[H,T] = dendrogram(Z,'ColorThreshold', 'default');

```

```
%print('images/single_jaccard','-djpeg')

figure
d = pdist(en2(1:100,2:3),'cosine');
Z = linkage(d); % single
[H,T] = dendrogram(Z,'ColorThreshold', 'default');
xtickangle(60);

%print('images/single_cosine','-djpeg')

figure
d = pdist(en2(1:100,2:3),'cosine');
Z = linkage(d, 'average');
[H,T] = dendrogram(Z,'ColorThreshold', 'default');
xtickangle(60);

%print('images/average_cosine','-djpeg')
```

Κώδικας 1.1: Script συσταδοποίησης των email του Enron

2 Διαμεριστική συσταδοποίηση με k-means

Σε αυτήν την ενότητα, θα γίνει συσταδοποίηση με k-means για διάφορες μεθόδους στα δεδομένα Iris και xV.

2.1 Διαχωριστική συσταδοποίηση για το σύνολο δεδομένων Iris

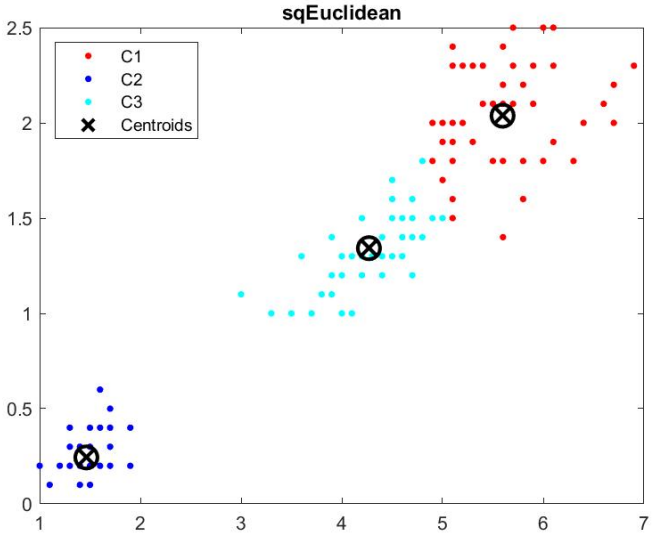
Στο script 2.1, υπολογίζεται ο k-means με τους εξής διαφορετικούς τρόπους υπολογισμού:

- Euclidean
- cityBlock
- cosine
- correlation

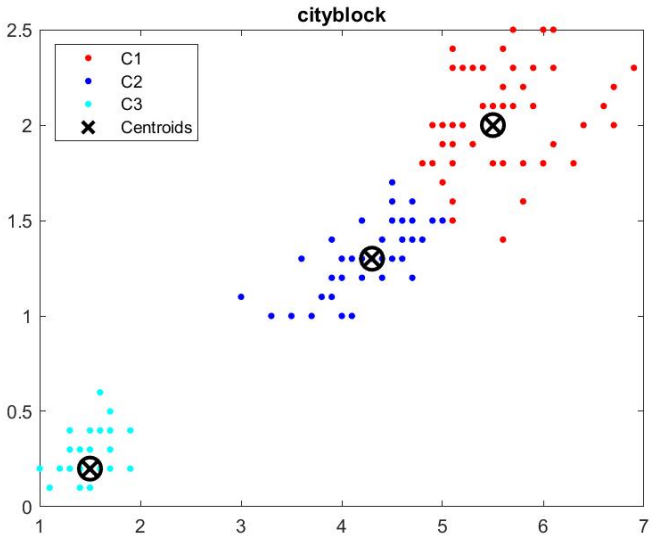
για κάθε ένα από τους εξής συνδιασμούς στηλών των δεδομένων Iris:

- τρίτη και τέταρτη
- πρώτη και δεύτερη
- πρώτη, δεύτερη, τρίτη και τέταρτη
- πρώτη, δεύτερη και τρίτη
- δεύτερη, τρίτη και τέταρτη

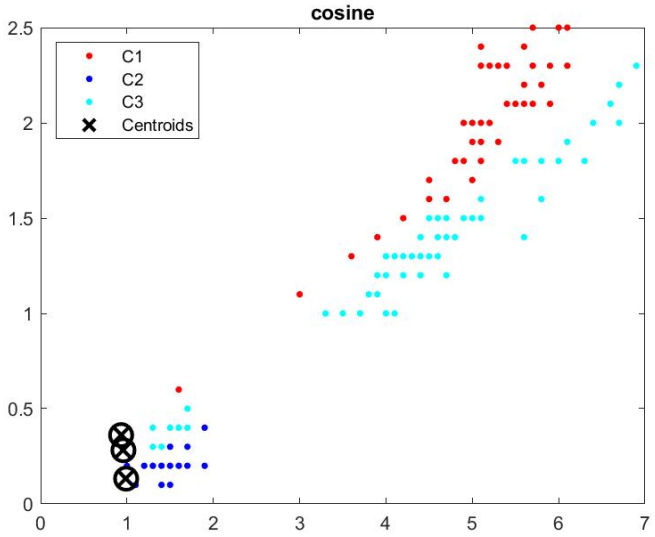
Από όλες τις διαφορετικές συσταδοποιήσεις, αυτή που φαίνεται να είναι η ποίο αποτελεσματική είναι η Euclidean και η CityBlock. Αυτό μπορούμε να το διακρίνουμε