HIERARCHICAL CLUSTERING FOR FRUITS USING AGGLOMERATIVE CLUSTERING

Ma May Zin Oo

Computer University (Mandalay)

Abstract

This thesis deals with the clustering of hierarchical method in data mining. Clustering is the process of grouping the data into classes of similar objects. A hierarchical clustering methods work by grouping data objects into tree of clusters. The system is intended to cluster fruits by using agglomerative hierarchical clustering method. The system can be used to known various fruits with same features. The system can be applied to find information quickly and less time-consuming about fruits that the user requests.

1.Introduction

Data mining has attracted a great deal of attention in the information industry and in society as a whole in recent years, due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. Data mining is the process of extracting or mining knowledge from large amount of data. Data mining applications can use a variety of parameter to examine the data. They include association, classification, clustering, forecasting and sequence of path analysis.

Clustering is one of the fundamental operations in data mining. Clustering can be defined as the process of organizing objects in a database into clusters such that objects within the same cluster have a high degree of similarity, while objects belonging to different clusters have a low degree of dissimilarity.

Hierarchical clustering methods work by grouping data objects into a tree of cluster. There are two methods, agglomerative clustering and divisive clustering, based on how the hierarchical decomposition is formed.

In this system, agglomerative hierarchical clustering is used to know group of fruits with the same features. The agglomerative approach builds the hierarchy from bottom-up. It starts with the data objects as individual clusters and successively merges the most similar pair of clusters until all the clusters are merged into one cluster which is the topmost level of the

hierarchy. The cluster is merged according to some principal; such as the minimum Euclidean distance between closets neighboring objects in the cluster.

2.Theory Background

Data mining is the process of digging or gathering information from various databases. The data mining should have been more appropriately named knowledge mining from data. Data mining is an iterative process of discovery. Data mining seeks to find new patterns hidden in the data stored in large databases. Data mining has two principal activities: finding patterns in data and describing those patterns clearly.

Data mining is the task of discovering interesting patterns from large amounts of data, where the data can be stored in databases, data warehouses, or other information repositories. It is a young interdisciplinary field, drawing from areas such as database systems, data warehousing, statistics, machine learning, data visualization information retrieval, and high-performance computing. Other contributing areas include neural networks, pattern recognition, spatial data analysis, image databases, signal processing, and many application fields, such as business, economics, and bioinformatics.

Data mining is one of the fastest growing fields in the computer industry. Since data mining is a natural activity to be performed on large data sets, one of the largest target markets is the entire data warehousing, decision-support community, encompassing professionals from such industries as retail, manufacturing, telecommunications, healthcare, insurance and transportation.

Data mining is the process of analyzing large data sets in order to find patterns that can help to isolate key variables to build predictive models for management decision-making. In essence, data mining helps businesses to optimize their processes so that their customers receive the most relevant services and the costs of serving them are proportionate to the value of the profits earned from them, and a company’s exposure to risk is proportionate to premiums earned, etc. Data mining enables companies to segment their customer base and to tailor products and services to needs and purchasing power of individual groups of customers.

Unlike classification and prediction, which analyzes class-labeled data objects, 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 intraclass similarity and minimizing the interclass similarity.

3.Agglomerative Hierarchical Clustering

Clustering is the process of grouping the data into classes or clusters so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters. Dissimilarities are assessed based on the attribute values describing the objects. Often distance measures are used.

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. Cluster analysis divides data into groups (cluster) that are meaningful, useful, or both. If meaningful groups are the goal, then the clusters should capture the nature structure of the data. In some cases, however, cluster analysis is only a useful starting point for other purposes, such as data summarization.

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 wide applications, including market or customer segmentation, pattern recognition, biological studies, spatial data analysis, web document classification. Cluster analysis can be used as a stand-alone data mining tool to gain insight into the data distribution or can serve as a preprocessing step for other data mining algorithms operating on the detected clusters.

The explosion of sensory and textual information available to us today has caused many data analysts to turn to clustering algorithms to make sense of the data. It has become a primary tool for so-called knowledge discovery, data mining, and intelligent data analysis. In fact, the massively-sized data sets of these applications have placed high demands on the performance of the computationally expensive clustering algorithms.

Clustering is used in various applications. In general, it can assist in:

- Formulating hypotheses concerning the origin of the data

- Describing the data in terms of a typology
- Predicting the future behavior of types of this data
- Optimizing a functional process

Hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters. Strategies for hierarchical clustering generally fall into two types: Agglomerative and Divisive. Agglomerative hierarchical clustering is a "bottom up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy. Divisive hierarchical clustering is a "top down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.The quality of a pure hierarchical clustering method suffers from its inability to perform adjustment once a merge or split decision has been executed. Recent studies have emphasized the integration of hierarchical agglomeration with iterative relocation methods.

