K-means Algorithm

Cluster Analysis in Data Mining

Presented by Zijun Zhang

Algorithm Description

- **What is Cluster Analysis?**

Cluster analysis groups data objects based only on information found in data that describes the objects and their relationships.

Goal of Cluster Analysis

The objects within a group be similar to one another and different from the objects in other groups

Algorithm Description

- **Types of Clustering**

  Partitioning and Hierarchical Clustering

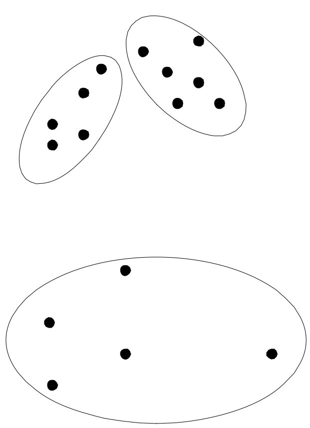
- **Hierarchical Clustering**

  - A set of nested clusters organized as a hierarchical tree

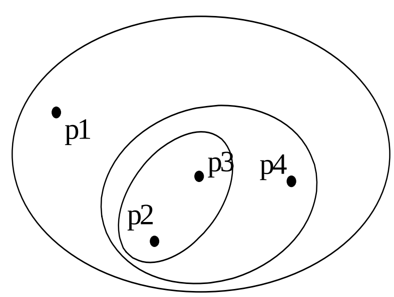
- **Partitioning Clustering**

  - A division data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset

Algorithm Description



A Partitional Clustering



Hierarchical Clustering

Algorithm Description

- **What is K-means?**

  1. Partitional clustering approach
  2. Each cluster is associated with a centroid (center point)
  3. Each point is assigned to the cluster with the closest centroid
  4. Number of clusters, K, must be specified

Algorithm Statement

- **Basic Algorithm of 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

Algorithm Statement

- **Details of K-means**

1. Initial centroids are often chosen randomly.
   - *Clusters produced vary from one run to another*
2. The centroid is (typically) the mean of the points in the cluster.
3. 'Closeness' is measured by **Euclidean distance**, cosine similarity, correlation, etc.
4. K-means will converge for common similarity measures mentioned above.
5. Most of the convergence happens in the first few iterations.
   - *Often the stopping condition is changed to 'Until relatively few points change clusters'*

Algorithm Statement

- **Euclidean Distance**

$$d(i,j)=\sqrt{|x_{i1}-x_{j1}|^2+|x_{i2}-x_{j2}|^2+...+|x_{ip}-x_{jp}|^2}$$

A simple example: Find the distance between two points, the original and the point (3,4)

$$d_E(O,A)=\sqrt{3^2+4^2}=5$$

Algorithm Statement

- **Update Centroid**

We use the following equation to calculate the n dimensional centroid point amid k n-dimensional points

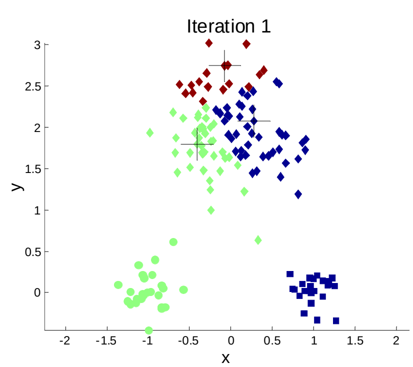
$$CP(x_1, x_2, \ldots, x_k) = \left(\frac{\sum_{i=1}^{k} x1st_i}{k}, \frac{\sum_{i=1}^{k} x2nd_i}{k}, \ldots, \frac{\sum_{i=1}^{k} xnth_i}{k}\right)$$

Example: Find the centroid of 3 2D points, (2,4), (5,2) and (8,9)

