A TIME & MEMORY EFFICIENT METHOD FOR EXTRACTING SEQUENTIAL PATTERNS AFTER DATA REDUCTION FROM ORIGINAL DATA SET

[1]Vishal Arun Pawar, [2]Amit Prakashrao Patil

Assistant Professor
MCA Department
RCPET's IMRD, Shirpur

Abstract: **The data mining is a process of analysis of the data and extraction of the essential patterns from the data. These patterns are used with the different applications for making decision making and prediction related task. The decision making and prediction is performed on the basis of the learning of algorithms. Sequential pattern mining plays a vital role in organizations in decision making. This paper presents a data set reduction based technique for mining all the sequential patterns from a transaction data set. The proposed method will save computation time and memory space.**

***Keywords*: Data Mining Pattern**

1 Introduction

The use of data mining [1,7] is placed in various decisions making task, using the analysis of the different properties and similarity in the different properties can help to make decisions for the different applications. Among them the prediction is one of the most essential applications of the data mining and machine learning. This work is dedicated to investigate about the decision making task using the data mining algorithms. Therefore an application of heart disease is reported for providing the fruitful results from the algorithms.

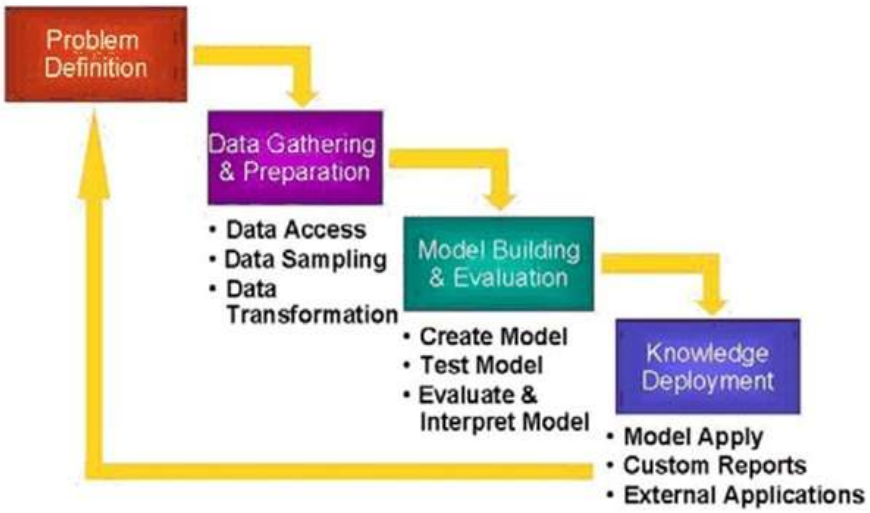

Fig. 1. Data Mining

Finally, it enables them to "drill down" into summary information to view detail transactional data. With data mining, a retailer could use point-of-sale records of customer purchases to send targeted promotions based on an individual's purchase history. By mining demographic data from comment or warranty cards, the retailer could develop products and promotions to appeal to specific customer segments.