καλύτερα κάνοντας plot το IDX που είναι η πρώτη τιμή από τις δύο όπου επιστρέφει η kmeans. Ποίο συγκεκριμένα, εφόσον ξέρουμε ότι είναι χωρισμένα ανά 50, είναι φυσιολογικό να έχει γίνει η συσταδοποίηση αντίστοιχα στα πρώτα 50, στα επόμενα 50 και στα τελευταία 50 στοιχεία των δεδομένων.



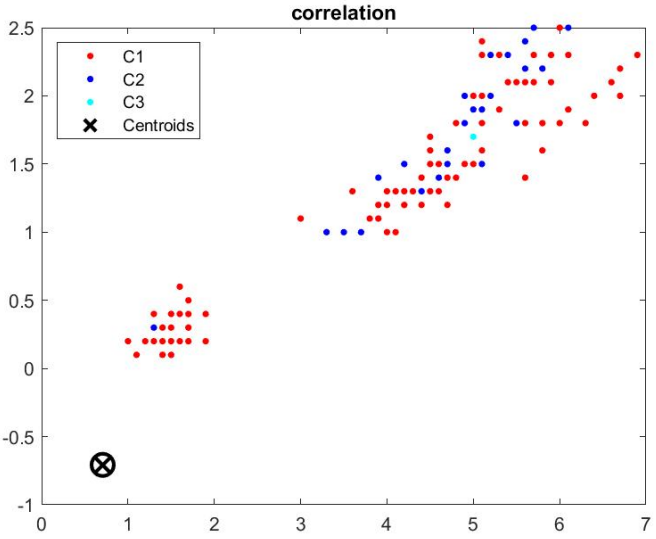
Σχήμα 2.1: Διαχωρισμός συστάδων με euclidean για την τρίτη και τέταρτη στήλη



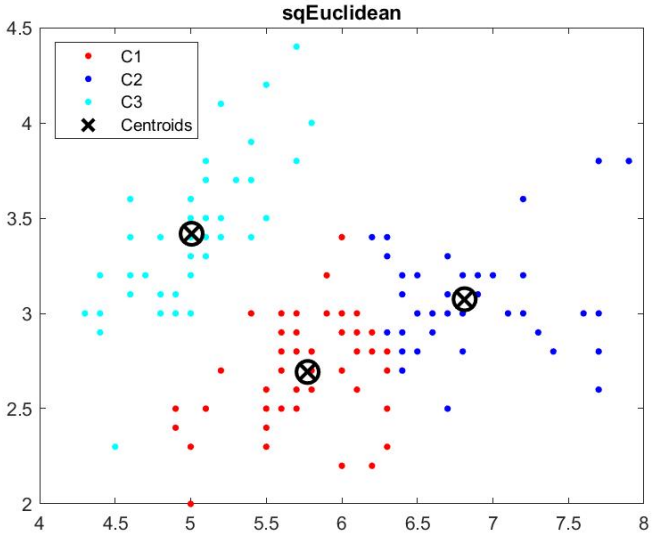
Σχήμα 2.2: Διαχωρισμός συστάδων με cityblock για την τρίτη και τέταρτη στήλη



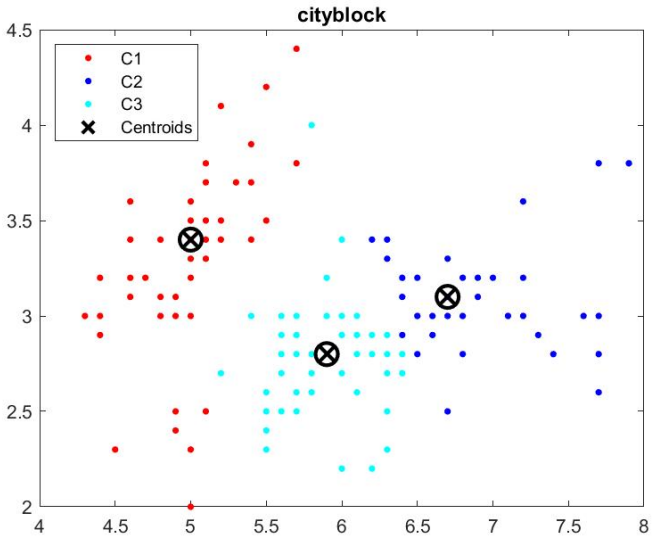
Σχήμα 2.3: Διαχωρισμός συστάδων με cosine για την τρίτη και τέταρτη στήλη



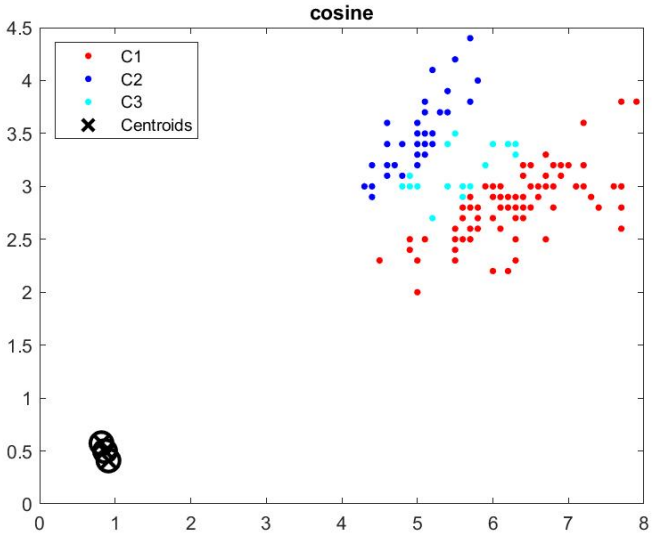