$$CP = \left(\frac{2+5+8}{3}, \frac{4+2+9}{3}\right) = (5,5)$$

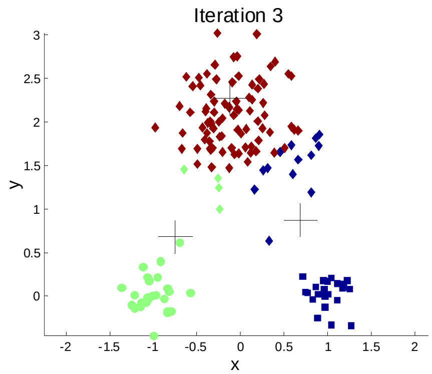
Example of K-means

- **Select three initial centroids**



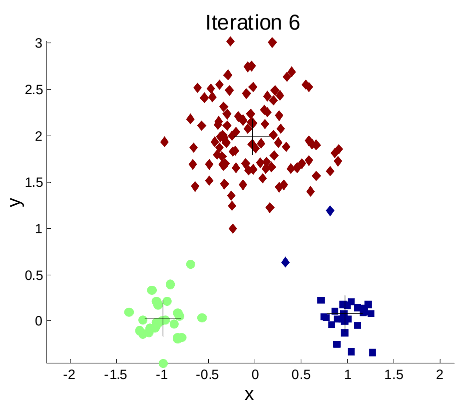
Example of K-means

- **Assigning the points to nearest K clusters and re-compute the centroids**

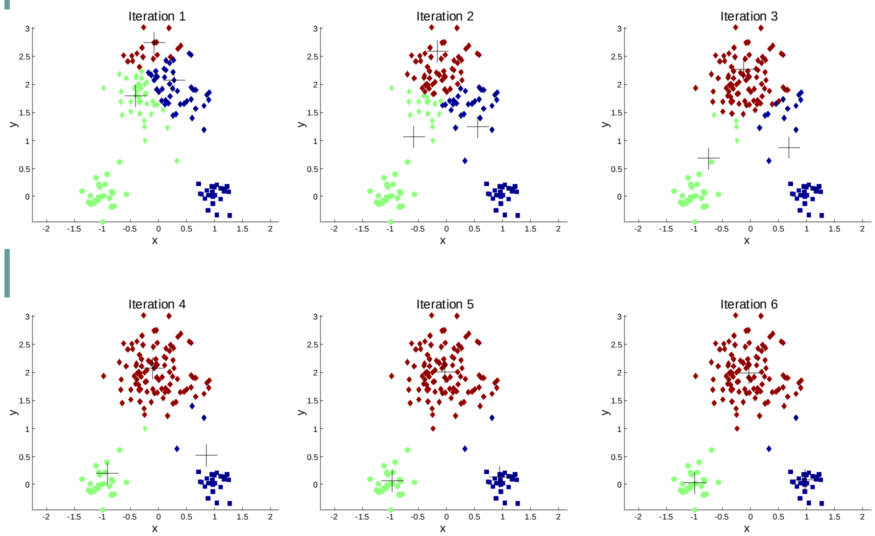


Example of K-means

- **K-means terminates since the centroids converge to certain points and do not change.**

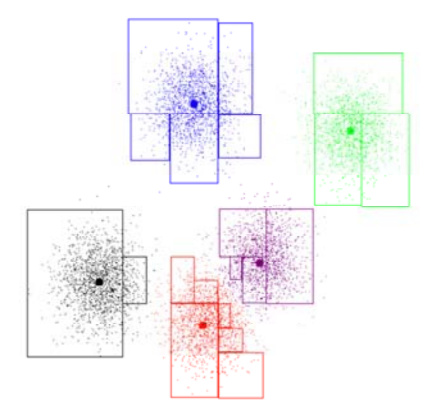


Example of K-means



Example of K-means

- **Demo of K-means**



Evaluating K-means Clusters

- Most common measure is Sum of Squared Error (SSE)
  - For each point, the error is the distance to the nearest cluster
  - To get SSE, we square these errors and sum them.

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

  - *x* is a data point in cluster *C*i and *mi* is the representative point for cluster *C*i
    - can show that *mi* corresponds to the center (mean) of the cluster
  - Given two clusters, we can choose the one with the smallest error
  - One easy way to reduce SSE is to increase K, the number of clusters
    - A good clustering with smaller K can have a lower SSE than a poor clustering with higher K

Problem about K

- **How to choose K?**

1. Use another clustering method, like EM.

2. Run algorithm on data with several different values of K.

3. Use the prior knowledge about the characteristics of the problem.

Problem about initialize centers

- **How to initialize centers?**

  - Random Points in Feature Space

  - Random Points From Data Set

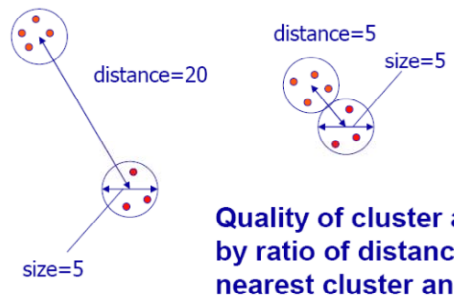
  - Look For Dense Regions of Space

  - Space them uniformly around the feature space

Cluster Quality

- Since any data can be clustered, how do we know our clusters are meaningful?
  - The size (diameter) of the cluster vs. The inter-cluster distance
  - Distance between the members of a cluster and the cluster's center
  - Diameter of the smallest sphere
- The ability to discover some or all of the hidden patterns

