www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 3, Issue 10 October, 2014 Page No. 9032-9037

Data Mining Using PAFI

Omkar Pralhad Thakur[1], Minakshi Pawar[2]

[1]Atharva College Of Engineering,Mumbai
omkar.gamer@gmail.com

[2]Atharva College Of Engineering,Mumbai
minakshipwr4@gmail.com

Abstract: ***Data mining is the process of extracting patterns from data. Data mining is becoming an increasingly important tool to transform this data into information. It is commonly used in a wide range of profiling practices, such as marketing, surveillance, fraud detection and scientific discovery. Aim is to make such an application which has lots of inbuilt features and functions, so that many tasks can be performed on same platform, so as to save time, and also enable in curbing cost with the use of technology. At the end our goal is to come up with a good system, well-computerized and embedded with the latest technology in order to give a better service to customers, give a competitive edge to the business***.

Keywords: Association rule,Frequent item sets,Apriori Algorithm,Clusters,Dynamic Environment

1. Introduction

An Association rule plays an important role in recent data mining techniques. The purchasing of one product along with another related product represents an association rule. Association rules are used to show the relationships between data items. Association rules are frequently used in marketing, advertising and inventory control. Association rules detect common usage of items. This problem is motivated by applications known as market basket analysis to find relationships between items purchased by customers. That is, what kinds of products tend to be purchased together? The associations between data are complicated and most of them are hidden. Association rule mining is the mostly used method in Association Knowledge Discovery which aim is to find out the hided information. The most famous is the Apriori algorithm which has been brought in 1993 by Agrawal, etl. But it has two deadly bottlenecks:

- It needs great I/O load when frequently scans database.

Each element in the candidates of frequent item sets Ck needs scan database one time to decide whether it can join the Lk (where Ck and Lk are variable used in apriori algorithm). It needs scan database ten times if the frequent item sets has ten elements.

- It may produce overfull candidates of frequent item sets.

The number of the candidate of frequent itemsets Ck which were produced by frequent itemsets Lk－1 increases in the speed of exponential. For example, 1000 1-frequent itemsets may produce 500000 candidates of 2-frequent itemsets. It is so large that is a challenge to the time and main Memory [2].

To solve the drawback of the Apriori algorithm, General idea is that it reduces passes of transaction database scans and shrinks number of candidates. In proposed system, we first partition the whole database into different clusters by using PAFI algorithm (Partition Algorithm for Mining Frequent Item sets). Any item set that is potentially frequent in database must be frequent in at least one of the partitions of database.

PAFI algorithm

- The transaction having largest number of items will be put in the first cluster CL1 .The transaction having the next highest number of items will be put in the next cluster CL2.
- This process is repeated until each cluster has at most one item set.
- Next all the transactions in the database are scanned and put the transaction into the cluster that has the highest similarity measures with the existing item set. The similarity is measured based on that the number of items that are in common with the existing item set. Then the number of transactions with in each cluster is counted [2].

After forming the cluster apply an improved Apriori algorithm based on the matrix [3] on each cluster. Process description of Matrix based method as follow:-

(1) Convert the affair database matrix contained i items and t affairs to a matrix which has i+1 row and i+1 column, the fist column notes items and the first row notes affairs.

(2) Finding out the largest K-frequent item sets uses the way of operating the matrix.

(3) Making logical AND (^) between each pair of frequent 1-itemsets.

(4) Count the Support of the item sets. If its support is not less than the minimum support, the item sets is a K-frequent item sets or else it is not.

Apply above steps in all cluster and finding out frequent item set.

For finding the large item sets it is enough to go through the transactions with in the clusters. There is no need to go through the entire database again. Hence it reduces the redundant database scan and improves the efficiency.

- **RELATED WORK**

In this topic we attempt to examine enhancement of the Organizational performance through strategic management: conceptual and theoretical approach. The performance of any business organization in the competitive economy is highly dependent upon the quality of its management i.e. proper implementation of strategic management.

Proper implementation of the strategic management along with other models of strategic planning in a business organization would provide a fresh approach to re-emphasizing responsibilities to manager. The study revealed that a genuine application of strategic management by manager will enhance organization performance.

Strategic management is both the process and philosophy for determining and controlling the organizational relationship in its dynamic environment. As a process, it attempts to define approaches and techniques to assist management adapt to the dynamic of today, through the use of objectives and strategies. Strategic management endeavors to achieve effective and efficient programs to accomplish the organization's mission. As a philosophy, it changes how manager looks at competitors, customers, markets and even the organization itself. Its objective is to stimulate management's awareness of the strategic implication of environmental events and internal decision.

