A. Dhanasekar *et al*, International Journal of Computer Science and Mobile Applications,
Vol.4 Issue. 2, February- 2016, pg. 15-24 **ISSN: 2321-8363**

Study and Analysis of Rule Mining from Heterogeneous Applications in Data Mining

A. Dhanasekar
Research Scholar, PG and Research Department of Computer Science,
Marudupandiyar College, Thanjavur, Tamilnadu, India

Dr. R.Mala
Asst.Prof & Research Advisor,
Department of Computer Science

Abstract- In this work we have identified the several issues and efficient approaches when we apply association rule mining is one of the data mining techniques for distributed data sets that are located on the distributed environments. In data mining, there are various application are highly available to extract the rules using association rule mining mechanisms. It takes much longer time to get expected results because since they have dynamic issues are generated every updated fields or applications due to the uncontrolled growth of developments in all areas. In this paper mainly focused on study and analysis process for various available applications and its merits, challenges and various available algorithms to association rule mining for the various organizations.

Key words: Data Mining, Applications, Techniques, Association Rule mining, Classification, Clustering.

I. INTRODUCTION

Data Mining is one of the wide areas of research in recent years to extract meaningful information from huge data sets from distributed environment. Data Mining is becoming popular in various fields because there is a need of efficient analytical methodology for detecting hidden and valuable information for every application. In many industries, Data Mining provides several benefits such as detection of the hidden information, availability of solution to the benefices in lower cost, detection of causes of issues and identification to solve the issue using related methods. It also helps the researchers for making efficient solutions, constructing recommendation systems to the society, developing pattern of the repositories.

Data mining consists of five major elements are [1]

- Extract, transform, and load transaction data onto the data warehouse system.
- Store and manage the data in a multidimensional database system.
- Provide data access to business analysts and information technology professionals.
- Analyze the data by application software.
- Present the data in a useful format, such as a graph or table.

II. WORKING STAGES OF KDD PROCESS AND TOOLS

A. Data Mining Process:

The KDD process [in figure1] which consists of the following steps:

1. **Selection:** It is the process of selecting data relevant for the task of analysis from the database.
2. **Pre-processing**: It Removes noise and inconsistent data and combines multiple data sources.
3. **Transformation:** It transforms data into appropriate forms to perform data mining.
4. **Data mining:** It chooses a data mining algorithm which is appropriate in extracting patterns.
5. **Interpretation/Evaluation:** It interprets the patterns into knowledge by removing redundant or irrelevant data and translating the useful patterns into terms that is understandable by human.