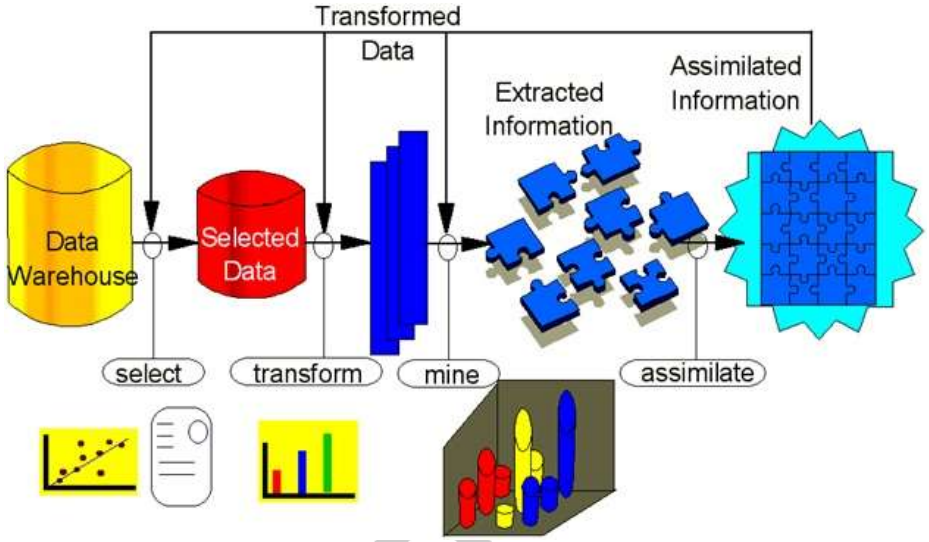

Fig. 2. key steps in data mining

The data mining is a process of analysis of the data and extraction of the essential patterns from the data. These patterns are used with the different applications for making decision making and prediction related task. The decision making and prediction is performed on the basis of the learning of algorithms. The data mining algorithms supports both kinds of learning supervised and unsupervised. In unsupervised learning only the data is used for performing the learning and in supervised technique the data and the class labels both are required to perform the accurate training. In supervised learning the accuracy [5,6] is maintained by creating the feedbacks form the class labels and enhance the classification performance by reducing the error factors from the learning model.

For example, Blockbuster Entertainment mines its video rental history database to recommend rentals to individual customers. American Express can suggest products to its cardholders based on analysis of their monthly expenditures [3].

Wal-Mart is pioneering massive data mining to transform its supplier relationships. Wal-Mart captures point-of-sale transactions from over 2,900 stores in 6

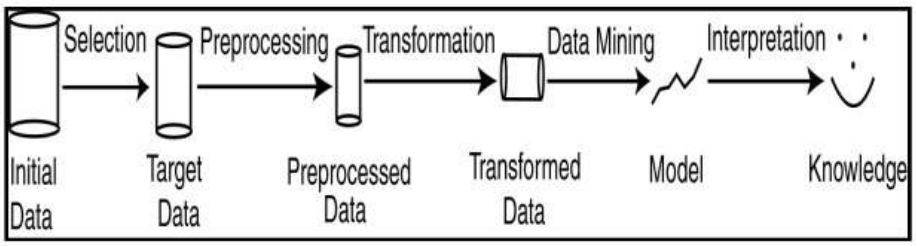

Fig. 3. KDD Process [4]

- **Selection**: Obtain data from various sources.
- **Preprocessing**: Cleanse data.
- **Transformation**: Convert to common format. Transform to new format.
- **Data Mining**: Obtain desired results.
- **Interpretation/Evaluation**: Present results to user in meaningful manner
- **Data visualization**: Generating graphs and charts for knowledge that is discovered.

In many cases it is useful to use low minimum support thresholds. But, unfortunately, the number of extracted patterns grows exponentially as we decrease. It thus happens that the collection of discovered patterns is so large to require an additional mining process that should filter the really interesting patterns. Various data bases scattered around the world are integrated in to a data ware house. It is huge data repository

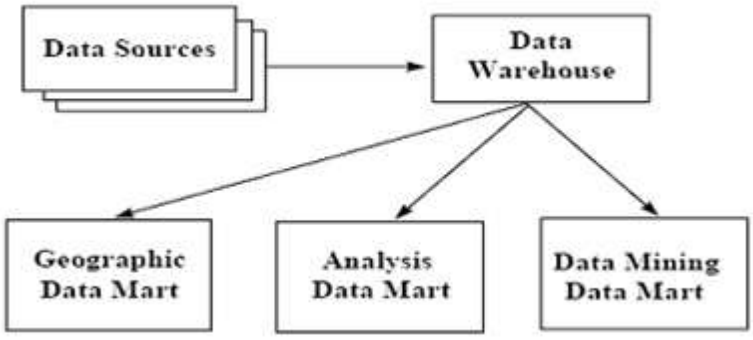

Fig. 4. Data Ware house and its Relations with Other Streams

This new database functions as a type of data mart.

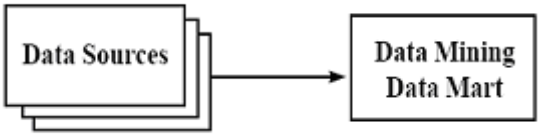

Fig. 5. Data Warehouse and Data Mart

The same holds with dense datasets, such as census data. These contain strongly correlated items and long frequent patterns. In fact, such datasets are hard to mine even with high minimum support threshold. The Apriori property [2] does not provide an effective pruning of candidates: every subset of a candidate is likely to be frequent. In conclusion, the complexity of the mining task becomes rapidly intractable by using conventional algorithms.

