CLUSTERING MENGGUNAKAN K-MEANS ALGORITHM
(K-MEANS ALGORITHM CLUSTERING)

Nur Wakhidah
Fakultas Teknologi Informasi dan Komunikasi Universitas Semarang

Abstract

Classification is the process of organizing object into groups whose members are similar in same way and a part of pattern recognition. Two kind of classification is supervised classification and unsupervised classification. K-Means is a type of unsupervised classification method which partitions data items into one or more clusters. K-Means tries to model a dataset into clusters so that data items in a cluster have similar characteristic and have different characteristics from the other clusters.

Keyword : pattern recognition, clustering, k-means

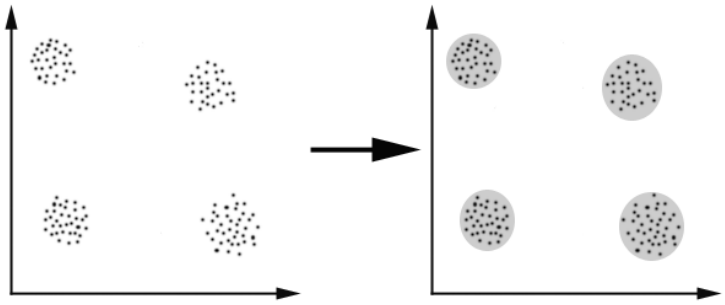
I. PENDAHULUAN

Dalam system klasifikasi terdapat 2 jenis yaitu *supervised classification* dan *unsupervised classification*. Pada *unsupervised classification*, dimana pembelajaran pola tentang pembagian class tidak diberikan, sehingga lebih banyak focus untuk memahami pola dalam cluster yang dapat dimengerti untuk menemukan persamaan dan perbedaan antar pola dan untuk memperoleh kesimpulan bermanfaat. Ide tersebut dapat dijumpai pada banyak bidang, seperti ilmu yang mempelajari hidup (biologi, zoology), ilmu pengetahuan medis (psychiatry, pathology), ilmu-ilmu sosial (sociology, archeology), ilmu pengetahuan bumi (geography, geology), dan rancang-bangun.

Misal binatang seperti domba, anjing, kucing termasuk kluster binatang menyusui; burung pipit, burung camar termasuk kluster burung; ular, kadal termasuk kluster binatang melata; ikan mas, ikan mullet merah, ikan hiu biru termasuk kluster ikan; dan kodok termasuk kluster binatang ampibi. Dalam mengelompokkan binatang-binatang tersebut ke dalam suatu cluster dibutuhkan penggambaran *clustering criterion*, hal ini sama halnya jika kita akan mengelompokkan cara binatang membawa keturunan mereka ke dalam sebuah cluster. Sebagai contoh domba, anjing, kucing, dan ikan hiu biru dapat dikelompokkan dalam satu cluster sedangkan binatang yang lain dapat dibentuk ke dalam cluster yang lain. Untuk jelasnya dapat dilihat pada gambar berikut.

Gambar 1. Beberapa cluster dari *clustering criterion*. Gambar (a) kelompok cara binatang membawa keturunannya, Gambar (b) kelompok paru-paru binatang

Clustering dapat dianggap yang paling penting dalam masalah *unsupervised learning,* karena setiap masalah semacam ini, ia berurusan dengan mencari *struktur* dalam kumpulan yang tidak diketahui datanya. Sehingga dapat didefinisikan bahwa clustering merupakan "proses mengatur objek menjadi anggota kelompok yang hampir sama dalam beberapa cara". Sebuah *cluster* merupakan kumpulan objek-objek yang "sama" di antara mereka dan "berbeda" pada objek dari cluster lainnya.



