



MODEL DATAMINING




SISTEM PENDUKUNG KEPUTUSAN

Rahadian Kurniawan, S.Kom., M.Kom.

Sistem Pendukung Keputusan Model Datamining


Sasaran

Mahasiswa dapat memahami dan mengaplikasikan **konsep klasifikasi** dan **clustering** dalam **model datamining**.



Referensi Utama

- Turban, Efraim; Aronson, Jay, E.; Liang, Ting-Peng. 2005. *Decision Support Systems and Intelligent Systems*. International Edition, Edisi 7. New Jersey: Pearson Prentice-Hall Education International



2013 © Rahadian Kurniawan

Sistem Pendukung Keputusan Model Datamining

Clustering

- Clustering** adalah proses pengelompokan objek yang didasarkan pada kesamaan antar objek.
- Tidak seperti proses klasifikasi yang bersifat *supervised* (terawasi) *learning*, pada clustering proses pengelompokan dilakukan atas dasar *unsupervised learning*.
- Pada proses klasifikasi, akan ditentukan lokasi dari suatu kejadian pada klas tertentu dari beberapa klas yang telah teridentifikasi sebelumnya.
- Sedangkan pada proses clustering, proses pengelompokan kejadian dalam klas akan dilakukan secara alami tanpa mengidentifikasi klas-klas sebelumnya.

2013 © Rahadian Kurniawan

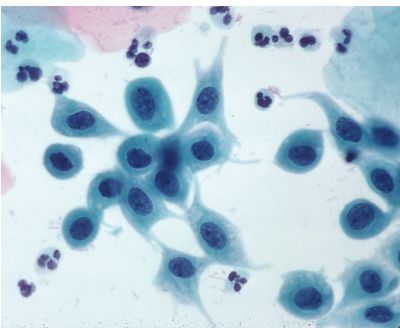
Sistem Pendukung Keputusan Model Datamining

Clustering

- Suatu metode clustering dikatakan baik apabila metode tersebut dapat menghasilkan cluster-cluster dengan kualitas yang sangat baik.
- Metode tersebut akan menghasilkan cluster-cluster dengan objek-objek yang memiliki tingkat kesamaan yang cukup tinggi dalam suatu cluster, dan memiliki tingkat ketidaksamaan yang cukup tinggi juga apabila objek-objek tersebut terletak pada cluster yang berbeda.
- Untuk mendapatkan kualitas yang baik, metode clustering sangat tergantung pada ukuran kesamaan yang akan digunakan dan kemampuannya untuk menemukan beberapa pola yang tersembunyi.