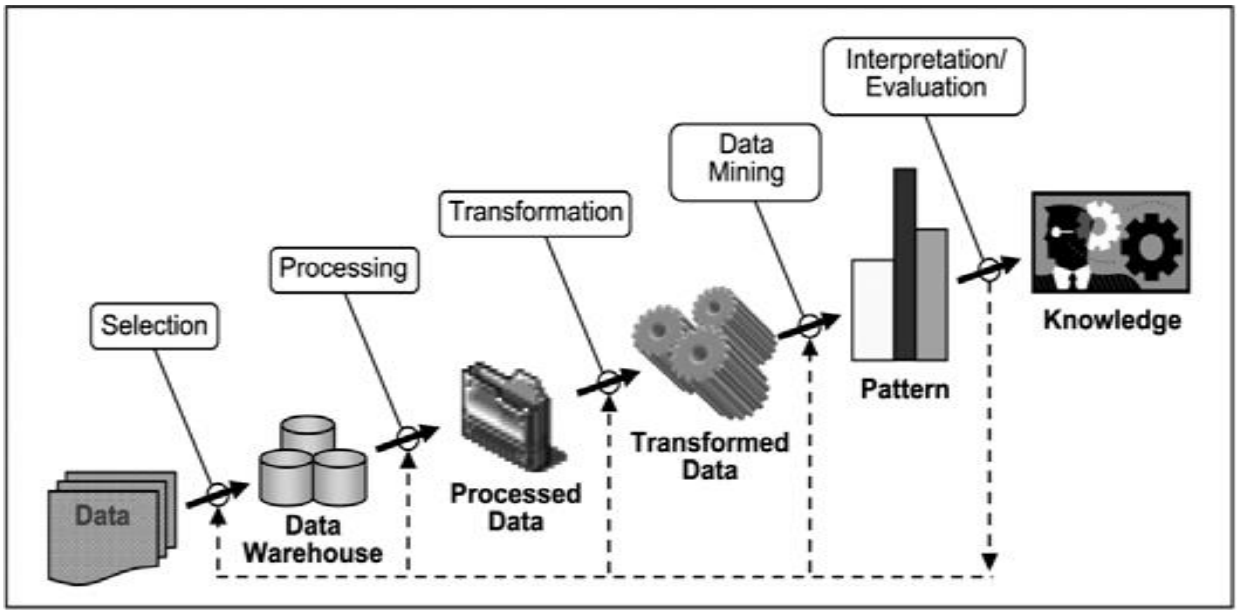

Figure 1. Stages of KDD Process

B. Data Mining Tools:

There are Different Data Mining tools available, which are as follows:

- SPSS Clementine
- SAS
- E-Miner
- MATLAB
- Oracle DM
- SQL server
- Open source software such as WEKA, R, and Orange [2].

III. DATA MINING TECHNIQUES

A. Association Rules Mining (ARM):

Association Rule algorithms can be able to generate rules with support and confidence values less than one. The number of possible Association Rules for a given dataset is commonly very large and a high proportion of the rules. Association rules provide the attributes value of conditions that may occur frequently together in a given item set. ARM computed from the data using "if-then" logic rules [3].

Association rule mining is the wide important technique in the part of data mining. ARM support the benefits to extract the frequent patterns, associations, correlations among sets of items. ARM tries to find the relationships among the attributes in the database which may be support in the task of decision making. But it is not restricted to a particular field only of association rule mining and also supports various important fields. It finds the new required rules in the transaction database and many fields, such as customer shopping analysis, additional sales, goods design, storage planning and classifying the users depending upon the buying patterns, etc. These association rules can be easily interpreted and communicated [4].

$$Rule: \ X \Rightarrow Y \quad \begin{cases} Support = \dfrac{frq(X,Y)}{N} \\ Confidence = \dfrac{frq(X,Y)}{frq(X)} \end{cases}$$

Figure 2. Two Phases of ARM

There are two phases [Figure2] in the problem of data mining association rules.

- **Support:** The support is the number of transactions that has all items in the antecedent and consequent parts of the rule. The support is sometimes act as a percentage of the total number of records in the database.

Support (XY) = Support count of (XY)/ Total number of transaction in D

- **Confidence:** Confidence is the ratio level in the number of transactions that contains all items in the consequent as well as the antecedent to the number of transactions that has all items in the antecedent [5].

Confidence (X|Y) = Support (XY)/ Support (X)