Closed item sets are a solution to the problems described above. These are obtained by partitioning the lattice of frequent item sets into equivalence classes according to the following property: two distinct item sets belong the same class if and only if they occur in the same set of transactions. Closed item sets are the collection of maximal item sets of these equivalence classes.

When a dataset is dense, the number of closed item sets extracted is order of magnitudes smaller than the number of frequent ones. This leverages the problem of the analyst of analyzing a large collection of patterns. Also, they reduce the complexity of the problem, since only a reduced search space has to be visited.

For example, the pattern found within the sales knowledge of a food market would indicate that if a client buys onions and potatoes along, he or she is probably going to additionally get hamburger meat. Such information are often used because the basis for decisions regarding marketing activities like, e.g., promotional evaluation or product placements. In addition to the above example from market basket analysis association rules are used these days in several application are as well as web us age mining, bioinformatics and intrusion detection. As against sequence mining, association rule learning generally doesn't take into account the order of things either inside a transaction or across transactions.

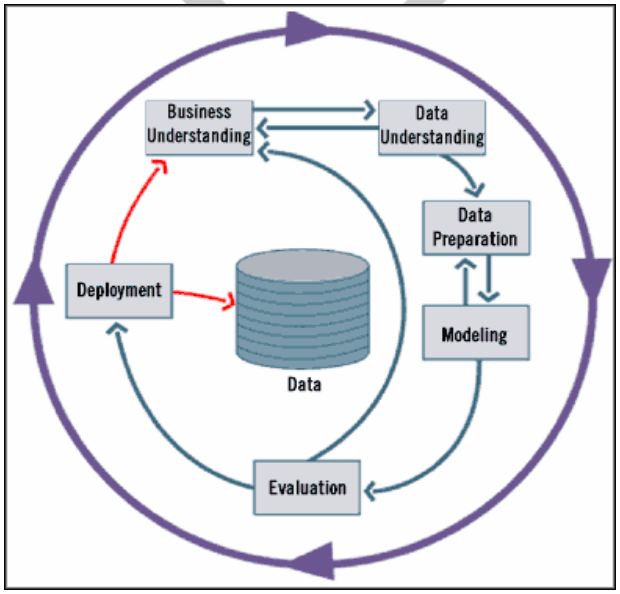

Fig. 6. Data mining stages

2 Data Mining Tasks:

Data mining software analyzes relationships and patterns in stored transaction data based on open ended user queries. Several types of analytical software are available: statistical, machine learning, and neural networks. Generally, any of four types of relationships are sought:

- **Classes**: Stored data is used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials.
- **Clusters**: Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities.
- **Associations**: Data can be mined to identify associations. The beer-diaper example is an example of associative mining.
- **Sequential patterns**: Data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes. Sequential pattern or sequential rules exploit this temporal information.

Object	Timestamp	Events
A	10	2, 3, 5
A	20	6, 1
A	23	1
B	11	4, 5, 6
B	17	2
B	21	7, 8, 1, 2
B	28	1, 6
C	14	1, 8, 7

Table 1. A Sequence Data Base

3 Related Work

Sequential rule mining has been applied in several domains such as stock market analysis [2, 3, 4, 8, 9, 10]. RPSP is also a very popular algorithm for sequential rule mining. The RPSP [6] algorithm first finds all Frequent Item sets. Here the pattern is detected by i^{th} projected databases, and after that constructs suffix database as well as prefix databases based on the famous apriori property. By reducing the minimum support, RPSP will increase the number of frequent patterns. When the founded frequent item set of prefix or suffix projected database of parent database is null then recursion will terminate. All the patterns which is generated by this algorithm that correspond to a particular i^{th} projected database of mapped or transformed database are formed into a unique set, which is not joint from all other sets. The union of disjoint subsets is the resultant set of frequent patterns. The algorithm was tested on the theoretical data and results obtained were found satisfactory. Thus, RPSP algorithm is good and it is applicable for many sequential data sets.

4 Proposed Algorithm

STEP 1: START

STEP 2: INPUT TRANSACTION DATA SET & MINIMUM SEQUENTIAL SUPPORT THRESHOLD

STEP 3: FIRST THE PROPOSED ALGORITHM SCANS THE TRANSACTION DATA BASE AND CALULATES THE SUPPORT OF EACH SINGLE SIZE ITEM.