Σχήμα 2.4: Διαχωρισμός συστάδων με correlation για την τρίτη και τέταρτη στήλη



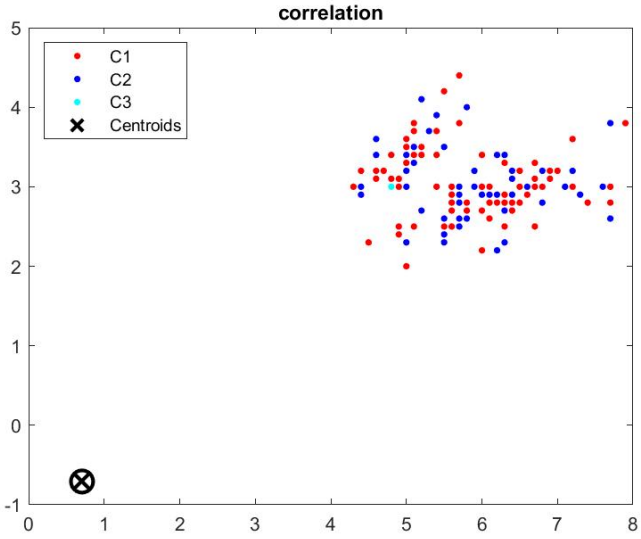
Σχήμα 2.5: Διαχωρισμός συστάδων με euclidean για την πρώτη και δεύτερη στήλη



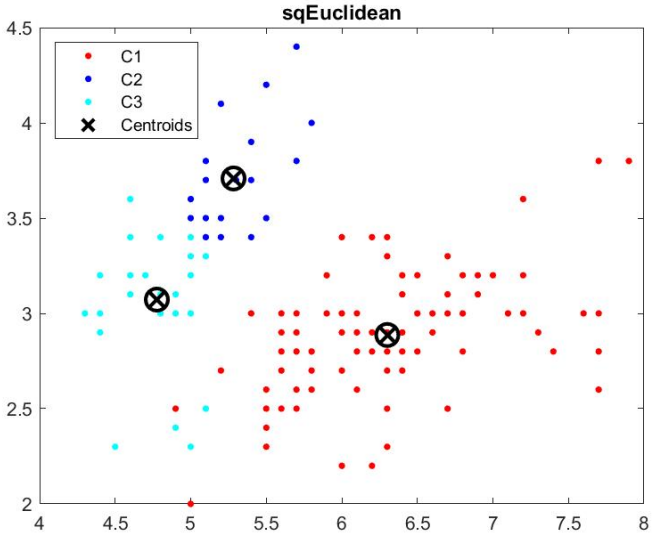
Σχήμα 2.6: Διαχωρισμός συστάδων με cityblock για την πρώτη και δεύτερη στήλη



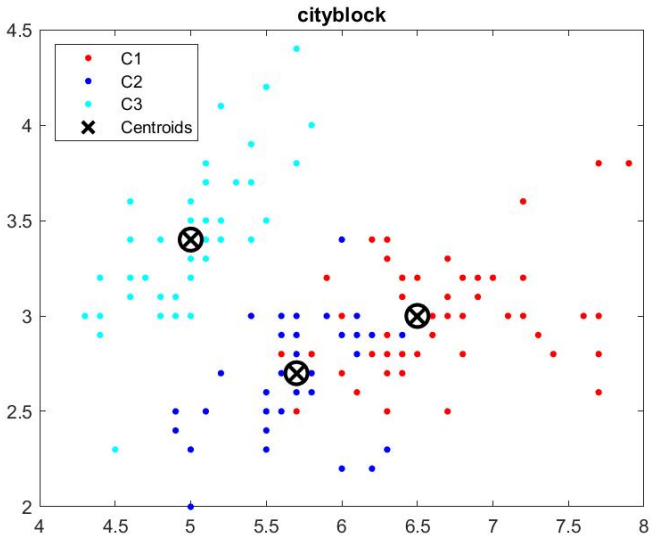
Σχήμα 2.7: Διαχωρισμός συστάδων με cosine για την πρώτη και δεύτερη στήλη



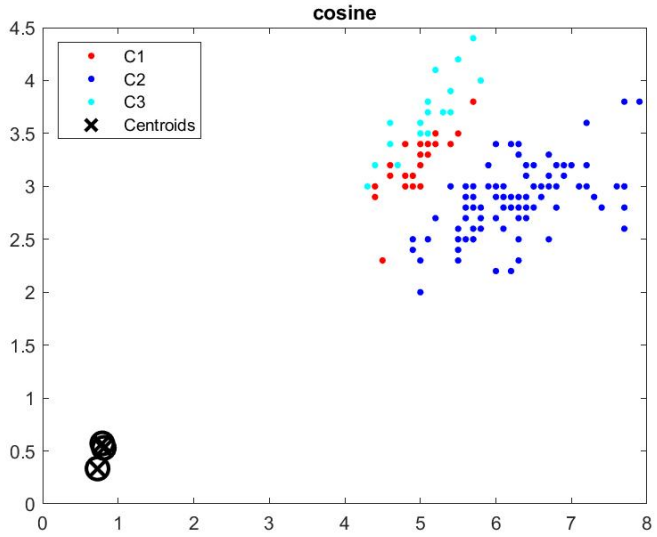
Σχήμα 2.8: Διαχωρισμός συστάδων με correlation για την πρώτη και δεύτερη στήλη



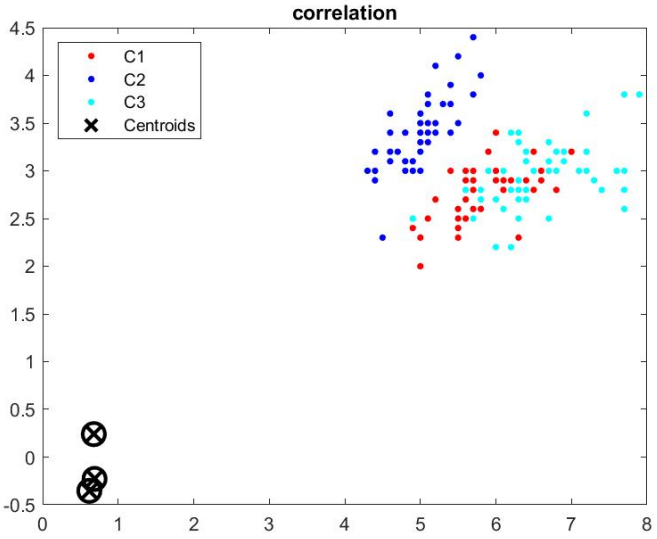
Σχήμα 2.9: Διαχωρισμός συστάδων με euclidean για την πρώτη έως και την τέταρτη στήλη



