Data Mining

Clustering I – K-Means Algorithm (Part A)

Dr. Jason T.L. Wang, Professor

Department of Computer Science

New Jersey Institute of Technology

/

NJIT

New Jersey's Science & Technology University

THE EDGE IN KNOWLEDGE

Clustering: Its Definition

Given a set of objects, the clustering process is to group the objects into clusters where similar objects are in the same cluster and dissimilar objects are in different clusters.

Clustering is an unsupervised learning process since it does not require any predefined training classes.

Applications of Clustering

- Market planning and research (clustering customers based on their profiles)
- Bioinformatics data analysis (clustering protein sequences into families)
- World Wide Web (clustering visitors based on their access patterns)

Input/Output of Clustering

The input of a clustering algorithm is
a set of objects and pairwise distances (dissimilarity or similarity) among the objects, and a user-specified parameter value k.

The output of the clustering algorithm is
k clusters.

Types of data in clustering analysis

- P-dimensional points (vectors) in Euclidean space
- Nominal variables
- Strings, trees, graphs and other objects
- Data of mixed types

Distance Between Euclidean Points

- *Minkowski distance*:

$$d(i,j) = \sqrt[q]{(|x_{i_1} - x_{j_1}|^q + |x_{i_2} - x_{j_2}|^q + ... + |x_{i_p} - x_{j_p}|^q)}$$

where $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ are two p-dimensional points.

- If $q = 1$, d is Manhattan distance.

$$d(i,j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + ... + |x_{i_p} - x_{j_p}|$$

Distance between Points

- *If q = 2*, *d* is Euclidean distance:

$$d(i,j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + ... + |x_{i_p} - x_{j_p}|^2)}$$

  - Metric properties
    - $d(i,j) \geq 0$
    - $d(i,i) = 0$
    - $d(i,j) = d(j,i)$
    - $d(i,j) \leq d(i,k) + d(k,j)$

Nominal Variables

- A nominal variable can take several states, e.g., red, yellow, blue, green.
- Distance:
  - *m*: # of matches, *p*: total # of variables

$$d(i, j) = \frac{p - m}{p}$$

End of

K-Means Clustering Algorithm Module (Part A)