- A hierarchical clustering algorithm does not generate a set of disjoint clusters. Instead, it generates a hierarchy of nested clusters that can be represented by a tree, called a dendrogram. The dendrogram is constructed as a sequence of partitions such that its root is a cluster covering all the points and the leaves are clusters containing only one point. In the middle, child clusters partition the points assigned to their common parent according to a dissimilarity level. The dendrogram is most useful up to a few levels deep, as the clustering becomes more trivial as the tree depth increases. Dendrogram illustrates the fusions or divisions made at each successive stage of analysis.
- In fact, the hierarchical methods are particularly favored by the biologists because they may give more insights to the structure of the clusters than the other methods.

4. System Design and Implementation

In this system, user can import fruit data table from the database. The system normalizes data to compute distance. And then, user can select attributes of fruits that user requests. The system calculates group of fruits by using the agglomerative algorithm. Then, user can see final result by tree view.

At the start of this system, main menu form displays as in Figure 4.13. There are four main menus in the system. They are Fruit, Clustering, About, Exit. Fruit menu has Fruit Data Table menu. If the user clicks Fruit Data Table for User menu, FruitList Form will appear as shown in Figure 4.14. If the user clicks Fruit Data Table for Admin, FruitList form will appear as shown in Figure4.15. In Clustering menu, user can cluster fruit datasets according to agglomerative approach. In About menu, user can see information of the system. It the user wants to exit the system, user needs to choose Exit menu.

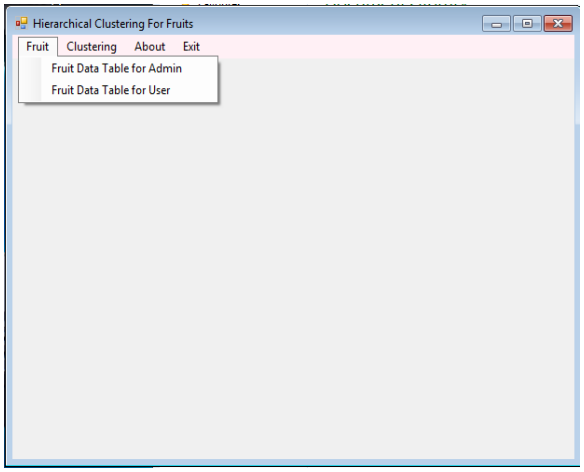

Figure 4.13 Main Menu Form

5. Conclusions, Limitations and Further Extension

The system presents hierarchical clustering for fruits by using Agglomerative algorithm, Single linkage method and Euclidean distance function. Agglomerative hierarchical clustering builds higher quality trees and is fast enough to be used even on very large data sets. The system is intended to implement clustered fruits. There are many books that are written about fruits. In the system, there are six attributes. They are fruits' name, taste, property, advantage, treatable disease and remark. The system is intended to support the user to know fruits with the same features. The advantage of the system is that user can create new database of fruits to view cluster output. When the doctors from the traditional hospital indicate suitable fruits for their patients, the system

can also be applied. This system provides a useful cluster group and helps the user to find information more quickly and less time-consuming about fruits that user requests. C#. NET is used to implement the system.

There are some limitations to implement this system. In this system, fruits are specified by five attributes although fruits have many attributes. According to Agglomerative theory, result will represent by dendrogram. This system is represented by tree view because of programming experience in graphic. This system is implemented only by using Microsoft Visual Studio. NET (or) Visual C#. NET. This system has been implemented on window-based operating system.

This system can be added with various clustering methods and complete clustering methods such as complete linkage, average linkage and centroid linkage. For further developments, it needs to be able to use for all other methods such as partitioning methods, density-based methods, grid-based methods, and model-based methods for the various datasets to compare hierarchical methods.

6.References

[1] A.M. Hafez, "Knowledge Discovery in Database", Department of Computer Science and Automatic Control Faculty of Engineering Alexandria University.

[2] Anna Szymkowiak and Jan Lars Kai Hansen, "Hierarchical Clustering for Data Mining."

[3] B.C.M.Fung, Ke Wang, and M.Ester, "Heierarchical Document Clustering", Simon raser University, Canada.

[4] C.J.C.Burges, "Data Mining and Knowledge Discovery."

[5] Hans-Joachin Mucha and Hizir Sofyan "Cluster Analysis."

[6] J.Han, M.Kamber, "Data Mining Concepts and Techniques."

[7] Ke Chen, “Hierarchical Clustering”

[8] Osmar R.Zaiane "Principles of Knowledge Discovery in Database Chapter 8: Data Clustering"

[9] Pang-Ning, Jan Micheal Steinbach and Vipin Kumar "Introduction to Data Mining."

[10] Vera Marinova-Boncheva " Using the Agglomerative Method of Hierarchical Clustering as a Data Mining Tool in Capital Market"