Lawrence and William (1988) defined strategic management as a stream of decisions and actions, which leads to the development of an effective strategy or strategies to help achieve corporate objectives. The strategic management process is the way in which strategists determine objectives and make strategic decisions.

Strategic management's main focus is the achievement of organizational goals taking into consideration the internal and external environmental factors.

Porter (1985) argues that the essence of formulating comprehensive strategy is relating a company to its environment. Strategic management permits the systematic management of change. It enables organization to purposefully mobilize resources towards a desired future.

Chandler (1962) also posited that any effective successful strategy is dependent on structure, thus to achieve any effective economic performance the organization needs to alter its structure.

The objective is to examine the conceptual and theoretical approach of strategic management on the performance of business organization.

The objectives are:

i. To examine the importance and relevance of strategic management in an organization.
ii. To examine the effectiveness of strategic management.
iii. To see how strategic management influence performance of the business organization.

- **CLASSIFICATION TECHNIQUES**

Data Mining Algorithms / Techniques

- Classification

  It is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".

- Prediction

  Attempts to find a function which models the data with the least error.

- Clustering

  It is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.

- Association Rules (Dependency modeling)

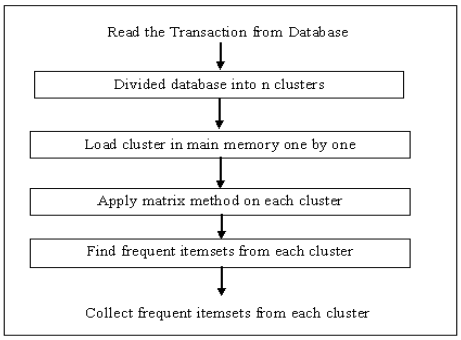
  It Searches for relationships between variables. For example a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.

- Sequential pattern mining

  It is used to find sets of data items that occur together frequently in some sequences. Sequential pattern mining, which extracts frequent subsequences from a sequence database, has attracted a great deal of interest during the recent data mining research because it is the basis of many applications, such as: web user analysis, stock trend prediction, DNA sequence analysis, finding language or linguistic patterns from natural language texts, and using the history of symptoms to predict certain kind of disease.

4. APPLICATION OF CLASSIFICATION

General idea used are that reduce passes of transaction database scans and shrink number of candidates so that it is easily fit into main memory even if database is large. Hence to reduce the number of candidate it is proposed that, we partition the whole database in to different cluster using PAFI algorithm after finding out the clusters. Then in second phase matrix method of transaction reduction applied on each cluster so that we do not need to scan database again.



Process description of proposed system

Hence time required to finding out frequent item sets required less time than apriori. Instead of whole database only clusters come into main memory so it can be easily fit into main memory.

- **TECHNIQUES**

Here we combine two algorithms called PAFI and Matrix based algorithm and which follow the following points:

Steps:

- For a set of transactions in the database D, it applies partition algorithm in order to find clusters based on the number of transactions. We are getting clusters CL1, CL2 and so on.
- ii. Partition the whole databases regardless to the size of cluster but cluster size is less than main memory size.
- iii. After finding out all the Clusters apply matrix based method on each cluster and find out the frequent item tests from each cluster.
- iv. List out all the frequent item sets of all the clusters of whole database.
- Here it did not scan whole database again hence achieved better time and space complexity.

PAFI algorithm

- The transaction having largest number of items will be put in the first cluster CL1 .The transaction having the next highest number of items will be put in the next cluster CL2.
- This process is repeated until each cluster has at most one item set.
- Next all the transactions in the database are scanned and put the transaction into the cluster that has the highest similarity measures with the existing item set. The similarity is measured based on that the number of items that are in common with the existing item set. Then the number of transactions with in each cluster is counted [2].

After forming the cluster apply an improved Apriori algorithm based on the matrix [3] on each cluster. Process description of Matrix based method as follow:-

(1) Convert the affair database matrix contained i items and t affairs to a matrix which has i+1 row and i+1 column, the fist column notes items and the first row notes affairs.

(2) Finding out the largest K-frequent item sets uses the way of operating the matrix.

(3) Making logical AND (^) between each pair of frequent 1-itemsets.

(4) Count the Support of the item sets. If its support is not less than the minimum support, the item sets is a K-frequent item sets or else it is not.

Apply above steps in all cluster and finding out frequent item set.

For finding the large item sets it is enough to go through the transactions with in the clusters. There is no need to go through the entire database again. Hence it reduces the redundant database scan and improves the efficiency.

Find frequent itemsets using Apriori algorithm:

The most famous is the Apriori algorithm [1] which has been brought in 1993 by Agrawal which uses association rule mining [5] [6].

