Chapter 2
Modified Single Pass Clustering Algorithm Based on Median as a Threshold Similarity Value

Mamta Mittal
G. B. Pant Govt. Engineering College, India

V.P. Singh
Thapar University, India

R. K. Sharma
Thapar University, India

Lalit Mohan Goyal
Bharati Vidyapeeth College of Enineering, India

ABSTRACT

Clustering is one of the data mining techniques that investigates these data resources for hidden patterns. Many clustering algorithms are available in literature. This chapter emphasizes on partitioning based methods and is an attempt towards developing clustering algorithms that can efficiently detect clusters. In partitioning based methods, k-means and single pass clustering are popular clustering algorithms but they have several limitations. To overcome the limitations of these algorithms, a Modified Single Pass Clustering (MSPC) algorithm has been proposed in this work. It revolves around the proposition of a threshold similarity value. This is not a user defined parameter; instead, it is a function of data objects left to be clustered. In our experiments, this threshold similarity value is taken as median of the paired distance of all data objects left to be clustered. To assess the performance of MSPC algorithm, five experiments for k-means, SPC and MSPC algorithms have been carried out on artificial and real datasets.

INTRODUCTION

Today, every organization is dealing with data repository systems like relational databases, data warehouses, temporal databases, transactional databases, spatial databases, multimedia databases or the World Wide Web, but a lot of them are not able to take advantage of their huge repositories. Data to be stored is often diverse in nature ranging from scientific to medical, geographic to demographic, financial to marketing as well as the volume of data is so high that human analyst cannot predict it without special tools. To automatically understand and analyze the data effectively and efficiently, the field of data min-

DOI: 10.4018/978-1-5225-0489-4.ch002

ing has emerged in recent years. One of the data mining tools that can be used to group the data objects into unknown classes is clustering. The goal of clustering is to discover the natural grouping among data objects such that the data objects in the same group are similar to one another and dissimilar to the data objects in other groups. Intensive research has been carried out in this field and many algorithms have been proposed. But, clustering is an NP-hard problem due to which the existing approaches have some limitations. To deal with the limitations of existing methods, research is continuously being done in this area. Beside this, collaborating Filtering which has its roots in data mining has now become recent research area for the researcher. It is a method of automatic filtering about the user's interest or likeness by collecting their preferences from many users (collaborating). Use of clustering helps a lot in the collaborating filtering as clustering gathers the information of similar liking users in one group and dissimilar liking users in others groups. When groups or clusters of same liking will be available then based on the user interest or user participation collaborating filtering further predict them.

Data Mining is so popular because it is used to mine interesting data from a large amount of data akin to the extraction of minerals from mineral ores. Most international organizations produce high amounts of information that could never be read by any person in a lifetime. The situation is even more alarming in the world wide networks. These days, gigabytes of data are distributed and exchanged over the world, the existing database management system allow retrieval of data but provide no tools to analyze it. Analysis is beneficial for unearthing the hidden relationships among the data. Data mining is one of the data analysis tools. It goes beyond the idea of conventional data analysis. It uses traditional analysis tools like statistics and graphics in conjunction with those associated with the field of artificial intelligence such as rule induction and artificial neural networks. It is an amalgam of all of these, but still somehow different.

Data mining is a distinctive approach towards the usual data analysis in the sense that the emphasis is not as much on extracting the facts as on generating the hypotheses. It is also capable of generating new business opportunities, the only condition being the provision of databases of sufficient size and acceptable quality. It is popular as it has the following capabilities:

- **Prediction of Trends and Behaviors**: It automates the tedious process of finding predictive behavior or information in huge databases. Before the emergence of data mining field, queries required excessive hands-on analysis, but now they can respond quickly and can be automated. Data mining analyzes past data to identify future trends. A common example is to know the future trends of the stock market.
- **Automated Discovery of Previously Unknown Patterns**: Data mining tools play a great role in identifying patterns which were previously hidden by sweeping through the database. A common example is to identify correlated products from the data of retail sales.

The field of data mining encompasses mainly three techniques: association rule mining, classification and clustering. These techniques are explained briefly as:

*A*ssociation *R*ule *M*ining (*ARM*) generates an implication between two or more data objects of a database. In a transactional database for the given items, an association rule, $X \overset{m\_sup}{\Rightarrow} Y$, is an implication where X and Y are disjoint sets of items; *m_sup* is the value of minimum support. The meaning of such an implication is that *m_sup*% transactions of a database which contain X also contain Y. For example, 90% (*m_sup*%) of the customers who purchase milk and eggs ($X = \{milk, eggs\}$) also purchase apples