The Clustering Approach for Nutrient Foods

Swe Swe Myint
University of Computer Studies, Mandalay
mday.242ko @gmail.com, sweswemyint2009@gmail.com

Abstract

Clustering is the process of grouping the data into classes of similar objects. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. Our system will implement the clustering of nutrient foods by using the k-mean partitioning method. Each cluster's center is presented by the mean value of the objects in the cluster. The k-means algorithm is by far the most widely used method for discovering clusters in data. This paper presents our system and shows how to accelerate it dramatically, while still always computing exactly the same result as the standard algorithm. The accelerated algorithm avoids unnecessary distance calculations by Euclidean distance measurements between each pair of objects. This system focuses on nutrient foods dataset, which contains 253 instances and eleven attributes from UCI machine learning repository.

1. Introduction

Knowledge Discovery in Databases (KDD) is the non-trivial extraction of implicit, previously unknown, and potentially useful information from databases [1].

Data mining is the process of digging or gathering information from various databases. The data mining should have been more appropriately names knowledge mining from data. Data mining involves the use of sophisticated data analysis tool to discover previously unknown, valid patterns and relationships in large data sets [2]. These tools can include statistical models, mathematical algorithms, and machine learning methods. Consequently, data mining consists of more then collecting and managing data; it also includes analysis and prediction.

Data mining can be performed on data represented in quantitative, textual, or multimedia forms. Data mining applications can use a variety of parameter to examine the data. They include association, sequence or path analysis, classification, clustering and forecasting [2].

2. Related Work and Motivation

Since k-means has historically been applied to small data sets, the state of the practice appears to be to try out various random starting points. Traditionally, k-means is used to initialize more expensive algorithms such as EM [B95]. In fact, other methods to initialize EM have been used, including hierarchical agglomerative clustering (HAC) [DH73, R92] to set the initial points. Hence out choice of comparing against the random starting points approach. However, regardless of where a starting point comes from, be it prior knowledge or some other initialization scheme, our method can be used as an effective and efficient scalable means to proceed to a solution.

In statistics, all schemes we are aware of appear to be memory-based. A book dedicated to the topic of clustering large data sets [KR89] presents algorithm CLARA for clustering “large databases”. However the algorithm is limited to 3500 cases maximum [KR89, p.126]. The only options available in this literature to scale to large databases are random sampling and on-line k-means [M67]. We compare against both these methods in Section 6. On-line k-means essentially works with a memory buffer of one case. As we show in the results section, this methods does not compare well with other alternatives. Other scalable clustering schemes include CLARANS [NH94] and DBSCAN [EKX95]. The latter two are targeted at clustering spatial data primarily. In [ZRL97] they compare against these schemes and demonstrate higher efficiency [3].

The major reason that data mining has attracted a great deal of attention in the information industry in recent years is due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. The information and knowledge gained can be used for applications ranging from business management, production control, and market analysis, to engineering design and science exploration.

Data mining can be viewed as a result of the natural evolution of information technology. An evolutionary path has been witnessed in the database industry in the development of the following functionalities: data collection and database

creation, data management and data analysis and understanding. Data can now be stored in many different types of databases. Once database architecture that has recently emerged is the data warehouse, a repository of multiple heterogeneous data sources, organized under a unified schema at a single site in order to facilitate management decision making.

Data warehouse technology includes data cleansing, data integration, and On-Line Analytical Processing (OLAP), that is, analysis techniques with functionalities such as summarization, consolidation, and aggregation, as well as the ability to view information from different angles. Although OLAP tools support multidimensional analysis and decision making, additional data analysis tools are required for in-depth analysis, such as data classification, clustering, and the characterization of data changes over time [2].

3. Clustering

Clustering analyzes data objects without consulting a known class label. In general, the class labels are not present in the training data simply because they are not known to begin with. Clustering can be used to generate such labels. The objects are clustered or grouped based on the principle of maximizing the intra-class similarity and minimizing the inter-class similarity. That is, clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters. Each cluster that is formed can be viewed as a class of objects, from which rules can be derived. Clustering can also facilitate taxonomy formation, that is, the organization of observations into a hierarchy of classes that group similar events together [2].

3.1. Cluster Analysis

Cluster analysis seeks to find groups of closely related observations so that observations that belong to the same cluster are more similar to each other than observations that belong to other clusters [4].