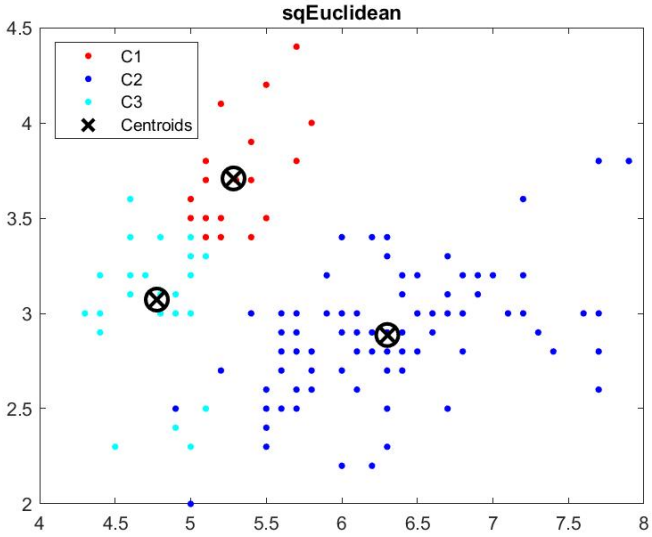
Σχήμα 2.10: Διαχωρισμός συστάδων με cityblock για την πρώτη έως και την τέταρτη στήλη



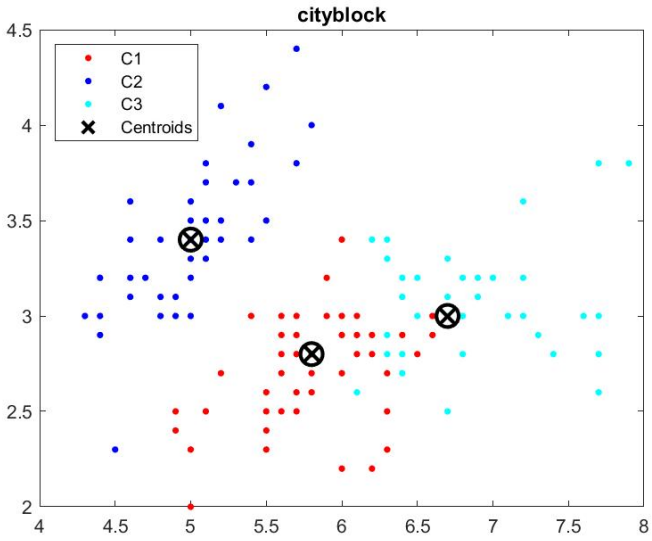
Σχήμα 2.11: Διαχωρισμός συστάδων με cosine για την πρώτη έως και την τέταρτη στήλη



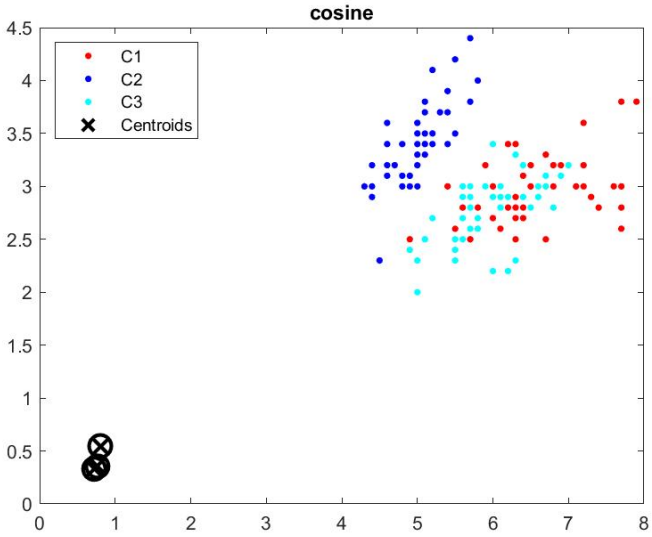
Σχήμα 2.12: Διαχωρισμός συστάδων με correlation για την πρώτη έως και την τέταρτη στήλη



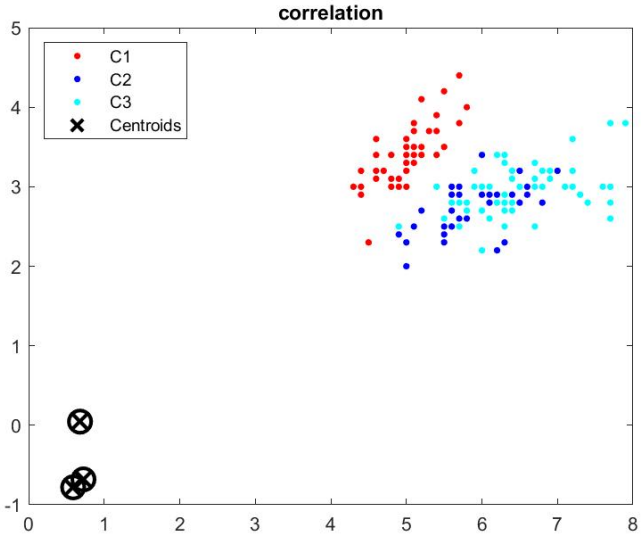
Σχήμα 2.13: Διαχωρισμός συστάδων με euclidean για την πρώτη έως και την τρίτη στήλη



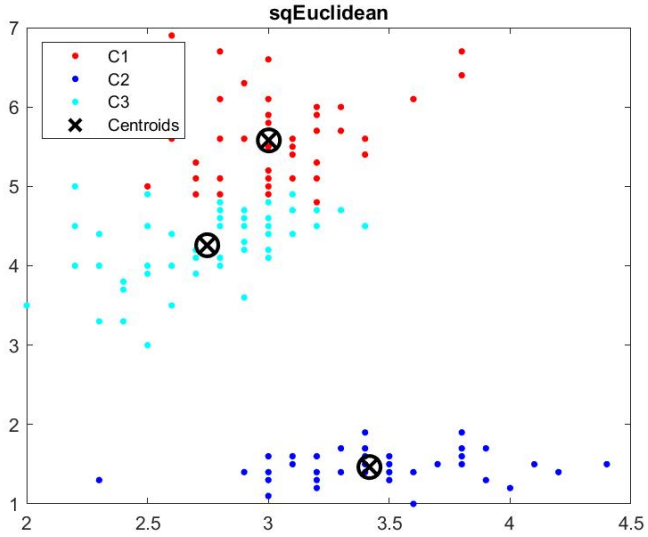
Σχήμα 2.14: Διαχωρισμός συστάδων με cityblock για την πρώτη έως και την τρίτη στήλη