Cluster Quality

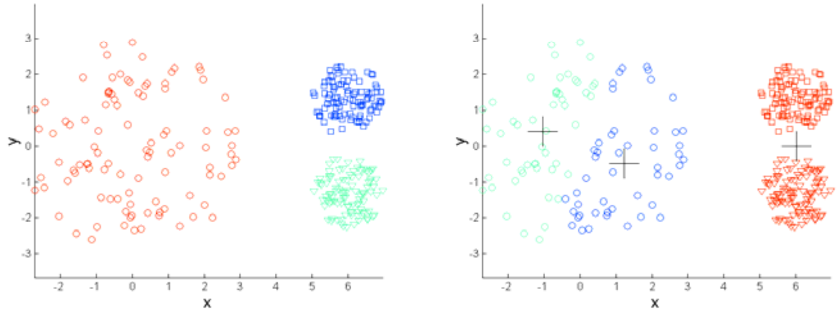


Quality of cluster assessed by ratio of distance to nearest cluster and cluster diameter

Limitation of K-means

- K-means has problems when clusters are of differing
  - Sizes
  - Densities
  - Non-globular shapes
- K-means has problems when the data contains outliers.

Limitation of K-means



Original Points

K-means (3 Clusters)

Application of K-means

- **Image Segmentation**

  The *k*-means clustering algorithm is commonly used in computer vision as a form of image segmentation. The results of the segmentation are used to aid border detection and object recognition.

K-means in Wind Energy

- Clustering can be applied to detect abnormality in wind data (abnormal vibration)
- Monitor Wind Turbine Conditions
- Beneficial to preventative maintenance
- K-means can be more powerful and applicable after appropriate modifications

K-means in Wind Energy

Modified K-means

Repeat until the criterion $d(k, \mathbf{x}, \mathbf{c}) - d(k-1, \mathbf{x}, \mathbf{c}) \le \xi$ is satisfied.

1) Assigning the value of k by $k+1$, the initialization of k is 2 and the maximum of k is 25.

2) Decompose the dataset to 10 sub-datasets with an equal size.

3) Repeat 10 times.

3.1) Randomly select 9 sub-datasets to conduct training dataset and left 1 sub-dataset as the test dataset.

3.2) Initializing k centroids.

3.3) Repeat until the centroids do not move.

3.3.1) Assigning data point to the closest cluster by $C_i^t = \{\mathbf{x}_j : \|\mathbf{x}_j - \mathbf{c}_i^t\| \le \|\mathbf{x}_j - \mathbf{c}_{i^*}^t\|, i^* = 1, 2, \cdots, k\}$.

3.3.2) Updating values of centroids by $\mathbf{c}_i = \sum_{\mathbf{x}_j \in C_i} \mathbf{x}_j / n$

3.4) Compute the clustering cost, d.

4) Estimate the average of clustering cost d in 10-fold cross-validation.

where the ξ is arbitrarily set as 0.001 in this research.

K-means in Wind Energy

- Clustering cost function

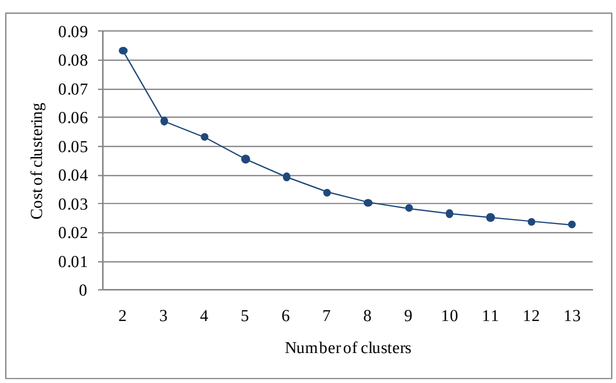
$$d(k,\mathbf{x},\mathbf{c}) = \frac{1}{n}\sum_{i=1}^{k}\left(\sum_{\mathbf{x}_j \in C_i}\left\|\mathbf{x}_j - \mathbf{c}_i\right\|^2\right)$$

$$n = \sum_{i=1}^{k} m_i$$

$$d(k,\mathbf{x},\mathbf{c}) = \frac{1}{\sum_{i=1}^{k} m_i}\sum_{i=1}^{k}\left(\sum_{\mathbf{x}_j \in C_i}\left\|\mathbf{x}_j - \mathbf{c}_i\right\|^2\right)$$