1. *Steps for generate Association Rules:*

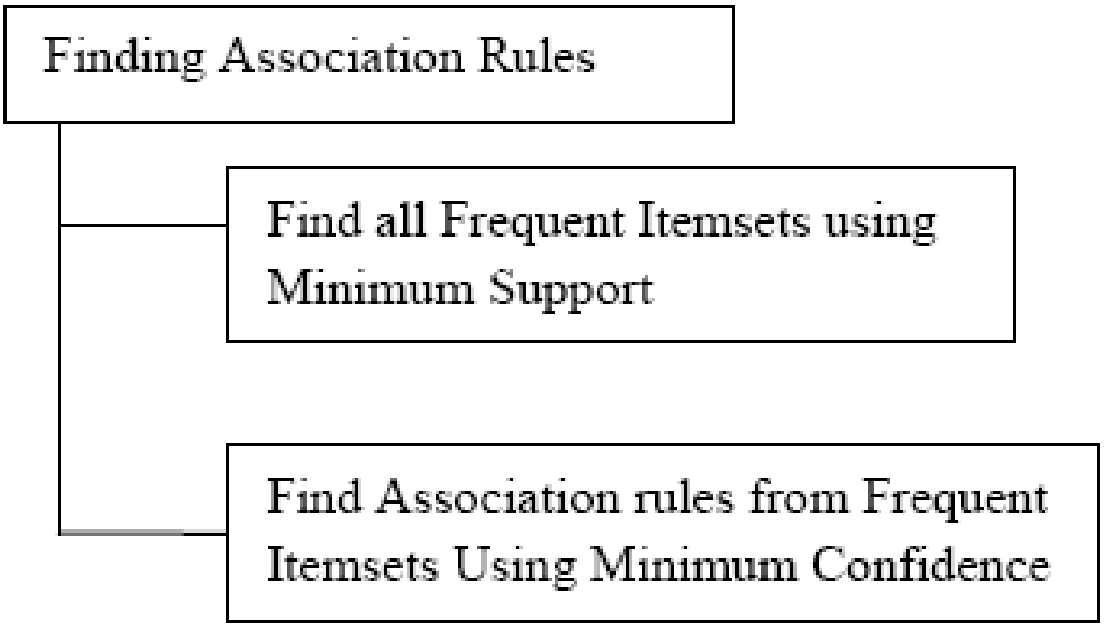

Figure 3. Generating Association Rules

2. *Distributed/parallel algorithms:*

Databases may store a large amount of data to be performed mining process. Mining association rules in such databases may require substantial processing capacity of power. A possible solution to this problem can be a distributed system. Many large databases are distributed in nature and may make it as more feasible by using distributed algorithms. Major part is a cost of mining association rules is the computation of the set of large item sets in the database. Distributed computing of large item sets has some new problems. One may compute locally large Item sets easily, but may not be globally large item set. Since it is highly expensive to broadcast the whole data set to other sites, one aspect is to show all the counts of all the item sets, no matter locally large or small, to other sites. A database may contain required combinations of item sets, and it will involve in more number of passes.

A distributed data mining algorithms (Fast Distributed Mining of association rules) are [6]

1. Distributed association rule learning
2. Collective decision tree learning
3. Collective PCA and PCA-based clustering
4. Distributed hierarchical clustering
5. Other distributed clustering algorithms
6. Collective Bayesian network learning

TABLE I. ADVANTAGES AND DISADVANTAGES OF VARIOUS CLASSIFICATION TECHNIQUES

Association Rule Mining Algorithm	**Advantages**	**Disadvantages**
AIS	✓ Estimation is used in the algorithm to prune item sets that have no hope to be large.	× Limited to only one item in the consequent. × Requires multiple passes over the database.
Apriori	✓ This algorithm has least memory consumption. ✓ Easy implementation. ✓ It uses Apriori property for pruning; item sets left for further support checking remain less.	× Requires many scans of database. × Requires only a single minimum support threshold. × Favorable for small database. × Explains the presence or absence of an item in the database.
FP-growth	✓ It is faster than other association rule mining algorithm. ✓ Repeated database scan is eliminated.	× Memory consumption is more. × Not used the interactive mining and incremental mining.

B. Classification:

Classification is the wide commonly used data mining technique, which processes a set of pre-classified samples to develop a model and it also classify the population of records at huge. It is the process of finding a function that allows the classification of data into several classes. Classification approach makes use of decision tree or neural network-based classification algorithms. The accuracy of the classification rules are estimated using test data. Applicable techniques if the required attribute is classified as decision tree, Bayesian classification, back propagation, based on concepts from ARM, k-nearest neighbor, reasoning, genetic algorithms, support vector machine and fuzzy set. If the target attribute is continuous then use of technique as linear, multiple and non-linear regression [7].

TABLE II. ADVANTAGES AND DISADVANTAGES OF VARIOUS CLASSIFICATION TECHNIQUES

Methods	**Advantages**	**Disadvantages**
K-NN	✓ Easy implementation ✓ Faster training	× Large Storage Space × Noise Sensitive × Slow Testing
Decision Tree	✓ No requirements of domain knowledge to construct decision Tree ✓ Minimizing Ambiguity ✓ Easy data process with high dimension ✓ Easy to interpret ✓ Handles numerical and categorical data	× Restricted to one output attribute × Categorical output × Unstable classifier × If dataset is numeric then Generate complex decision tree

Support Vector Machine	✓ Better Accuracy ✓ Easily handle complex data point ✓ Over fitting Problem is not much	× Computationally Expensive × Selection of Right kernel function × More time for training process × Breaking into two classes
Neural Network	✓ Identify complex relationship between dependent and independent variable ✓ Handle noisy data	× Local minima × Over fitting × Difficult to interpret
Bayesian Belief Network	✓ Easy computation process ✓ Better speed and accuracy	× Not accuracy in some variable dependent cases

C. Clustering:

Clustering is an unsupervised learning method and it is different from classification**.** Clustering is a process of partitioning a set of data into a set of required sub classes, called clusters. To based on users purpose the natural grouping or structure required in a data set. Clustering partitioned the data points based on the similarity measure .Clustering approach is used to identify same services between data points [8].

