ISSN:0975-9646

Data Mining Techniques in Association Rule : A Review

Rahul B. Diwate , Amit Sahu

Department of computer Science & Engineering,
G.H.R.C.E.M. Amravati, Amravati University, India

***Abstract*- Data mining may be seen as the extraction of data and display from wanted information for specific process intended to searching information. There are different systems, tools and software's which are used to extract a relative data from a specific group of data. The greater part of data mining methods can manage distinctive information sorts. The paper gives a review and relative study on a percentage of the most widely data mining techniques being used today in normal life and commercial business.**

***Keywords*- Data mining, Knowledge management, Data mining techniques, Data mining applications.**

I. INTRODUCTION

Data mining is an interdisciplinary subfield of computer science which includes computational procedure of expansive data sets' examples disclosure. The objective of this progressed investigation process is to concentrate data from a data set and change it into a reasonable structure for further utilization. The techniques used in artificial intelligence, machine learning, statistics & database management system. Data Mining is about taking care of issues by dissecting data recently show in databases **[1]**.

Data mining tools predict future trends and behaviors, helps organizations to make proactive knowledge-driven decisions **[8].** Data mining is generally misrepresented to mean any form of large data or data processing like extraction, warehousing and analysis yet is additionally globalized to sort of supportive network. In the 1960s, statisticians had an bad practice of analysising data without a from the earlier speculation termed "Data Fishing" or "Data dredging". The expression "Data Mining" showed up around 1990 in the database group which later on came to be more prevailing in the business and press neighborhoods. Presently, Data Mining and Information Finding are utilized reciprocally. Data mining is frequently recognized to be "A mix of detail, Ai and Database research.

The aim of this study is covered:

1) Provide an overview of existing techniques that can be used for extracting of useful information from databases.

2) Provide a feature classification technique that identifies important aspects to study knowledge discovery.

3) Investigate existing knowledge discovery and data mining software tools using the proposed feature classification scheme.

II. RELATED WORK

Basic step towards the data mining techniques is KDD. Data mining popularly referred to as Knowledge Discovery in Database (KDD) **[7]**, identifies your nontrivial extraction associated with implied, in the past not known in addition to likely useful info coming from facts inside sources. While data mining in addition to Knowledge Discovery in Database (or KDD) are often cared for since word alternatives **[3]** Data mining is actually the part of Knowledge Discovery in Database, (Fig1) shows the data mining steps in knowledge discovery in database.

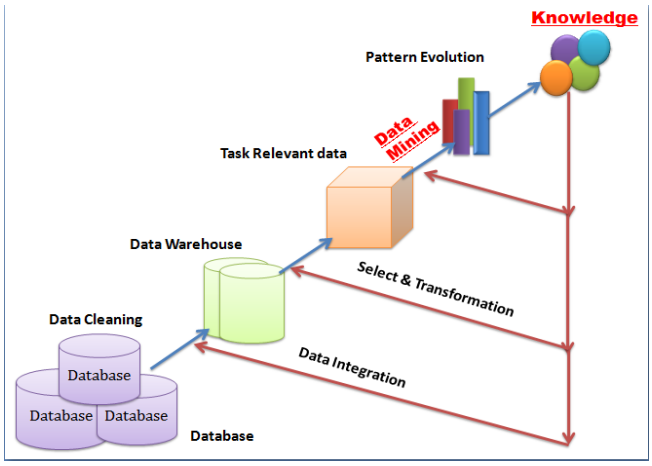

Fig 1 : Knowledge Discovery Process

The iterative process consists of the following steps for Knowledge Discovery Process Descriptions:

Developing an understanding of the application domain and the goals of the data mining process **[2]**

- selecting a target data set
- Integrating and checking the data set
- Data cleaning, preprocessing, and transformation
- Model development and hypothesis building
- Choosing suitable data mining algorithms
- Result interpretation and visualization
- Result testing and verification

The iterative process comprises of steps:

Data cleaning: also known as data cleansing, it is a phase in which noise data and irrelevant data are removed from the collection.

Data integration: at this stage, multiple data sources, often heterogeneous, may be combined in a common source.

Data selection: at this step, the data relevant to the analysis is decided on and retrieved from the data collection.

Data transformation: also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.
Data mining: it is the crucial step in which clever techniques are applied to extract patterns potentially useful.
Pattern evaluation: in this step, strictly interesting patterns representing knowledge are identified based on given measures.
Knowledge representation: is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.
It is common to combine some of these steps together. For instance, data cleaning and data integration can be performed together as a pre-processing phase to generate a data warehouse. Data selection and data transformation can also be combined where the consolidation of the data is the result of the selection, or, as for the case of data warehouses, the selection is done on transformed data.