Cluster analysis is an important human activity. Early in childhood, one learns how to distinguish between cats and dogs, or between animals and plants, by continuously improving subconscious clustering schemes. Cluster analysis has been widely used in numerous applications, including pattern recognition, data analysis, image processing, and market research. By clustering, one can identify dense and sparse regions and, therefore, discover overall distribution patterns and interesting correlation among data attribution [2]. Clustering methods partition a set of objects into clusters such that objects in the same cluster are more similar to each other than objects in different clusters according to some defined criteria [5].

The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. A cluster of data objects can be treated collectively as one group in many applications.

3.2. Euclidean Distance

Interval-scaled variables are continuous measurements of a roughly linear scale. Typical examples include protein, fat, carbohydrate, calcium, iron, vitamin A, vitamin B1, vitamin B2, vitamin B3 (Niacin), and vitamin C. The measurement unit used can affect the clustering analysis.

The dissimilarity (or similarity) between the objects described by interval-scaled variables is typically computed based on the distance between each pair of objects [1]. The most popular distance measure is Euclidean distance**,** which is defined as

$$d(i,j)=\sqrt{|x_{i1}-x_{j1}|^2+|x_{i2}-x_{j2}|^2+...+|x_{ip}-x_{jp}|^2} \qquad (1)$$

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

4. Overview of the System

Clustering partitions a set of objects into non-overlapping subsets called clusters such that the objects inside each cluster are similar to each other and the objects from different clusters are not similar. The set of non-overlapping clusters is called a partition. The idea behind clustering was that if certain documents match a user query, the documents in the same cluster also are likely to be relevant.

The user can view the analysis result of clustering as original import dataset. The system accepts k - value from the user and defines initialized clusters centre. The system can process until cluster means have no change. Then, the system reassigns objects to clusters based on the distance between the objects and the clusters means. All the points are assigned; recalculate the cluster's means.

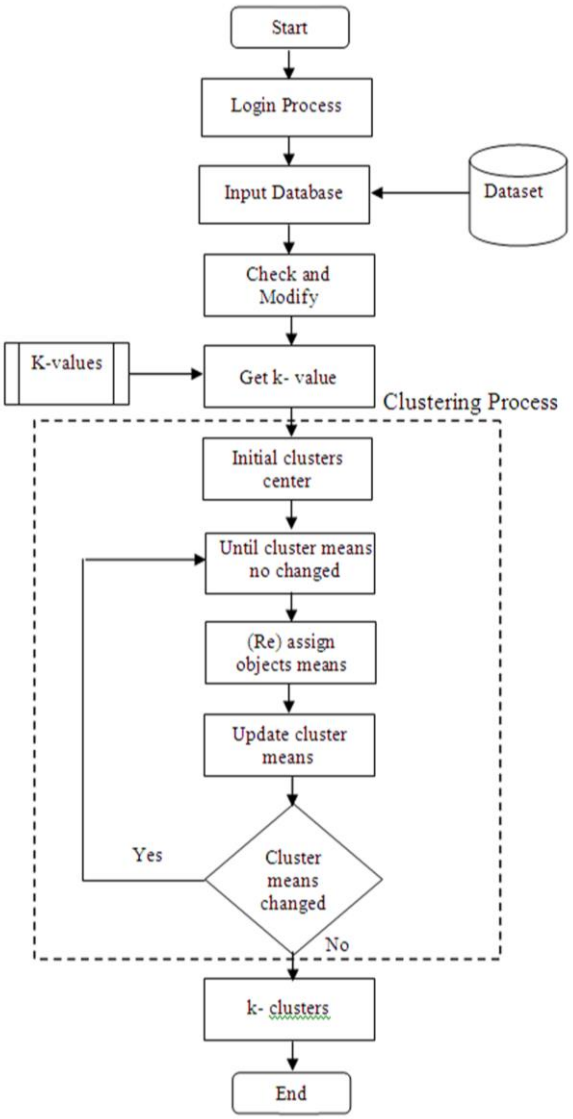

Figure 1. System flow diagram

5. Algorithm of the System

Algorithm: k-mean. The k-means algorithm for partitioning based on the mean value of the objects in the cluster.
Input: The number of clusters k and a database containing n objects.
Output: A set of k clusters that minimizes the squared-error criterion.

Begin
initial cluster ← arbitrarily choose k-objects;
repeat
assign each object to the cluster to which the object is the most similar (based on the mean value of the objects in the cluster);
update the cluster means; (calculate the mean value of the objects for each cluster)
until no change;
End [5]

5.1. Attributes Information of the Training Dataset

This system focuses on nutrient foods dataset, which contains 253 instances and eleven attributes from UCI machine learning repository. This dataset contains food name and their attributes are minerals and vitamins.