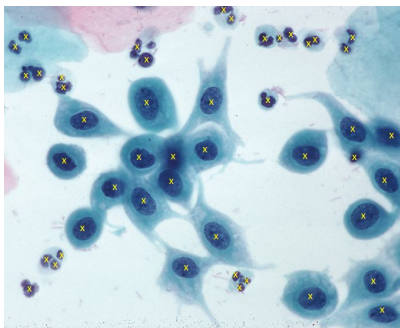
2013 © Rahadian Kurniawan

Sistem Pendukung Keputusan Model Datamining

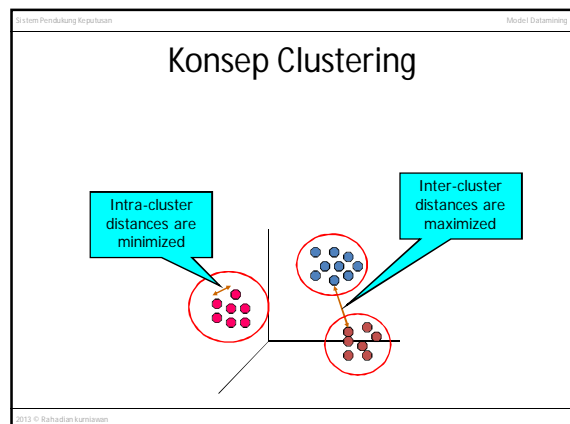
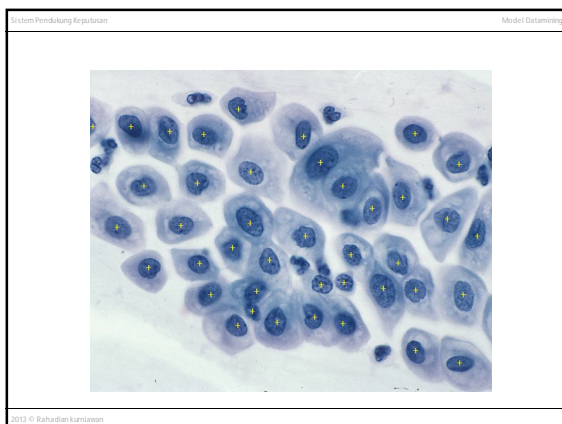
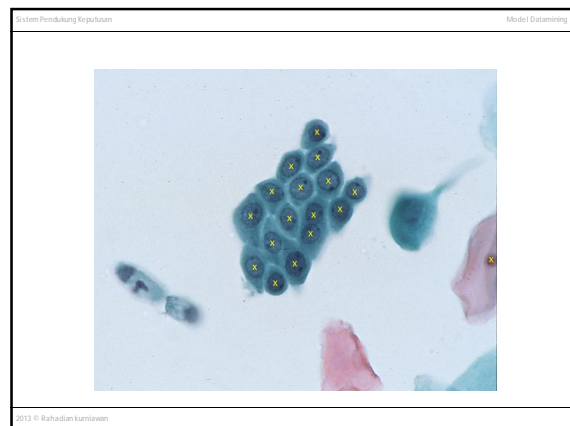
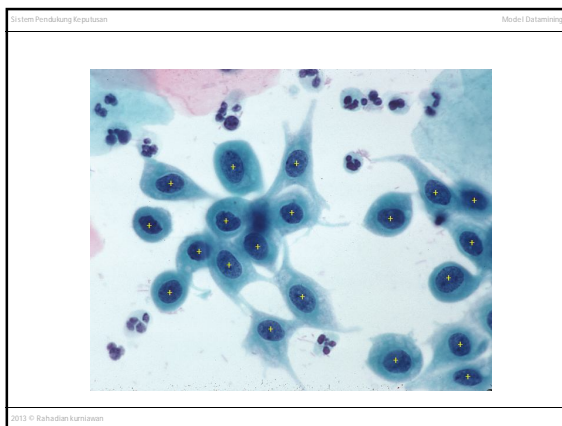
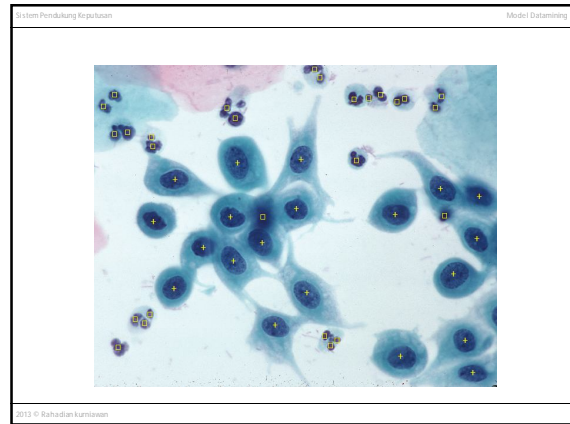
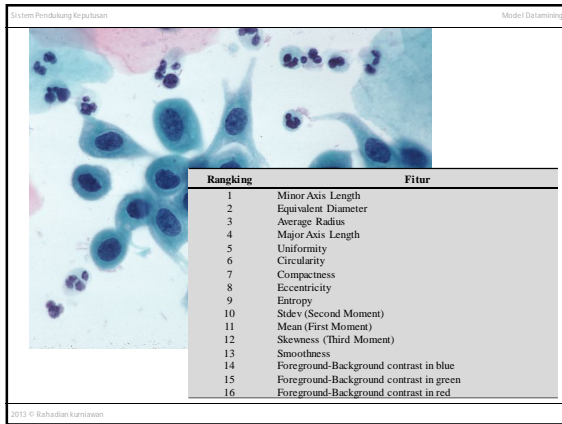


2013 © Rahadian Kurniawan

Sistem Pendukung Keputusan Model Datamining



2013 © Rahadian Kurniawan



Sistem Pendukung Keputusan Model Data Mining

K-Means

- Konsep dasar dari **K-Means** adalah pencarian pusat cluster secara iteratif.
- Pusat cluster ditetapkan berdasarkan jarak setiap data ke pusat cluster.
- Proses clustering dimulai dengan mengidentifikasi data yang akan dicluster, x_{ij} ($i=1, \dots, n$; $j=1, \dots, m$) dengan n adalah jumlah data yang akan dicluster dan m adalah jumlah variabel.