Clustering is used as identification of similar classes of objects. By using clustering techniques we can further identify object space and discover overall distribution pattern and correlations among data attributes. Classification approach can also be used for high performance means of distinguishing classes of object but it becomes highly expensive so clustering utilized as preprocessing approach for the required attribute further subset selection and classification [9].

TABLE III. ADAVATAGES AND DISADVANTAGES OF VARIOUS CLUSTERING TECHNIQUES

Methods	**Advantages**	**Disadvantages**
K-means Clustering	✓ Simple ✓ Efficient ✓ Less complex method	× Require number of cluster × Problem with handling categorical attributes × Not discover × Various result
Hierarchical Clustering	✓ Easy implementation ✓ Good visualization capacity ✓ No need to specify cluster counting	× Cubic time complexity × Decision to select split point × Not work well with noise × Not scalable
Density Based Clustering	✓ No need to specify the cluster ✓ Easily handling cluster ✓ Worked well with presence of noise	× Not handle the data points × Depend results

IV. DATA MINING APPLICATIONS OR DOMAINS

There are many applications are recently available as follows

- Market Basket Analysis
- Education System
- Medical Diagnosis (Healthcare)
- Census Data
- CRM of Credit Card Business
- Protein Sequences

A. Market Basket Analysis:

Data mining technique is used in Market Basket Analysis. Where they customer want to buying some products then this technique helps us extract the associations rule between different items that the customer put in their shopping buckets. Here the discovery of this association that develops the business technique. In this way the retailers uses the data mining technique so that they can identify the customers' intension grouping (buying the different pattern) based on regular purchases. In this way this technique is used for profits of any business and also support to purchase the branded related items.

B. Web Education:

Data mining methods are used in the web Education in the education system which is used to develop aware about courseware. The relationships are discovered among the usage data utilized up during students' sessions. This knowledge is very useful for the teacher or the author of the any kinds of course, who could decide what updating will be the highly required to improve the effectiveness of the online course. In the earlier stage the beginners are using the data mining techniques which are one of the best learning methods. This makes it possible to increase the awareness of learners. Web Education is rapidly growth in the application of data mining methods to educational communications which is both feasible and can be improvement in learning environments in the earlier system [10].

C. Medical diagnosis:

Applying ARM in medical field can be used for assisting physicians to cure patients. The general issue of the induction of related diagnostic association rules is harder because theoretically no induction process by itself to responsible to give the correctness of leading hypotheses. Practically diagnosis is not an easy process as it involves non related diagnosis tests and the presence of noise in training samples. This may result in hypotheses with unsatisfactory correctness of prediction which is highly unwanted for critical medical applications. Serban has proposed a technique based on relational database ARM and supervised learning methods to support for identify the probability of illness in a certain disease. This interface can be simply extended new symptoms additionally and its types for the given disease, and by defining new relations between given symptoms.

D. Census Data:

Censuses make a huge variety of common statistical information on society available to both researchers and the general public. The data reliable to population and economic census fields and it can be forecasted in planning public services such as education, health, transport, funds and in public business are setup new factories, shopping malls or banks and even marketing any products. The application of data mining techniques to census data and more generally to official data has honestly great potential in supporting good public policy and the effective functioning of a democratic society. Otherwise it is not necessary to demanding and requires exigent methodological study, which is still in the preliminary stages.

E. CRM of Credit Card Business:

Customer Relationship Management (CRM), through the banks hope to identify the preference of different customer classified groups based on their purchased products and services accessed to their liking to enhance the cohesion between credit card customers and the bank, has become a more interest. Shaw mainly describes how to incorporate data mining into the framework of good marketing management. The collective application of association rule techniques reinforces the informative management process and allows marketing dealing to know their customers well to provide better quality services. Song proposed a method to illustrate change of customer behavior at different time snapshots from customer profiles and marketing data. The basic aspect is to discover changes from two datasets and generate rules from each dataset to carry out rule matching [11].