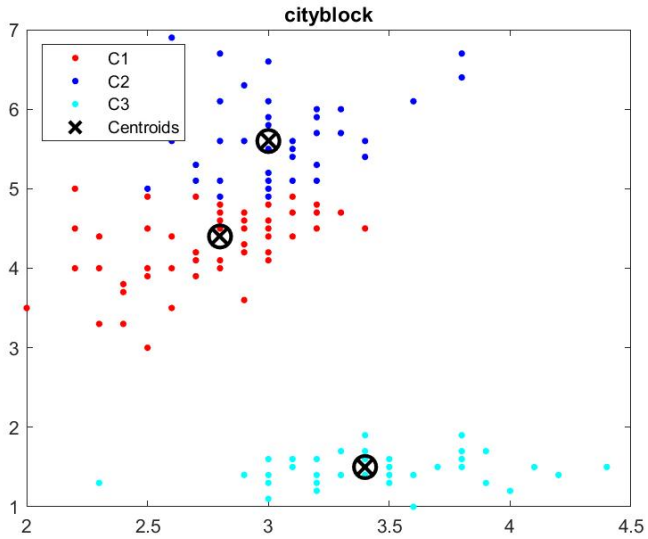
Σχήμα 2.15: Διαχωρισμός συστάδων με cosine για την πρώτη έως και την τρίτη στήλη



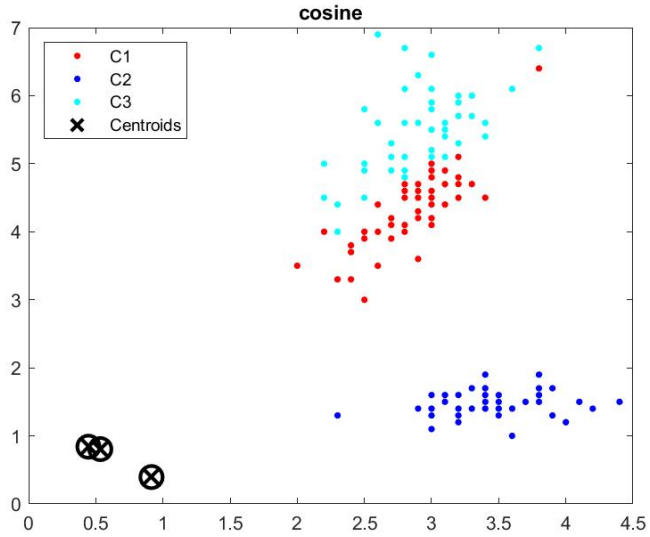
Σχήμα 2.16: Διαχωρισμός συστάδων με correlation για την πρώτη έως και την τρίτη στήλη



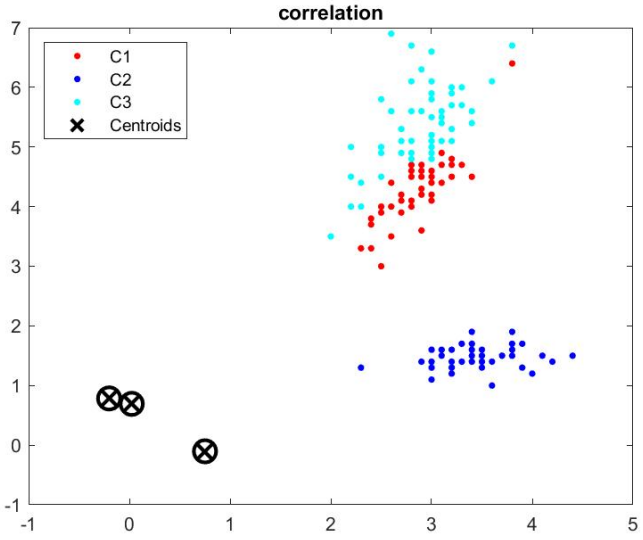
Σχήμα 2.17: Διαχωρισμός συστάδων με euclidean για την δεύτερη έως και την τέταρτη στήλη



Σχήμα 2.18: Διαχωρισμός συστάδων με cityblock για την δεύτερη έως και την τέταρτη στήλη



Σχήμα 2.19: Διαχωρισμός συστάδων με cosine για την δεύτερη έως και την τέταρτη στήλη



Σχήμα 2.20: Διαχωρισμός συστάδων με correlation για την δεύτερη έως και την τέταρτη στήλη

