Introduction to K-means Algorithm

Wenyan Li (Emily Li)
Sep. 15, 2009

Outline

- Introduction to Clustering Analysis
- K-means Algorithm Description
- Example of K-means Algorithm
- Other Issues of K-means Algorithm
- K-means Algorithm in STATISTICA
Introduction to Clustering Analysis

- **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.

- **Types of 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

![Hierarchical Clustering](image1)

![Partitioning Clustering](image2)
Outline

- Introduction to Clustering Analysis
- K-means Algorithm Description
- Example of K-means Algorithm
- Other Issues of K-means Algorithm
- K-means Algorithm in STATISTICA

K-means Algorithm Description

- What is K-means Algorithm?
  - Partitioning clustering approach;
  - Each cluster is associated with a centroid (center point);
  - Each point is assigned to the cluster with the closest centroid;
  - Number of clusters, K, must be specified
K-means Algorithm Description

- 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

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

- **Euclidean Distance**

  \[ d(i,j) = \sqrt{(x_i - x_j)_1^2 + (x_i - x_j)_2^2 + \ldots + (x_i - x_j)_p^2} \]

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

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

- **Update Centroid**

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

  \[ CP(x_1,x_2,\ldots,x_k) = \left( \frac{\sum x_1st_i}{k}, \frac{\sum x_2nd_i}{k}, \ldots, \frac{\sum xnth_i}{k} \right) \]

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

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

- Introduction to Clustering Analysis
- K-means Algorithm Description
- Example of K-means Algorithm
- Other Issues of K-means Algorithm
- K-means Algorithm in STATISTICA

Example of K-means Algorithm

- Select three initial centroids

![Diagram of K-means Algorithm iteration 1](image-url)
Example of K-means Algorithm

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

Example of K-means Algorithm

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

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.
  \[ SSE = \sum_{i=1}^{K} \sum_{x \in C_i} \text{dist}^2(m_i, x) \]
  - x is a data point in cluster \( C_i \) and \( m_i \) is the representative point for cluster \( C_i \)
  - Can show that \( m_i \) 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 Centers

- How to choose \( K \)?
  - Use another clustering method, like EM.
  - Run algorithm on data with several different values of \( K \).
  - Use the prior knowledge about the characteristics of the problem.

- 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 Algorithm

- K-means has problems when clusters are of differing
  - Sizes
  - Densities
  - Non-globular shapes

- K-means has problems when the data contains outliers.
A Real Example

Using STATISTICA Software

Prepare the Data

- Statistica can read from Excel, .txt and many other types of files
Open an Excel File

- Click the “Import selected sheet to Spreadsheet”
- Select the desired Excel sheet where your data is stored
- Get variable names from the first row

Clustering

- Use the wind energy data set
Clustering

- Select k-Means and choose the variables

- Choose the distance metrics and initial cluster centers
Clustering

- 5 clusters and see the results

![Clustering Diagram]

Clustering

- Centroids (cluster means)

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Operator Station</th>
<th>ARS (Comfort Package)</th>
<th>Remote Cylinder</th>
<th>Power Take-Off</th>
<th>Hitch Quick Coupler and Drawbar</th>
<th>Number of cases</th>
<th>Percentage(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>989 1005</td>
<td>1210</td>
<td>2400</td>
<td>2410</td>
<td>2530</td>
<td>396</td>
<td>30.76059</td>
</tr>
<tr>
<td>2</td>
<td>989 1005</td>
<td>1210</td>
<td>2400</td>
<td>2410</td>
<td>2530</td>
<td>420</td>
<td>32.04469</td>
</tr>
<tr>
<td>3</td>
<td>991 1005</td>
<td>1210</td>
<td>2400</td>
<td>2410</td>
<td>2530</td>
<td>150</td>
<td>11.95173</td>
</tr>
<tr>
<td>4</td>
<td>986 1005</td>
<td>1210</td>
<td>2300</td>
<td>2410</td>
<td>2530</td>
<td>152</td>
<td>11.99909</td>
</tr>
<tr>
<td>5</td>
<td>989 1005</td>
<td>1210</td>
<td>2300</td>
<td>2420</td>
<td>2530</td>
<td>163</td>
<td>12.94469</td>
</tr>
</tbody>
</table>
Clustering