K-means in Wind Energy

- Determination of *k* value



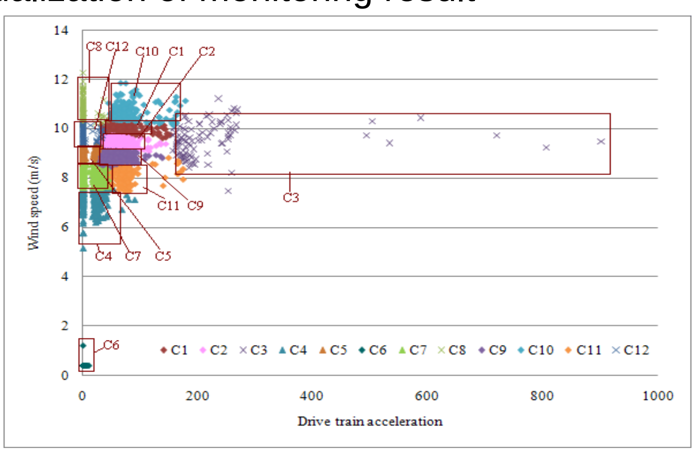
K-means in Wind Energy

- Summary of clustering result

No. of Cluster	c_1 (Drive train acc.)	c_2 (Wind speed)	Number of points	Percentage (%)
1	71.9612	9.97514	313	8.75524
2	65.8387	9.42031	295	8.25175
3	233.9184	9.57990	96	2.68531
4	17.4187	7.13375	240	6.71329
5	3.3706	8.99211	437	12.22378
6	0.3741	0.40378	217	6.06993
7	18.1361	8.09900	410	11.46853
8	0.7684	10.56663	419	11.72028
9	62.0493	8.81445	283	7.91608
10	81.7522	10.67867	181	5.06294
11	83.8067	8.10663	101	2.82517
12	0.9283	9.78571	583	16.30769

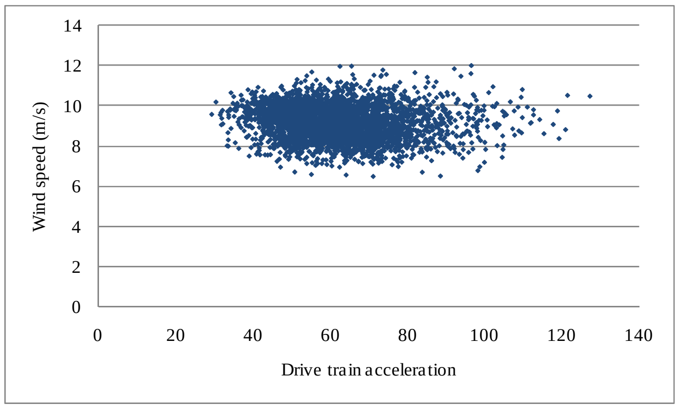
K-means in Wind Energy

- Visualization of monitoring result



K-means in Wind Energy

- Visualization of vibration under normal condition



Reference

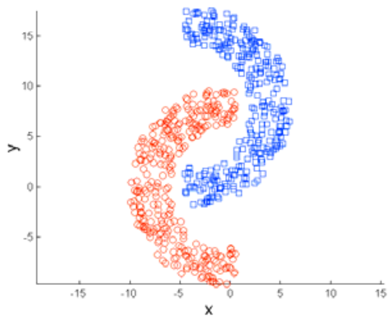
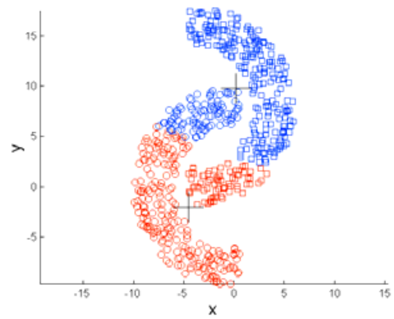
1. Introduction to Data Mining, P.N. Tan, M. Steinbach, V. Kumar, Addison Wesley

2. An efficient k-means clustering algorithm: Analysis and implementation, T. Kanungo, D. M. Mount, N. Netanyahu, C. Piatko, R. Silverman, and A. Y. Wu, IEEE Trans. PatternAnalysis and Machine Intelligence, 24 (2002), 881-892

3. http://www.cs.cmu.edu/~cga/ai-course/kmeans.pdf

4. http://www.cse.msstate.edu/~url/teaching/CSE6633Fall08/lec16%20k-means.pdf

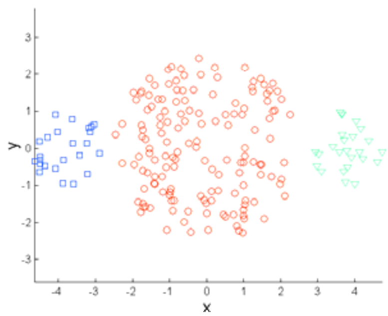
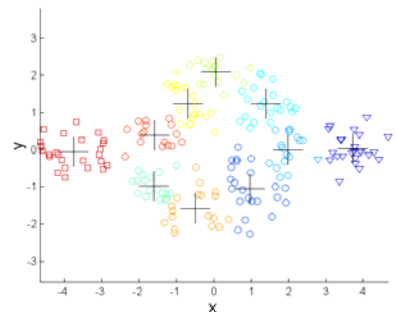
Appendix One

Original Points

K-means (2 Clusters)

Appendix Two

Original Points

K-means Clusters

One solution is to use many clusters.
Find parts of clusters, but need to put together.