F. Protein Sequences:

Proteins are essential contributions of cellular machinery of any organism. DNA technologies have provided tools for the immediate determination of DNA sequences and, by inference, the amino acid sequences of proteins from the particular of structural genes. Proteins are sequences made up of 20 types of amino acids. Each protein contains a unique 3-dimensional structure, which no independently on amino-acid sequence. A small change in sequence of protein may change the great functioning of protein. They are highly related with protein functioning on its amino acid sequence pattern has a subject of great anxiety. The research has gone into understanding the composition and nature of proteins; still many services used to be required satisfactorily. Now it is usually believed that amino acid sequences of proteins are not random. The authors are Nitin Gupta, Nitin Mangal, Kamal Tiwari, and Pabitra Mitra has deciphered the nature of associations between different amino acids and it present in a protein. Such association rules are advantageous for enhancing our knowledge data of protein composition to clarified and hold the potential to give clues regarding the global interactions amongst some particular sets of amino acids taking place in proteins. Knowledge of these association rules or constraints is highly hope for synthesis of un natural proteins [12].

V. CONCLUSION

Association Rule Mining is the one of wide range of technique in data mining and it performs with the distributed data items in data bases. Recently, there are many applications required the services of data mining techniques to produce the solution in the social effective result in all fields. ARM applied for heterogeneous application in the global to extract the newly generate rules. In this paper, we also focused on usage of various algorithms for various applications in data mining.

REFERENCES

[1] Nikita Jain, "Data Mining Techniques: A Survey Paper", International Journal of Research in Engineering and Technology (IJRET), Vol.2, Issue.11, Nov 2013, eISSN: 2319-1163, pISSN: 2321-7308.

[2] Parvathi I, "Survey on Data Mining Techniques for the Diagnosis of Diseases in Medical Domain", International Journal of Computer Science and Information Technologies (IJCSIT), Vol.5 (1), 2014, 838-846, ISSN: 0975-9646.

[3] Brijesh Kumar Baradwaj, "Mining Educational Data to Analyze Students' Performance", International Journal of Advanced Computer Science and Applications (IJACSA), Vol.2, No.6, 2011

[4] Kainaz B. Sherdiwala, "Association Rule Mining: An Overview", International Multidisciplinary Research Journal (IMRJ), Vol.2, Issue.4, Apr 2015, ISSN (O): 2349-7637.

[5] S. Venkata Krishna Kumar, "A Survey on Association Rule Mining", International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), Vol.5, Issue.9, Sep 2015, ISSN: 2277-128X.

[6] Pallavi Dubey, "Association Rule Mining on Distributed Data", International Journal of Scientific & Engineering Research (IJSER), Vol.3, Issue.1, Jan 2012, ISSN: 2229-5518.

[7] Prakash Mahindrakar, "Data Mining in Healthcare: A Survey of Techniques and Algorithms with its Limitations and Challenges", International Journal of Engineering Research and Applications (IJERA), Vol.3, Issue.6, Nov-Dec 2013, pp.937-941, ISSN: 2248-9622.

[8] Sheetal L. Patil, "Survey of Data Mining Techniques in Healthcare", International Research Journal of Innovative Engineering (IRJIE), Vol.1, Issue.9, Sep 2015, ISSN: 2395-0560.

[9] S. R. Pande, "Data Clustering Using Data Mining Techniques", International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE), Vol.1, Issue.8, Oct 2012, ISSN (O): 2278-1021, ISSN (P): 2319-5940.

[10] Neelamadhab Padhy, "The Survey of Data Mining Applications and Feature Scope", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.3, June 2012, DOI : 10.5121/ijcseit.2012.2303.

[11] Jagmeet Kaur, "Association Rule mining: A Survey", International Journal of Hybrid Information Technology, Vol.8, No.7 (2015), pp.239-242, ISSN: 1738-9968, http://dx.doi.org/10.14257/ijhit.2015.8.7.22.

[12] Jeetesh Kumar Jain, "A Survey: On Association Rule Mining", International Journal of Engineering Research and Applications (IJERA), Vol.3, Issue.1, Jan-Feb 2013, pp.2065-2069, ISSN: 2248-9622.