Association rules are usually required to satisfy a user-specified minimum support and a user-specified minimum confidence at the same time. Association rule generation is usually split up into two separate steps:

- Minimum support is applied to find all frequent item-sets in a database.
- These frequent item-sets and the minimum confidence constraint are used to form rules.

Advantage of this algorithm it is easy to find frequent item sets if database is small but it has two deadly bottlenecks. First, It needs great I/O load when frequently scans database and Second, It may produce overfull candidates of frequent item-sets.

- **Find frequent itemsets using PAFI as well as Apriori algorithm:** D.Kerana Hanirex and Dr.M.A.Dorai Rangaswamy [2] proposed efficient algorithm for mining frequent item sets using clustering techniques. They presents an efficient Partition Algorithm for Mining Frequent Item sets (PAFI) using clustering. This algorithm finds the

frequent itemsets by partitioning the database transactions into clusters and after clustering it finds the frequent itemsets with the transactions in the clusters directly using improved Apriori algorithm which further reduces the number of scans in the database as well as easy to manage and available easily, hence improve the efficiency as well as new algorithm better than the Apriori in the space complexity but again it uses apriori algorithm hence efficiency not increase as much as required.

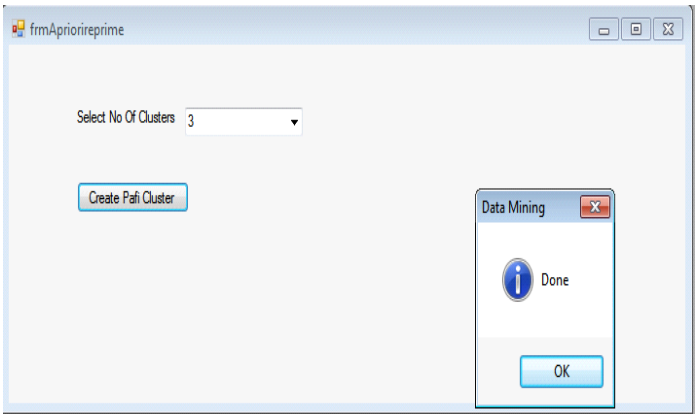


Fig1:-Result of Apriori

- **Find frequent itemsets using Improved Apriori algorithm based on matrix:** Feng WANG and Yong-hua [3] proposed An improved Apriori algorithm based on the matrix. To solve the bottleneck of the Apriori algorithm, they introduce an improved algorithm based on the matrix [8]. It uses the matrix effectively indicate the affairs in the database and uses the "AND operation" to deal with the matrix to produce the largest frequent itemsets and others. The algorithm based on matrix don't scan database frequently, which reduce the spending of I/O. So the new algorithm is better than the Apriori in the time complexity.