Gambar 2. Identifikasi Kelompok

Dengan memperhatikan gambar di atas, kita dengan mudah mengidentifikasikan 4 kelompok menjadi data yang dapat dibagi yaitu kesamaan dengan kriteria jarak antara dua atau lebih benda dalam klaster yang sama jika mereka dekat dan sesuai dengan jarak yang diberikan. Hal ini disebut *distance-based clustering.* Lain halnya untuk jenis pengelompokan *konseptual clustering,* dimana dua atau lebih benda dalam klaster yang sama dengan mendefinisikan konsep secara umum untuk semua benda, dengan kata lain objek dikelompokkan menurut konsep deskriptif. Tujuan dari clustering adalah untuk mengklasifikasikan data, dengan cara menentukan pengelompokan dalam satu set data yang tidak diketahui. Tetapi bagaimana untuk menentukan clustering yang baik? Dapat menunjukkan tidak ada criteria absolut "terbaik" yang akan bergantung pada tujuan akhir dari clustering. Akibatnya, pengguna yang harus menyertakan kriteria ini, sehingga hasil clustering akan memenuhi kebutuhan mereka. Syarat yang harus dipenuhi dalam clustering algoritma adalah skalabilitas; berhadapan dengan berbagai jenis atribut; menemukan bentuk kelompok persyaratan minimal adalah domain pengetahuan untuk menentukan parameter masukan; kemampuan untuk menangani gangguan; dimensi tinggi; serta interpretability dan usability.

II. APLIKASI

Clustering algoritma dapat diterapkan dalam berbagai bidang, misalnya:

- *Pemasaran:* mencari kelompok pelanggan yang mirip dengan perilaku, diberikan database yang besar berisi data pelanggan mereka memperoleh properti dan catatan masa lalu;
- *Biologi:* klasifikasi tanaman dan binatang;
- *Perpustakaan:* katalog buku;
- *Asuransi:* mengidentifikasi kelompok pemegang polis asuransi motor dengan rata-rata klaim biaya tinggi;
- *Perencanaan kota:* mengidentifikasi kelompok rumah sesuai dengan tipe rumah, nilai dan lokasi geografis;

III. KLASIFIKASI

Clustering algoritma dapat diklasifikasikan sebagai berikut:

1. Exclusive Clustering
   - Data dikelompokkan ke dalam suatu cara yang eksklusif, sehingga jika suatu fakta milik suatu cluster maka tidak dapat dipakai (menjadi anggota) di cluster lain
2. Overlapping Clustering
   - Menggunakan fuzzy set untuk cluster data sehingga titik kemungkinan

memiliki dua atau lebih kelompok yang berbeda sesuai derajat keanggotaannya. Dalam hal ini data akan dihubungkan dengan nilai keanggotaannya.

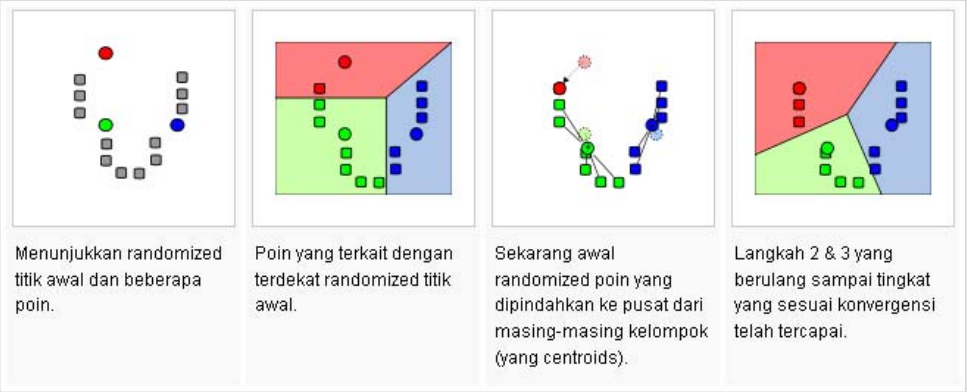
3. Hierarchical Clustering
   - Didasarkan pada kesatuan antara dua kelompok terdekat. Permulaan kondisi diwujudkan dengan menetapkan setiap datum sebagai cluster. Setelah beberapa iterasi mencapai final kelompok yang diinginkan.
4. Probabilistic Clustering
   - Sepenuhnya menggunakan pendekatan probabilistic

Terdapat empat algoritma yang paling sering digunakan dalam clustering, yaitu:

- K-means (exclusive clustering)
- Fuzzy C-means (overlapping clustering)
- Hierarchical clustering
- Mixture of Gaussians (probabilistic clustering)

IV. K-MEANS

K-Means merupakan algoritma untuk cluster n objek berdasarkan atribut menjadi k partisi, dimana k < n. Gambar berikut ini menunjukkan k-means clustering algoritma dalam tindakan, untuk kasus dua dimensi. Pusat awal yang dihasilkan secara acak untuk menunjukkan tahapan lebih rinci. Background ruang partisi hanya untuk ilustrasi dan tidak dihasilkan oleh algoritma k-means.