```
load('iris.dat')

method = {'sqEuclidean', 'cityblock',...
    'cosine', 'correlation'};
plotColors = {'r', 'b', 'c'};
combinations = [3,4; 1,2; 1,4; 1,3; 2,4];
```

```
k = 3;

for j =1:length(combinations)
   X = iris(:,[combinations(j,1):combinations(j,2)]);

   for i = 1:length(method)
      [IDX, C] = kmeans(X,k, 'distance', method(i));

      figure
      plot(IDX,'o')

      figure
      for z=1:k
         plot( X(IDX==z,1), X(IDX==z,2),...
            'LineStyle','none',...
            'Marker','.',...
            'color',plotColors{z},...
            'MarkerSize',12)
         hold on
      end

      title(method(i));

      plot(C(:,1),C(:,2),'kx', 'MarkerSize', 12, 'LineWidth', 2)
      plot(C(:,1),C(:,2),'ko', 'MarkerSize', 12, 'LineWidth', 2)
      legend('C1','C2','C3','Centroids','Location','NW')

%       filename = strcat('images/iris/comb(',...
%          num2str(combinations(j,1),'%d'), ',',...
%          num2str(combinations(j,2),'%d'), ')_',...
%          method{i});
%       print(filename,'-djpeg')
   end
end
```

Κώδικας 2.1: Script συσταδοποίησης με k-means για διάφορους μεθόδους στα δεδομένα Iris

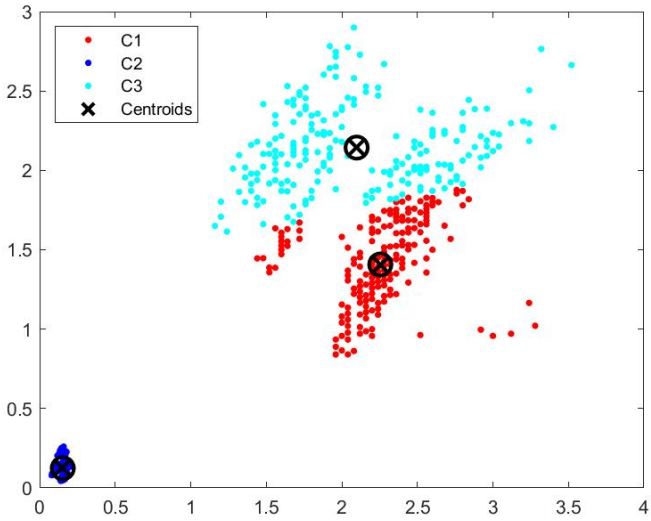
2.2 Διαχωριστική συσταδοποίηση για το σύνολο δεδομένων xV

Στην εικόνα 2.21, η συσταδοποίηση που έχει γίνει είναι αποδεκτή. Ωστόσο, οπτικά φαίνεται πως υπάρχει καλύτερος διαχωρισμός συσταδοποίησης για τα C1 και C3.

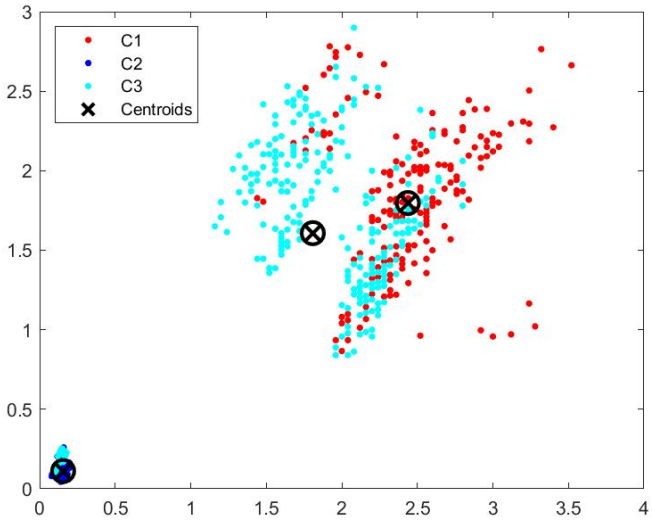
Ο διαχωρισμός με βάση όλες τις στήλες (εικ. 2.22) δεν είχε θεμιτό αποτέλεσμα και η συσταδοποίηση που υπάρχει δεν είναι ικανοποιητική.

Τέλος, η συσταδοποίηση με βάση τις στήλες 296 και 305 (εικ. 2.23), δεν είναι καθόλου εμφανές ο διαχωρισμός. Ενδεχομένως, να ήταν καλύτερος ο διαχωρισμός αν ο κεντροειδής του C3 ήταν δεξιότερα και του C1 ποίο ψηλά και δεξιά, δηλαδή κοντά στο σημείο (0.2, 0.2).

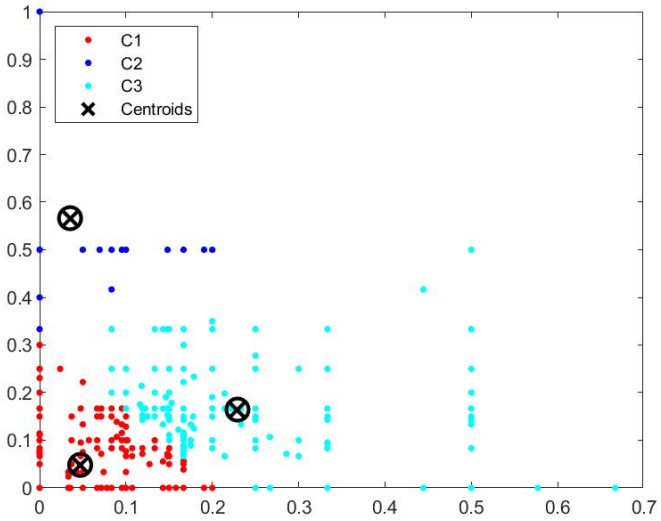
Είναι εμφανές, πως η καλύτερη συσταδοποίηση έγινε με τα δύο πρώτα χαρακτηριστικά (εικ. 2.21) όπου επέστρεψε οπτικά των καλύτερο διαχωρισμό.



Σχήμα 2.21: Διαχωρισμός συστάδων με βάση τις πρώτες δύο στήλες



Σχήμα 2.22: Διαχωρισμός συστάδων με βάση όλες τις στήλες



Σχήμα 2.23: Διαχωρισμός συστάδων με βάση τις στήλες 296 και 305

