Performance Analysis & Comparison b/w Enhanced K-Means & Orthogonal Partitioning (OC), based on proposed New Approach: "DRID"

Rimmy Chuchra
M.tech (Computer Science)
Lovely Professional University
Phagwara, India

Sanjay Sood
(Department of IT)

Karanvir Kaur
Assistant Professor (Computer Science)
Lovely Professional University
Phagwara, India

Summary
Clustering provides grouping together of similar data items. This technique provides a high level view of the database. Clustering technique is a technique that merges and combines techniques from different disciplines such as mathematics, physics, math-programming, statistics, computer sciences, artificial intelligence and databases etc. Variety of clustering algorithms exists and belongs to several different categories. Two Prominent Categories are Distance based and Grid based (example: - K-Means and OC partitioning clustering algorithms) respectively. K-means clustering algorithm is fast, easy to implement, converges to local optima almost surely but there is one drawback behind k-means clustering algorithm that is it is easily effected by noise. On the other hand, O-Cluster algorithm creates a hierarchical grid-based clustering model, that is, it creates axis-parallel (orthogonal) partitions in the input attribute space. The resulting hierarchical structure represents an irregular grid that tessellates the attribute space into clusters. O-Cluster separates areas of high density by placing cutting planes through areas of low density. O-Cluster needs multi-model histograms (peaks and valleys). If an area has projections with uniform or monotonically changing density, then faces difficulty in partition. In this research paper we proposes a new algorithm for merging of clusters and proposed algorithm name is "DRID" which follows a common strategy in two different- different environments and then performance analysis of a distance based algorithm and Grid based algorithm and compare the results. The reason for merging of clusters is to improve the quality of clusters, reduce noise problem and increase the performance and reduce the execution time.

Keywords: *Clustering algorithms, text mining*.

1. Introduction

Data mining is a technique which is used for the extraction of hidden predictive information from large databases. The goal of data mining is to extract knowledge from a data set in a human-understandable language. There are various types of data mining available like Text mining, Graph mining, Spatial mining, Web mining, Structure mining, Content mining, Link mining etc. Here, we are working on Text type data mining which holds natural form of text and the process of deriving high quality data from it is called "Text mining". Basically Text mining gives a structure to the input text and derives patterns from the structured data. Text Mining consists of various tasks like Text Clustering, Text Classification, Document Summarization, Sentiment analysis etc. Two major advantage of using Text mining are: Visualization, In visualization, either complete text is in structured from or non-structured form and the second advantage is that the user can create own custom stop list for performing a better text mining. With increasing the growth of technical world various data mining techniques are available like Clustering, Classification, Prediction, and Association Rules. Clustering defines identification of similar classes of objects. Cluster analysis holds several types of methods like Partitioning methods, Hierarchical methods, Grid Based methods, Model based methods etc. Every Method in cluster analysis supports different algorithms. I am using Enhanced K-means Clustering algorithm which comes under Partitioning methods having distance based environment and OC (Orthogonal Partitioning) Clustering algorithm comes under Grid Based methods having grid based environment. Both algorithms belongs two different environments. The Purpose of Enhanced K-Means Clustering algorithm is to find out better initial centroids with reduced time complexity and whole working of this algorithm is based on K-Means Clustering algorithm where K-Means Clustering Algorithm is Distance based Clustering algorithm which defines distance measures from data instances and also find partitions of the distances as like distance between objects within same clusters is minimized and between different clusters is maximized. The purpose of Orthogonal Partitioning (OC)Clustering algorithm creates a hierarchical grid-based clustering model, which means, it creates axis-parallel(orthogonal) partitions in the input attribute space. O-Cluster separates areas of high density by placing cutting planes through areas of low density. The clusters discovered by O-Cluster

Manuscript received March 5, 2012
Manuscript revised March 20, 2012

are used to generate a Bayesian probability model that is then used during scoring (model apply) for assigning data points to clusters. It reads the data in batches (the default batch size is 50000). The advantages of using Grid based Environment is objects are represented in multi- resolution grid form with higher processing time and independent number of objects.

2. Theoratical Consideration

Clustering algorithms are to be categorizes into several categories. Two prominent categories are distance based and density based environments where distance based uses K-means algorithm and density based uses DBSCAN algorithm. Distance based clustering easy to implement, fast and easily affected by noise where as density based clustering helps to find arbitrary shape clusters and also handle noise but its speed is too slow when search neighborhood data element.