STEP 4: IN THIS STEP A LIST OF FREQUENT SEQUENTIAL ITEM AND INFREQUENT ITEM IS PREPARED ON THE BASIS OF MINIMUM SEQUENTIAL SUPPORT THRESHOLD.
IF AN ITEM IS HAVING SUPPORT GREATER THAN THE MINIMUM SEQUENTIAL SUPPORT THRESHOLD THEN ITEM IS PLACED IN SEQUENTIAL FREQUENT ITEM LIST AND ALSO IN EXPANSION LIST AS WELL. OTHERWISE IT IS PLACED IN INFREQUENT ITEM LIST

STEP 5: REMOVE THE TRANSACTION FROM THE DATA SET D WHICH DOES NOT CONTAIN ANY FREQUENT ITEM

STEP 6: IN THIS STEP, ALL THE MEMBERS OF THE INFREQUENT ITEM LIST ARE REMOVED FROM THE TRANSACTION DATA BASE BECAUSE THEY WILL NOT APPEAR IN ANY SEQUENTIAL FREQUENT ITEM SET. IN THIS WAY, THE ORIGINAL TRANSACTION DATA BASE IS CONVERTED INTO REDUCED SIZE DATA BASE. NOW THIS REDUCED DATA BASE WILL BE USED IN THE CALCULATION OF LARGER SIZE SEQUENTIAL FREQUENT ITEM SETS.

STEP 7: WHILE EXPANSION LIST IS NOT EMPTY

- CHOOSE AN ELEMENT FROM LIST
- PERFORM LEFT EXPANSION OF SMALLER SIZE ITEMS TO GENERATE HIGHER SIZE ITEMS AND THEN REPEAT STEP 4 FOR THEM
- PERFORM RIGHT EXPANSION OF ELEMENTS AND THEN REPEAT STEP 4 FOR THEM.

STEP 8: WRITE THE LIST OF SEQUENTIAL FREQUENT ITEM SETS

STEP 9: STOP

Conclusion

The data mining is helpful for analysis the data, when the manually analysis of the data is not feasible then the data mining techniques are applied for analysis. The data mining techniques are the computer based algorithms which identify the relationship among the data and extraction of the similar pattern data on which they are trained. Basically the data mining techniques are analyze data in two different manner in first the training of the algorithm is not required which is called as the unsupervised learning techniques and the techniques in which the training before use of the algorithm is necessary is known as the supervised learning technique. This paper has given a new methodology for mining all the sequentially frequent patterns from a sequential data set. This proposed method is based on the concept of transaction elimination. It is time and memory efficient.

REFERENCES

[1] Tan, kumar "Introduction to data mining".
[2] Arun Pujari " Introduction to data mining"
[3] Das. G., Lin, K.-I., Mannila, H., Renganathan, G., and Smyth, P. Rule Discovery from Time Series. In *Proc. 4th Int. Conf. on Knowledge Discovery and Data Mining* (New York, USA, August 27-31, 1998), 16-22.
[4] Harms, S. K., Deogun, J. and Tadesse, T. 2002. Discovering Sequential Association Rules with Constraints and Time Lags in Multiple Sequences. In *Proc. 13th Int. Symp. on Methodologies for Intelligent Systems* (Lyon, France, June 27-29, 2002), pp. .373-376.
[5] Mannila, H., Toivonen and H., Verkano, A.I. Discovery of frequent episodes in event sequences. *Data Mining and Knowledge Discovery*, 1, 1 (1997), 259-289
[6] Dr P padmaja, P Naga Jyoti, m Bhargava "Recursive Prefix Suffix Pattern Detection Approach for Mining Sequential Patterns" IJCA September 2011
[7] R. Agrawal and R.Srikant. Fast algorithms for mining association rules. In ProceedingsofInternationalConferenceonVeryLargeDataBases,pages487–499,1994.
[8] R.AgrawalandR.Srikant.Miningsequentialpatterns.In*ProceedingsofInter-nationalConferenceonDataEngineering*,pages 3–14,1995.
[9] R. Agrawal and E. Wimmers. A framework for expressing and combining preferences. In Proceedings of the ACM SIGMOD International Conference on Management of Data, pages 297–306, 2000.
[10] J.Ayres, J.Gehrke, T.Yiu, and J.Flannick. Sequential pattern mining using a bitmap representation. In Proceedings of ACMSIGKDD International Conference on Knowledge Discovery and Data Mining, pages 429–435,2002