Gambar 3. K-means clustering dalam tindakan (2 dimensi)

Kelemahan dari K-Means

K-means memiliki banyak kelemahan, antara lain:

- Bila jumlah data tidak terlalu banyak, mudah untuk menentukan cluster awal.
- Jumlah cluster, sebanyak K, harus ditentukan sebelum dilakukan perhitungan.
- tidak pernah mengetahui real cluster dengan menggunakan data yang sama, namun jika dimasukkan dengan cara yang berbeda mungkin dapat memproduksi cluster yang berbeda jika jumlah datanya sedikit.
- tidak tahu kontribusi dari atribut dalam proses pengelompokan karena dianggap bahwa setiap atribut memiliki bobot yang sama.

Salah satu cara untuk mengatasi kelemahan itu adalah dengan menggunakan K-means clustering namun hanya jika tersedia banyak data.

Algoritma K-Means

Langkah-langkah dalam algoritma K-means clustering adalah :

1. Menentukan jumlah cluster
2. Menentukan nilai centroid
   Dalam menentukan nilai centroid untuk awal iterasi, nilai awal centroid dilakukan secara acak. Sedangkan jika menentukan nilai centroid yang merupakan tahap dari

iterasi, maka digunakan rumus sebagai berikut

$$\overline{v}_{ij} = \frac{1}{N_i} \sum_{k=0}^{N_i} x_{kj} \ ,$$

dimana :

v_{ij} adalah centroid/ rata-rata cluster ke-I untuk variable ke-j

N_i adalah jumlah data yang menjadi anggota cluster ke-i

i,k adalah indeks dari cluster

j adalah indeks dari variabel

x_{kj} adalah nilai data ke-k yang ada di dalam cluster tersebut untuk variable ke-j

3. Menghitung jarak antara titik centroid dengan titik tiap objek

   Untuk menghitung jarak tersebut dapat menggunakan Euclidean Distance, yaitu

$$D_e = \sqrt{(x_i - s_i)^2 + (y_i - t_i)^2} \ ,$$

   dimana :

   D_e adalah Euclidean Distance

   i adalah banyaknya objek,

   (x,y) merupakan koordinat object dan

   (s,t) merupakan koordinat centroid.

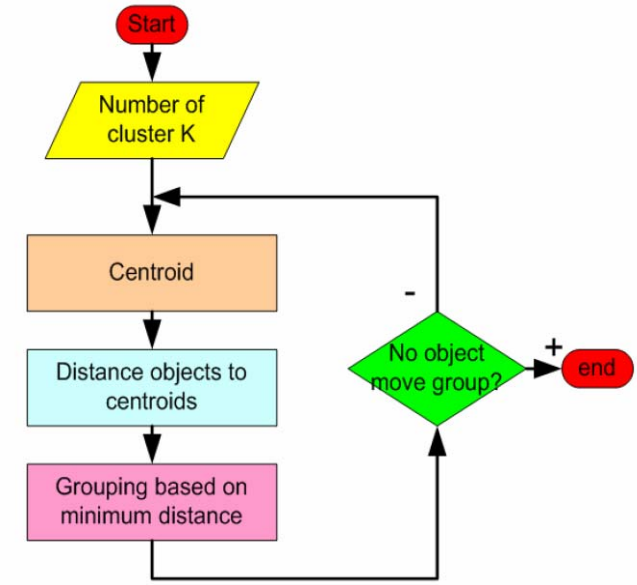
4. Pengelompokan object

   Untuk menentukan anggota cluster adalah dengan memperhitungkan jarak minimum objek. Nilai yang diperoleh dalam keanggotaan data pada distance matriks adalah 0 atau 1, dimana nilai 1 untuk data yang dialokasikan ke cluster dan nilai 0 untuk data yang dialokasikan ke cluster yang lain.

5. Kembali ke tahap 2, lakukan perulangan hingga nilai centroid yang dihasilkan tetap dan anggota cluster tidak berpindah ke cluster lain.