- Members and their distance to the centroids

<table>
<thead>
<tr>
<th>Case No.</th>
<th>Operator Station</th>
<th>AHS</th>
<th>Comfort Package</th>
<th>Remote Cylinder Control</th>
<th>Power Take-Off</th>
<th>Hitch Quick Coupler and Drawbar</th>
<th>Distance to centroid</th>
</tr>
</thead>
<tbody>
<tr>
<td>67</td>
<td>3</td>
<td>991</td>
<td>1002</td>
<td>1200</td>
<td>2400</td>
<td>2410</td>
<td>2550 1.00000</td>
</tr>
<tr>
<td>68</td>
<td>2</td>
<td>999</td>
<td>1005</td>
<td>1200</td>
<td>2400</td>
<td>2410</td>
<td>2550 1.00000</td>
</tr>
<tr>
<td>69</td>
<td>2</td>
<td>999</td>
<td>1005</td>
<td>1200</td>
<td>2400</td>
<td>2410</td>
<td>2550 1.00000</td>
</tr>
<tr>
<td>70</td>
<td>2</td>
<td>999</td>
<td>1005</td>
<td>1200</td>
<td>2400</td>
<td>2410</td>
<td>2550 1.00000</td>
</tr>
<tr>
<td>71</td>
<td>2</td>
<td>999</td>
<td>1005</td>
<td>1200</td>
<td>2400</td>
<td>2410</td>
<td>2550 1.00000</td>
</tr>
<tr>
<td>72</td>
<td>2</td>
<td>999</td>
<td>1005</td>
<td>1200</td>
<td>2400</td>
<td>2410</td>
<td>2550 1.00000</td>
</tr>
<tr>
<td>73</td>
<td>2</td>
<td>999</td>
<td>1005</td>
<td>1200</td>
<td>2400</td>
<td>2410</td>
<td>2550 1.00000</td>
</tr>
<tr>
<td>74</td>
<td>2</td>
<td>999</td>
<td>1005</td>
<td>1200</td>
<td>2400</td>
<td>2410</td>
<td>2550 1.00000</td>
</tr>
<tr>
<td>75</td>
<td>2</td>
<td>999</td>
<td>1005</td>
<td>1200</td>
<td>2400</td>
<td>2410</td>
<td>2550 1.00000</td>
</tr>
<tr>
<td>76</td>
<td>5</td>
<td>999</td>
<td>1005</td>
<td>1200</td>
<td>2400</td>
<td>2410</td>
<td>2550 1.00000</td>
</tr>
<tr>
<td>77</td>
<td>2</td>
<td>999</td>
<td>1005</td>
<td>1200</td>
<td>2400</td>
<td>2410</td>
<td>2550 1.00000</td>
</tr>
<tr>
<td>78</td>
<td>2</td>
<td>999</td>
<td>1005</td>
<td>1200</td>
<td>2400</td>
<td>2410</td>
<td>2550 1.00000</td>
</tr>
<tr>
<td>79</td>
<td>2</td>
<td>999</td>
<td>1005</td>
<td>1200</td>
<td>2400</td>
<td>2410</td>
<td>2550 1.00000</td>
</tr>
<tr>
<td>80</td>
<td>2</td>
<td>999</td>
<td>1005</td>
<td>1200</td>
<td>2400</td>
<td>2410</td>
<td>2550 1.00000</td>
</tr>
<tr>
<td>81</td>
<td>2</td>
<td>999</td>
<td>1005</td>
<td>1200</td>
<td>2400</td>
<td>2410</td>
<td>2550 1.00000</td>
</tr>
</tbody>
</table>

Software Demonstration
Thank You