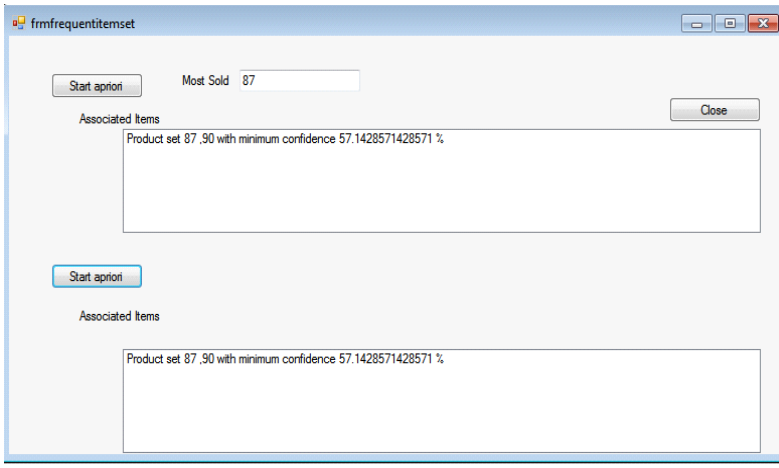


Fig2:-Pafi frequent itemset

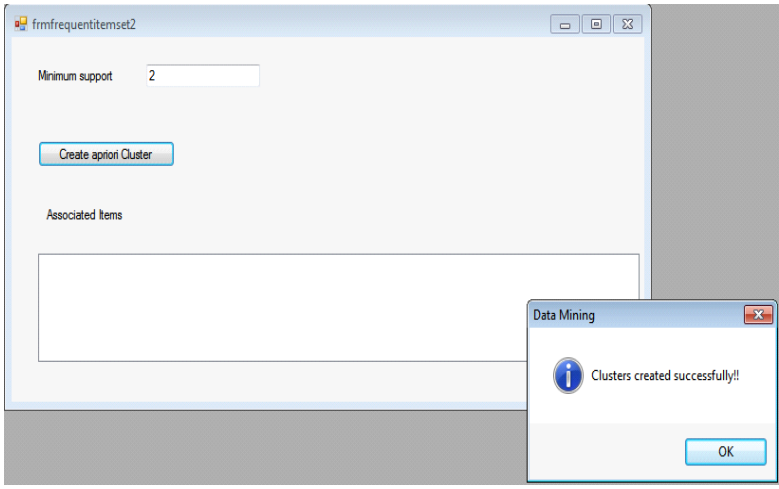


Fig :- Frequent item set 2

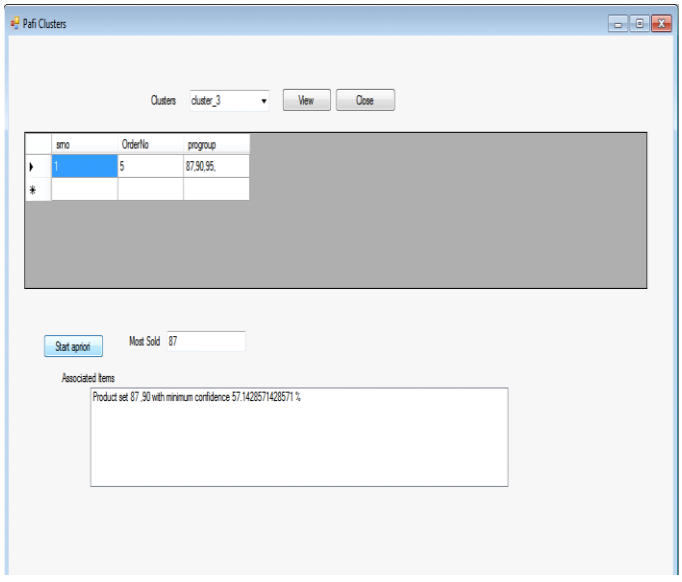


Fig :- PAFI Cluster Algorithm

- **Future Scope**

An enhancement of business organization performance through strategic management will depend on management's recognition of the following functions: setting objectives, establishing policies with which to work towards objectives, assign responsibilities and provide for coordinated action, selecting and developing key personnel, helping them adjust to change, motivating and stimulating them to think creatively and measuring progress and evaluating results.

References

[1] T. Chang, "Texture Analysis of Digitized Fingerprints for Singularity Detection," Proc. Fifth ICPR, pp. 478-480, 1980.

[2] P.E. Danielsson and Q. Z. Ye, "Rotation-Invariant Operators Applied to Enhancement of Fingerprints," Proc. Ninth ICPR, pp. 329-333, Rome, 1988.

[3] J.G. Daugman, "Uncertainty Relation for Resolution in Space, Spatial-Frequency, and Orientation Optimized by TwoDimensional Visual Cortical Filters," J. Optical Soc. Am., vol. 2, pp. 1,160-1,169, 1985.

[4] L. Hong, A.K. Jain, S. Pankanti, and R. Bolle, "Fingerprint Enhancement," Proc. First IEEE WACV, pp. 202-207, Sarasota, Fla.,1996.

[5] D.C. Huang, "Enhancement and Feature Purification of Fingerprint Images," Pattern Recognition, vol. 26, no. 11, pp. 1,661-1,671,1993.

[6] A. Jain, L. Hong, and R. Bolle, "On-Line Fingerprint Verification,"IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 19, no. 4, pp. 302-314, 1997.

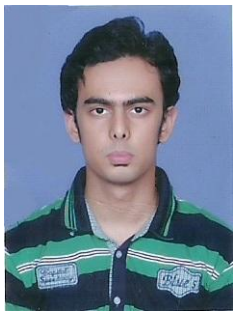
[7] A.K. Jain and F. Farrokhnia, "Unsupervised Texture Segmentation Using Gabor Filters," Pattern Recognition, vol. 24, no. 12, pp. 1,167- 1,186, 1991.

[8] T. Kamei and M. Mizoguchi, "Image Filter Design for Fingerprint Enhancement," Proc. ISCV' 95, pp. 109-114, Coral Gables, Fla., 1995.

[9] K. Karu and A.K. Jain, "Fingerprint Classification," Pattern Recognition, vol. 29no. 3, pp. 389-404, 1996.

[10] M. Kass and A. Witkin, "Analyzing Oriented Patterns," Computer Vision, Graphics, and Image Processing, vol. 37, no. 4, pp. 362-385, 1987

Author Profile



Omkar Pralhad Thakur received the B.E degree in Computer Engineering from Atharva College Of Engineering,Mumbai in 2014.

Minakshi Pawar received the B.E degree in Computer Engineering from Atharva College Of Engineering,Mumbai in 2014.