Flowchart K-Means Clustering

Berikut penggambaran algoritma k-means clustering menggunakan flowchart :



Gambar 4. Flowchart K-means Clustering

V. Kasus :

Misalnya kita memiliki 4 objek sebagai titik data pelatihan dan setiap obyek memiliki 2 atribut . Tiap atribut mewakili koordinat dari objek, yaitu

Objek Atribut 1 (X): bobot indeks

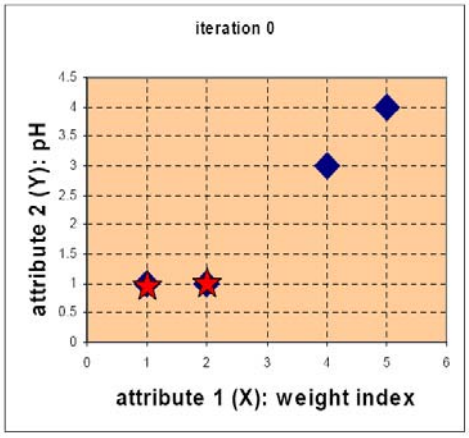
Objek Atribut 2 (Y): pH

Tabel 1. Data Kasus

Object	Atribut 1 (X) : bobot index	Atribut 2 (Y) : pH
Medicine A	1	1
Medicine B	2	1
Medicine C	4	3
Medicine D	5	4

Untuk menyelesaikan permasalahan tersebut, kita dapat melakukan beberapa tahap, yaitu :

1. Menentukan Jumlah Cluster
   Dengan memperhatikan data tersebut, kita dapat mengelompokkan object tersebut ke dalam dua cluster sesuai dengan atributnya (yaitu cluster 1 dan cluster 2). Masalahnya adalah bagaimana menentukan medicine tersebut merupakan anggota dalam cluster 1 atau cluster 2.
   Dari data yang diperoleh, dapat ditentukan bahwa 4 object tersebut memiliki 2 atribut (bobot index dan pH), dimana tiap-tiap medicine mewakili satu titik dengan 2 atribut (X,Y). Untuk lebih jelasnya dapat dilihat pada gambar berikut.



Gambar 5. Iteration 0

2. Menentukan nilai centroid
   Untuk menentukan nilai awal centroid dilakukan secara acak. Disini, dimisalkan titik koordinat medicine A adalah cluster 1 (C1) dan medicine B adalah cluster 2 (C2) sebagai nilai centroid awal.
   - C1 = (1,1)
   - C2 = (2,1)
3. Menghitung jarak antara titik centroid dengan tiap titik object.
   Untuk menghitung jarak antara titik centroid dengan tiap titik object, kita dapat menggunakan rumus Euclidean Distance yaitu

$$D_e = \sqrt{(x_i - s_i)^2 + (y_i - t_i)^2}$$

dimana :
De adalah Euclidean Distance
i adalah banyaknya objek,
(x,y) merupakan koordinat object
(s,t) merupakan koordinat centroid
Sehingga pada iterasi 0, dengan titik centroid C1 = (1,1) dan C2 = (2,1)

	Medicine A	Medicine B	Medicine C	Medicine D
X	1	2	4	5
Y	1	1	3	4

Berikut adalah cara untuk menghitung distance dari tiap object :

- Medicine A = (1,1) dengan C1=(1, 1)

  → $= \sqrt{(1-1)^2 + (1-1)^2} = 0$

  dengan C2=(2,1)

  → $= \sqrt{(1-2)^2 + (1-1)^2} = 1$

- Medicine B = (2,1) dengan C1=(1, 1)

  → $= \sqrt{(2-1)^2 + (1-1)^2} = 1$

  dengan C2=(2,1)

  → $= \sqrt{(2-2)^2 + (1-1)^2} = 0$

- Medicine C = (4,3) dengan C1=(1, 1)

  → $= \sqrt{(4-1)^2 + (3-1)^2} = 3.61$

  dengan C2=(2,1)

  → $= \sqrt{(4-2)^2 + (3-1)^2} = 2.83$

- Medicine D = (5,4) dengan C1=(1, 1)

  → $= \sqrt{(5-1)^2 + (4-1)^2} = 5$

  dengan C2=(2,1)