2013 © Rahadian kurniasari

Sistem Pendukung Keputusan Model Data Mining

K-Means

- Pada awal iterasi, pusat setiap cluster ditetapkan secara bebas (sembarang), c_{kj} ($k=1, \dots, K$; $j=1, \dots, m$).
- Kemudian dihitung jarak antara setiap data dengan setiap pusat cluster.
- Untuk melakukan penghitungan jarak data ke- i (x_i) pada pusat cluster ke- k (c_k), diberi nama (d_{ik}), dapat digunakan formula Euclidean, yaitu:

$$d_{ik} = \sqrt{\sum_{j=1}^m (x_{ij} - c_{kj})^2}$$

2013 © Rahadian kurniasari

Sistem Pendukung Keputusan Model Data Mining

K-Means

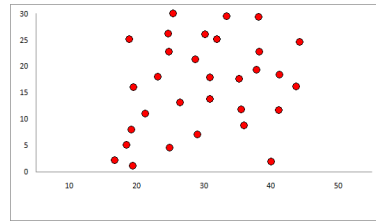
- Suatu data akan menjadi anggota dari cluster ke- J apabila jarak data tersebut ke pusat cluster ke- J bernilai paling kecil jika dibandingkan dengan jarak ke pusat cluster lainnya.
- Selanjutnya, kelompokkan data-data yang menjadi anggota pada setiap cluster.
- Nilai pusat cluster yang baru dapat dihitung dengan cara mencari nilai rata-rata dari data yang menjadi anggota pada cluster tersebut, dengan rumus:

$$c_{kj} = \frac{\sum_{i=1}^p y_{ij}}{p}; y_{ij} = x_{ij} \in \text{cluster ke-} k$$

2013 © Rahadian kurniasari

Sistem Pendukung Keputusan Model Data Mining

Konsep K-means



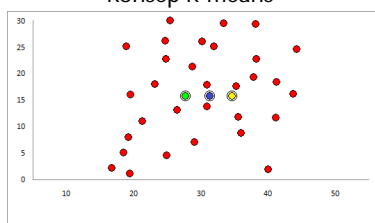
Algorithm 1 Basic K-means Algorithm.

- 1: Select K points as the initial centroids.
- 2: **repeat**
- 3: Form K clusters by assigning all points to the closest centroid.
- 4: Recompute the centroid of each cluster.
- 5: **until** The centroids don't change

2013 © Rahadian kurniasari

Sistem Pendukung Keputusan Model Data Mining

Konsep K-means



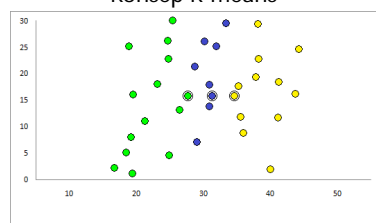
Algorithm 1 Basic K-means Algorithm.

- 1: Select K points as the initial centroids.
- 2: **repeat**
- 3: Form K clusters by assigning all points to the closest centroid.
- 4: Recompute the centroid of each cluster.
- 5: **until** The centroids don't change