[13] R. Sridevi, "A General Survey on Multidimensional and Quantitative Association Rule Mining Algorithms", International Journal of Engineering Research and Applications (IJERA), Vol.3, Issue 4, Jul-Aug 2013, pp.1442-1448, ISSN: 2248-9622.

[14] Anshuman Singh Sadh, "Association Rules Optimization: A Survey", International Journal of Advanced Computer Research (IJACR), Vol.3, No.1, Issue.9, March 2013, ISSN (print):2249-7277, ISSN (online): 2277-7970.

[15] Divya Tomar, "A Survey on Data Mining approaches for Healthcare", International Journal of Bio-Science and Bio-Technology (IJ), Vol.5, No.5 (2013), pp.241-266, http://dx.doi.org/10.14257/ijbsbt.2013.5.5.25.

[16] Vijaykumar S, Dr. M. Balamurugan, Ranjani K, Big Data: Hadoop Cluster Deployment on ARM Architecture, International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE), Vol. 4, Special Issue 1, June 2015, ISSN 2278-1021 & 2319-5940.

[17] Sona Baby, "A Survey Paper of Data Mining in Medical Diagnosis", International Journal of Research in Computer and Communication Technology (IJRCCT), ISSN (O): 2278-5841, ISSN (P): 2320-5156.

[18] K. Pazhani kumar, "Association Rule Mining and Medical Application: A Detailed Survey", International Journal of Computer Applications (IJCA), Vol.80, No.17, Oct 2013, 0975-8887.

[19] T. Karthikeyan, "A Survey on Association Rule Mining, International Journal of Advanced Reasearch in Computer and Communication Engineering (IJARCCE)", Vol.3, Issue.1, Jan 2014, ISSN (O): 2278-1021, ISSN (P): 2319-5940.

[20] R. Revathi, "Mining Techniques in Health Care: A Survey of Immunization", International Journal of Computer Trends and Technology (IJCTT), Vol.10, No.2, Apr 2014, ISSN: 2231-2803.

[21] S. Sharath, "A Survey on the Principle of Mining Clinical Dataset by Utilizing Data Mining Technique", International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE), Vol.2, Issue.4, Apr 2014, ISSN (O): 2320-9801, ISSN (P): 2320-9798.

[22] S. Vijaykumar, M. Balamurugan, S.G. Saravanakumar, Unique Sense: Smart Computing Prototype, Procedia Computer Science, Volume 50, 2015, Pages 223-228, ISSN 1877-0509, http://dx.doi.org/10.1016/j.procs.2015.04.056.

[23] Monali Dey, "Study and Analysis of Data Mining Algorithms for Healthcare Decision Support System", International Journal of Computer Science and Information Technologies (IJCSIT), Vol.5 (1), 2014, 470-477, ISSN: 0975-9646.

[24] James Malone, "Data Mining using Rule Extraction from Kohonen Self-Organisaing Maps", pp.1-16.

[25] Qiankun Zhao, "Association Rule Mining: A Survey", Technical Report, CAIS, Nanyang Technological University, Singapore, No. 2003116, 2003.

[26] M. Mayilvaganan, "Cognitive Skill Analysis for Students through Problem Solving Based on Data Mining Techniques", Procedia Computer Science 47, Science Direct, Elsevier, 2015, 62-75.

[27] Vijaykumar, S., Saravanakumar, S., & Balamurugan, M. (2015). Unique Sense: Smart Computing Prototype for Industry 4.0 Revolution with IOT and Bigdata Implementation Model. *Indian Journal Of Science And Technology*, *8*(35). doi:10.17485/ijst/2015/v8i35/86698

[28] Mohammed Abdul Khaleel, "A Survey of Data Mining Techniques on Medical Data for Finding Locally Frequent Diseases", International journal of Advaced Research in Computer Science and Software Engineering (IJARCSSE), Vol.3, Issue.8, Aug 2013, ISSN: 2277-128X.

[26] Dr. M. Hemalatha, "Mining Techniques in Health care: A Survey of Immunization", Journal of the Theoretical and Applied Information Technology (JATIT), Vol.25, No.2, March 2011, ISSN: 192-8645.