```
load('xV.mat')

notNaN = ~isnan(xV);
xV(~notNaN) = 0;

k = 3;
plotArgs = {'r', 'b', 'c'};

X = xV(:,1:2);
[IDX, C] = kmeans(X,k);


figure
for j=1:k
   plot( X(IDX==j,1), X(IDX==j,2),...
      'LineStyle','none',...
      'Marker','.',...
      'color',plotArgs{j},...
      'MarkerSize',12)
   hold on
end

plot(C(:,1),C(:,2),'kx', 'MarkerSize', 12, 'LineWidth', 2)
plot(C(:,1),C(:,2),'ko', 'MarkerSize', 12, 'LineWidth', 2)
legend('C1','C2','C3','Centroids','Location','NW')

% print('images/xV/comb(1,2)_sqEuclidean','-djpeg')

```

```
X = xV;
[IDX, C] = kmeans(X,k);


figure
for j=1:k
   plot( X(IDX==j,1), X(IDX==j,2),...
      'LineStyle','none',...
      'Marker','.',...
      'color',plotArgs{j},...
      'MarkerSize',12)
   hold on
end

plot(C(:,1),C(:,2),'kx', 'MarkerSize', 12, 'LineWidth', 2)
plot(C(:,1),C(:,2),'ko', 'MarkerSize', 12, 'LineWidth', 2)
legend('C1','C2','C3','Centroids','Location','NW')

% print('images/xV/comb(all)_sqEuclidean','-djpeg')


X = [xV(:,296),xV(:,305)];
[IDX, C] = kmeans(X,k);


figure
for j=1:k
   plot( X(IDX==j,1), X(IDX==j,2),...
      'LineStyle','none',...
      'Marker','.',...
      'color',plotArgs{j},...
      'MarkerSize',12)
   hold on
end

plot(C(:,1),C(:,2),'kx', 'MarkerSize', 12, 'LineWidth', 2)
plot(C(:,1),C(:,2),'ko', 'MarkerSize', 12, 'LineWidth', 2)
legend('C1','C2','C3','Centroids','Location','NW')

% print('images/xV/comb(296,305)_sqEuclidean','-djpeg')
```

Κώδικας 2.2: Script συσταδοποίησης με k-means για διάφορους μεθόδους στα δεδομένα xV