This paper proposes a “BRIDGE” that efficiently merges both of these by exploiting the advantages of one counter and the limitations of the other and so on.

In our contribution, we will integrate two different-different environments and propose a new algorithmic approach name as “DRID”, which follows a common strategy in every environment. “DRID” helps to merge two clusters, which are placed close to each other. The major objectives are to increase the performance and reduce the execution time. And compare the k-means clustering algorithm and Orthogonal partitioning clustering algorithm, analysis the performance which one is best by using the tool of ORACLE 11G and apply mining, which helps to show the result in graphical manner by using Sql developer tool.

3. Experimental Consideration

Table 1 K-Means Clustering Algorithm

EMPID	ENAME	DEPTNAME	PENSION	CLUST ID
1	VIVEK	ACCOUNT	2000	18
2	AMIT	FINANCE	2000	19
3	SUKHPAL	FINANCE	3000	17
4	IQBAL	ACCOUNT	3000	14
5	SANJAY	FINANCE	5000	12

Table 2 Orthogonal Clustering algorithm

EMPID	ENAME	DEPTNAME	PENSION	CLUST ID
1	VIVEK	ACCOUNT	2000	1
2	AMIT	FINANCE	2000	1
3	SUKHPAL	FINANCE	3000	1
4	IQBAL	ACCOUNT	3000	1
5	SANJAY	FINANCE	5000	1

Here, when we perform performance analysis on two different clustering algorithms we are considering same data sets values that indicates the performance of K-Means clustering algorithm is better than orthogonal clustering algorithm because of K-means algorithm takes any kind of data sets values where orthogonal is not that much suitable this only use a sensitivity parameter which only helps to differentiate a high density peak valleys from the low density peak valleys.

As per above tables CLUS ID column helps to measure the distance between two clusters in K-Means algorithm as per taking related combinations that can be set by the users. In case of orthogonal Clustering CLUSID is to be automatically generated by the system which shows the all clusters combinations are to be placed in same cluster id. So, we can see the results and say that Performance of k-means clustering algorithm is better than orthogonal clustering algorithm.

4. CONCLUSIONS

We propose a “DRID – A new algorithmic approach” that merge two different-different environments, Distanced based environment and Grid based environment that results by merging of two clusters, which are placed close to each other increase the performance, reduce noise problem and reduce the execution time. A common strategy is to be followed in two different-different environments. Distance based algorithm(K-means algorithm) is suitable for optimization type of problem and Grid based algorithm(Orthogonal Partitioning clustering) algorithm is suitable for the set of cutting of hyper planes problems.

ACKNOWLEDGEMENT

There are a bunch of people to thank for this paper, including Dr. Sanjay Sood and Karanvir kaur. This paper would not exist but for their faith in me, and I offer them my heartfelt thanks.

REFERENCES

[1] Jiann-Cherng Shieh, Yung-Shun Lin. Bibliomining User Behaviors in the Library. Journal of Educational Media & Library Sciences.2006.

[2] Ping YU, Data mining in library Reader Management.2011 International Conference on Network Computing and Information Security.

[3] Dr. J. Akilandeswari, A survey of partitioning clustering algorithms. International Journal of Enterprise Computing and Business Systems.

[4] Manoranjan dash, Huan liu, xiaowei Xu. Merging distance and density based clustering.

[5] ZhaoHui Tang, Jamie MacLennan. Data Mining with SQL Server 2005.Wiley Publishing Inc, 2005.

[6] P. S. Bradley, U. Fayyad, and C. Reina. Scaling clustering algorithms to large databases. In Proceedings of the 4th International Conference on Knowledge Discovery & Data Mining (KDD'98).

[7] Jiawei Han and Micheline Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, 2000.

[8] Hsiao-Tieh Pu. Explore improving the utilization of library resources by bibliomining. Journal of Library Association of the Republic of China.2006.

[9] Nicholson, S. The Bibliomining Process: Data Warehousing and Data Mining for Library Decision-Making. Information Technology and Libraries. 2003.

[10] Seth Paul, Jamie MacLennan, Zhaohui Tang. Data Mining Tutorial. Microsoft Corporation.2005.