  → $= \sqrt{(5-2)^2 + (4-1)^2} = 4.24$

dari perhitungan diatas, diperoleh distance matriksnya, yaitu

$$D^0 = \begin{matrix} A & B & C & D \\ \end{matrix} \begin{pmatrix} 0 & 1 & 3.61 & 5 \\ 1 & 0 & 2.83 & 4.24 \end{pmatrix} \begin{matrix} \rightarrow \text{C1=(1,1)} \\ \rightarrow \text{C2=(2,1)} \end{matrix}$$

4. Pengelompokan Object.

Setelah menghitung distance matriks, kita menentukan anggota cluster menurut jarak minimum dari centroid. Dengan merujuk pada distance matriks, medicine A termasuk cluster 1, sedangkan medicine B, C dan D termasuk cluster 2. Hal ini dapat dilihat pada perolehan nilai sebagai berikut :

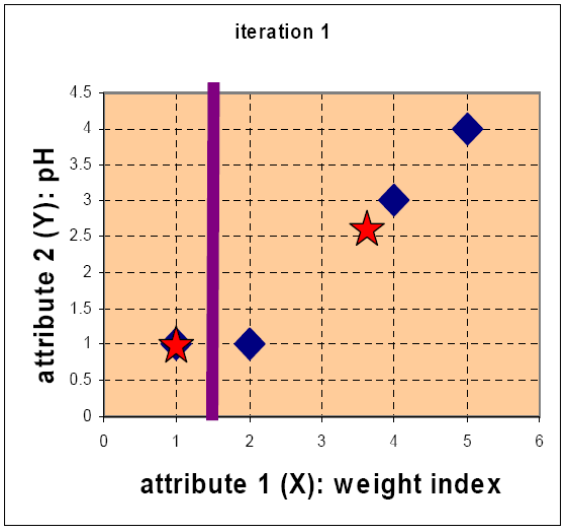
$$G^0 = \begin{matrix} A & B & C & D \\ \end{matrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{pmatrix} \begin{matrix} \rightarrow \text{Cluster 1} \\ \rightarrow \text{Cluster 2} \end{matrix}$$

5. Iterasi 1, menentukan centroid baru.

Himpunan yang terbentuk pada tahap sebelumnya, telah diketahui anggota tiap cluster. Untuk cluster 1 mempunyai anggota medicine A saja, sedangkan cluster 2 mempunyai anggota medicine B, C dan D. Dari data tersebut, hitung kembali centroid untuk menentukan centroid baru. Karena pada cluster 1 hanya mempunyai 1 anggota, maka untuk centroid baru masih berada di C1 = (1,1). Sedangkan pada C2 dengan menghitung nilai rata-ratanya dapat diperoleh nilai centroid barunya, yaitu :

$$C2 = \left(\frac{2+4+5}{3}, \frac{1+3+4}{3} \right)$$

$$C2 = \left(\frac{11}{3}, \frac{8}{3} \right)$$



Gambar 6. Iteration 1

6. Iterasi 1, menghitung jarak antara titik centroid baru dengan tiap titik object.
   Pada tahap menghitung jarak antara object dengan centroid baru. Hal ini hampir sama dengan tahap 3, yaitu menghitung jarak dengan C2

$$C2 = \left(\frac{11}{3}, \frac{8}{3}\right)$$

   Dengan cara perhitungan yang sama pada tahap 3, maka diperoleh distance matriksnya, yaitu

$$D^1 = \begin{matrix} A & B & C & D \\ \end{matrix}$$

$$\begin{pmatrix} 0 & 1 & 3.61 & 5 \\ 3.14 & 2.36 & 0.47 & 1.89 \end{pmatrix} \rightarrow \textbf{C1 = (1,1)} \quad \rightarrow C2 = \left(\frac{11}{3}, \frac{8}{3}\right)$$

7. Iterasi 1, melakukan pengelompokan object
   Hampir sama dengan tahap 4, yaitu menentukan anggota cluster dengan menghitung jarak minimum tiap object dengan centroid baru. Hasil yang diperoleh :

$$G^1 = \begin{matrix} A & B & C & D \\ \end{matrix}$$

$$\begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix} \begin{matrix} \rightarrow \textbf{Cluster 1} \\ \rightarrow \textbf{Cluster 2} \end{matrix}$$