To be more detail, each nutrient foods data items contains the attributes protein, fat, carbohydrate, calcium, iron, vitamin A, vitamin B1, vitamin B2, vitamin B3 (Niacin), vitamin C. This data entry dataset must have the same data unit such as milligram used in our system. For example, the amount of protein in Bean sprout is 13 mg, fat is 0.81 mg, carbohydrate is 19.59 mg, calcium is 0.109 mg, iron is 0.0082 mg, vitamin A is 0 mg, vitamin B1 is 0.08 mg, vitamin B2 is 0.08 mg, vitamin B3 (Niacin) is 0.53 mg, vitamin C is 15 mg.

The k-means algorithm proceeds as follows. First, it randomly selects k of the objects, each of which initially represents a cluster mean or centre. For each of the remaining objects, an object is assigned to the cluster to which it is the most similar, based on the distance between the object and the cluster mean. It then computes the new mean for each cluster. This process iterates until the criterion function converges. Typically, the Euclidean distance is used.

6. Implementation

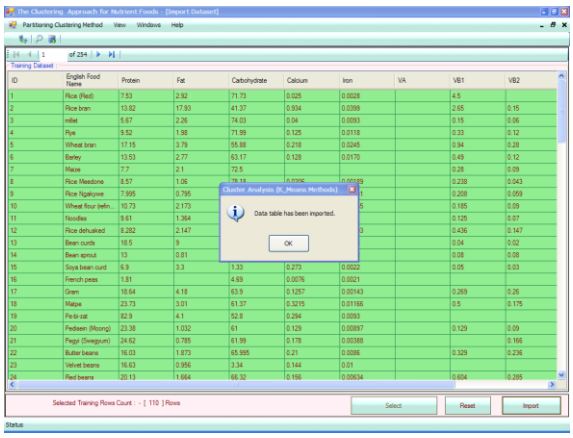

Figure 2. Import Dataset

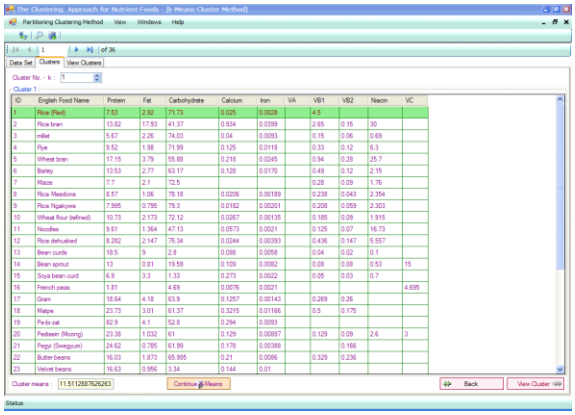

Figure 3. Clustering analysis

Figure 4. Clustering the data

The system can clustering based on the user input cluster count as described in figure 2.

Figure 2. show the user import dataset and can select the user's requirement records from the training dataset.

If the input training dataset is correct, the system can calculate the k-means.

7. Conclusion

In this paper, we study the concepts of clustering analysis and the k-means algorithm. And then, analyze the data objects without consulting a known class label and realize the clustering of the large amount of data.

This paper represents a simple and efficient clustering algorithm based on the k-mean method. This algorithm is easy to implement, requiring a simple data structure to keep some information in each iteration to be used in the next iteration. The experimental results demonstrated that nutrient foods can improve the computational speed of algorithms by the magnitude in the total number of distant calculation and the overall time of computation.

References

[1] Martin Ester, Hans-Peter Kriegel, Xiaowei Xu, "A Database Interface for Clustering in Large Spatial Databases", *Institute for Computer Science, University of Munich. 1st International Conference* on Knowledge Discovery and Data Mining (KDD-95).

[2] Jiawei Han & Micheline Kamber, *Data Mining Concepts and Techniques*, ISBN 1- 55860-489-8, Morgan Kaufmann Publishers, 2001.

[3] P.S. Bradley, Usama Fayyad, and Cory Reina "Scaling Clustering Algorithms to Large Databases", *Miscrosoft Research*, Redmond, WA 98052, USA {bradley, fayyad, coryr} @miscrosoft.com.

[4] Pang-Ning Tan Michale Steinbach Vipin Kuma, *Introduction to Data Mining*, March 25, 2006.

[5] Zhexue Huang, "A fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining", CSIRO *Mathematical and Information Sciences*, GPO Box 664, Canberra2601, AUSTRALIA.