2013 © Rahadian kurniasari

Sistem Pendukung Keputusan Model Data Mining

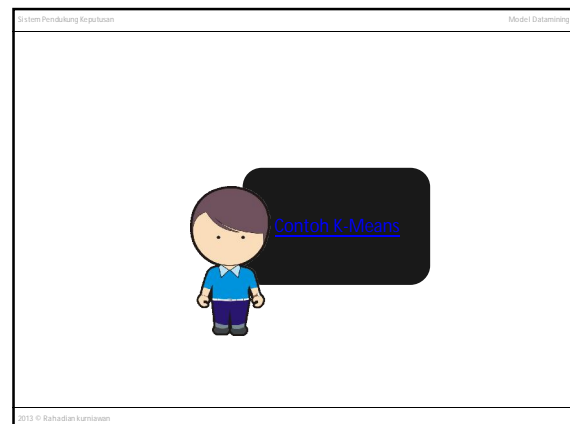
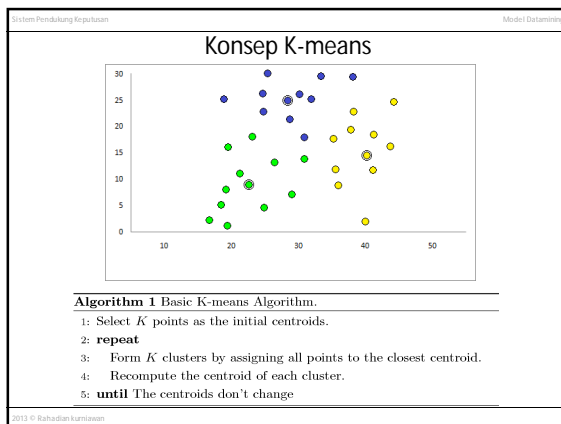
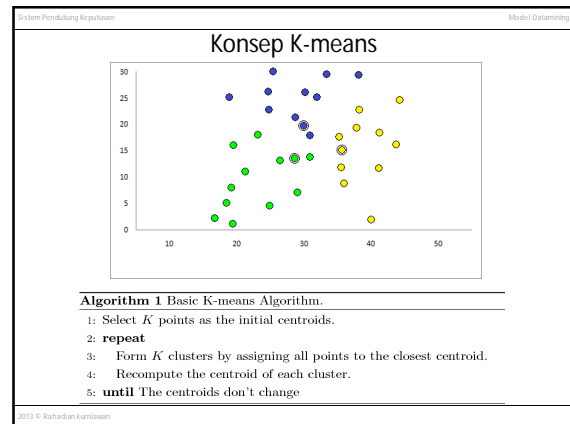
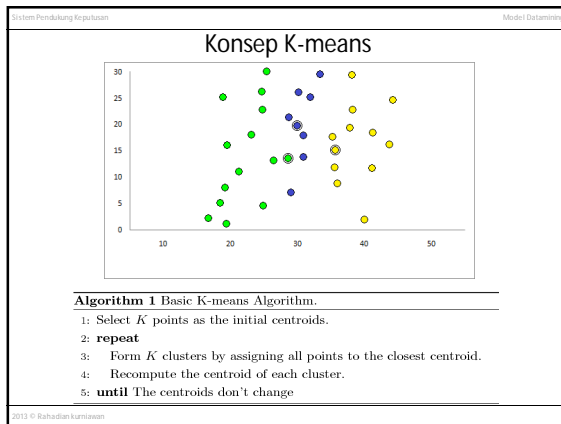
Konsep K-means



Algorithm 1 Basic K-means Algorithm.

- 1: Select K points as the initial centroids.
- 2: **repeat**
- 3: Form K clusters by assigning all points to the closest centroid.
- 4: Recompute the centroid of each cluster.
- 5: **until** The centroids don't change

2013 © Rahadian kurniasari



- Sistem Pendukung Keputusan Model Datamining
- ### Penentuan Jumlah Cluster
- Salah satu masalah yang dihadapi pada proses clustering adalah pemilihan jumlah cluster yang optimal.
 - Kauffman dan Rousseeuw (1990) memperkenalkan suatu metode untuk menentukan jumlah cluster yang optimal, metode ini disebut dengan *silhouette measure*.
 - Misalkan kita sebut A sebagai cluster dimana data X_i berada, hitung a_i sebagai rata-rata jarak X_i ke semua data yang menjadi anggota A.
 - Anggaplah bahwa C adalah sembarang cluster selain A.
- 2013 © Rahadian kurniasari

- Sistem Pendukung Keputusan Model Datamining
- ### Penentuan Jumlah Cluster
- Hitung rata-rata jarak antara X_i dengan data yang menjadi anggota dari C, sebut sebagai $d(X_i, C)$.
 - Cari rata-rata jarak terkecil dari semua cluster, sebut sebagai $b_i, b_i = \min(d(X_i, C))$ dengan $C \neq A$.
 - Silhouette dari X_i , sebut sebagai s_i dapat dipandang sebagai berikut (Chih-Ping, 2005):
- $$s_i = \begin{cases} 1 - \frac{a_i}{b_i}, & a_i < b_i \\ 0, & a_i = b_i \\ \frac{b_i}{a_i} - 1, & a_i > b_i \end{cases}$$
- 2013 © Rahadian kurniasari

Sistem Pendukung KeputusanModeli Gadamirney

Penentuan Jumlah Cluster

- Rata-rata s_j untuk semua data untuk k cluster tersebut disebut sebagai rata-rata silhouette ke- k .
- Nilai rata-rata silhouette terbesar pada jumlah cluster (katakanlah: k) menunjukkan bahwa k merupakan jumlah cluster yang optimal.

2013 © Rahadian Kusnawan