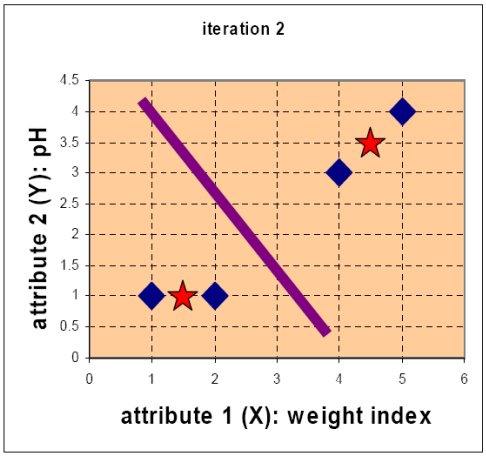
8. Iterasi 2, menentukan centroid baru.
   Tahap ini mengulang kembali tahap 5, yaitu menghitung centroid baru. Dari cluster 1 yang mempunyai 2 anggota yaitu medicine A dan B, dan cluster 2 yang mempunyai 2 anggota yaitu medicine C dan D, maka hasil centroid baru yang diperoleh adalah :

$$C1 = \left(\frac{1+2}{2}, \frac{1+1}{2}\right)$$

$$C1 = \left(\frac{3}{2}, 1\right)$$

$$C2 = \left(\frac{4+5}{2}, \frac{3+4}{2}\right)$$

$$C2 = \left(\frac{9}{2}, \frac{7}{2}\right)$$



Gambar 7. Iteration 2

9. Iterasi 2, menghitung jarak antara titik centroid baru dengan tiap titik object.
   Tahap ini juga hampir sama dengan tahap 3, yaitu menghitung jarak dengan Centroid baru

$$C1 = \left(\frac{3}{2}, 1\right)$$

$$C2 = \left(\frac{9}{2}, \frac{7}{2}\right)$$

Dengan cara perhitungan yang sama pada tahap 3, maka diperoleh distance matriksnya, yaitu

$$D^2 = \begin{array}{c} \begin{matrix} A & B & C & D \end{matrix} \\ \begin{pmatrix} 0.5 & 0.5 & 3.20 & 4.61 \\ 4.30 & 3.54 & 0.71 & 0.71 \end{pmatrix} \end{array} \begin{array}{l} \rightarrow C1 = \left(\frac{3}{2}, 1\right) \\ \rightarrow C2 = \left(\frac{9}{2}, \frac{7}{2}\right) \end{array}$$

10. Iterasi 2, melakukan pengelompokan object

Hampir sama dengan tahap 4, yaitu menentukan anggota cluster dengan menghitung jarak minimum tiap object dengan centroid baru yang telah dihasilkan. Hasil yang diperoleh :

$$G^2 = \begin{array}{c} \begin{matrix} A & B & C & D \end{matrix} \\ \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix} \end{array} \begin{array}{l} \rightarrow \text{Cluster 1} \\ \rightarrow \text{Cluster 2} \end{array}$$

Berdasarkan hasil anggota cluster yang diperoleh tetap sama antara $G^1 = G^2$, maka iterasi dihentikan.

VI. KESIMPULAN

Dari 4 objek yang digunakan dalam kasus tersebut, dapat disimpulkan bahwa :

1. K-means Algoritma merupakan algoritma yang sederhana
2. K-means clustering mampu menyelesaikan permasalahan yang ada
3. Terdapat 2 cluster yang dihasilkan, untuk cluster 1 mempunyai anggota medicine A dan B, sedangkan cluster 2 mempunyai anggota C dan D

Untuk hasil yang diperoleh, dapat dilihat pada table berikut.

Tabel 2. Hasil Clustering

Object	Atribut 1 (X) : bobot index	Atribut 2 (Y) : pH	Cluster (result)
Medicine A	1	1	1
Medicine B	2	1	1
Medicine C	4	3	2
Medicine D	5	4	2

VII. Daftar Pustaka

E.S. Gopi, “Algorithm Collections for Digital Signal Processing Applications Using Matlab”, Spinger: National Institute of Technology, Tiruchi, India,

Chen Yu, “K-Means Clustering”, Indiana University

Sergios Theodoridis, Konstantinos Koutroumbas : “Pattern Recognition”, Elsevier Academic Press

Teknomo, Kardi. K-Means Clustering Tutorials. http://people.revoledu.com/kardi/tutorial/kMean/

http://en.wikipedia.org/wiki/k-means_algorithm

http://home.dei.polimi.it/matteucc/clustering/tutorial_ht