III. DATA MINING TECHNIQUES

There are several major data mining techniques have been developed and used in data mining projects recently including association, classification, clustering, prediction and sequential patterns etc., are used for knowledge discovery from databases.

ASSOCIATION

Association is one of the best known data mining technique. In association, a pattern is discovered based on a relationship of a particular item on other items in the same transaction **[10]**. Generally association rule is applied on the large amount of data. For example, the association technique is used in market basket analysis to identify what products that customers frequently purchase together. Based on this data businesses can have corresponding marketing campaign to sell more products to make more profit.
The goal is to find all association rules with support at least *s* and confidence at least *c*, for some given minimal support size *s* and confidence level *c*, $_{that}$ hold in the unified database, while minimizing the information disclosed about the private databases held by those users. The information that we would like to protect in this context is not only individual transactions in the different databases, but also more global information such as what association rules are supported locally in each of those databases.
Applications: market basket data analysis, cross-marketing, catalog design, loss-leader analysis, etc.
Factors of Association Rule:
Generally a data is gathered from some specific technique, but which technique or which rule/method is suitable for retrieving the appropriate data, and how many time it appears. **Support** means fraction of transaction that contains item-set. **Confidence** means percentage of data in item-set. Both give the occurrence of data which is followed by group of data. Following example shows the support and confidence of data.
Example: -Given a set of transactions, find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction

Table 1:-Market Basket Data Table

TID	Items
1	Bread, Milk
2	Bread, Cloths, Beer, Eggs
3	Milk, Cloths, Beer, Coke
4	Bread, Milk, Cloths, Beer
5	Cloths, Bread, Milk, Coke

{Cloths}→{Beer}
{Milk, Bread}→{Eggs, Coke}
{Beer, Bread} → {Milk}
Where, Implication relates co-occurrence, not causality
DEFINITION OF FREQUENT ITEM SET

- **Item-set**
  - A collection of one or more items
    - Example: {Milk, Bread, Cloths}
  - k-item-set
    - An item-set that contains k items
- **Support count (σ)**
  - Frequency of occurrence of an item-set
  - E.g. σ({Milk, Bread, Cloths}) = 2
- **Support**
  - Fraction of transactions that contain an item-set
  - E.g. s({Milk, Bread, Cloths}) = 2/5
- **Frequent Item-set**
  - An item-set whose support is greater than or equal to a *minsup* threshold
  -

DEFINITION OF ASSOCIATION RULE
An implication expression of the form X → Y, where X and Y are item-sets
Example:
{Milk, cloths} → {Beer}
Rule Evaluation Metrics
Support (s)
◆ Fraction of transactions that contain both X and Y
Confidence (c)
◆ Measures how often items in Y
appear in transactions that contain X
From Table 1:
{Milk, Cloths}=>Beer
S= σ (Milk, Cloths, Beer) /|T|=2/5=0.4
C = σ (Milk, Cloths, Beer)/ σ(Milk, Cloths) =2/3=0.67

IV. CLASSIFICATION

Classification is concentrating on data mining techniques that are being used for such purposes **[5].** Being generic as it must try to integrate as many learning algorithms as possible. Meanwhile the system must be capable of generating, by means of meta-learning, a decision mechanism and so being able to decide the most adequate algorithm for each data mining task, depending on basic features of the data set, requirements of the user and the background knowledge acquired on previous data mining sessions [9].
Goal: Provide an overview of the classification problem and introduce some of the basic algorithms. Classification is a classic data mining technique based on machine learning. Basically classification is used to classify each

item in a set of data into one of predefined set of classes or groups. For Example, Teachers classify students' grades as A, B, C, D, or F. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics. In classification, we make the software that can learn how to classify the data items into groups. For example, we can apply classification in application that "given all past records of employees who left the company, predict which current employees are probably to leave in the future." In this case, we divide the employee's records into two groups that are "leave" and "stay". And then we can ask our data mining software to classify the employees into each group.

Classification Techniques

- Regression
- Distance
- Decision Trees
- Rules
- Neural Networks

Clustering

Clustering is "the process of organizing objects into groups whose members are similar in some way". A *cluster* is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters **[10]**. We can take library as an example. In a library, books have a wide range of topics available. The challenge is how to keep those books in a way that readers can take several books in a specific topic without irritate. By using clustering technique, we can keep books that have some kind of similarities in one cluster or one shelf and label it with a meaningful name. If readers want to grab books in a topic, he or she would only go to that shelf instead of looking the whole in the whole library.

Prediction

The prediction as it name implied is one of a data mining techniques that discovers relationship between independent variables and relationship between dependent and independent variables. In data mining independent variables are attributes already known and response variables are what we want to predict unfortunately, many real-world problems are not simply prediction For instance, sales volumes, stock prices, and product failure rates are all very difficult to predict because they may depend on complex interactions of multiple predictor variables. Therefore, more complex techniques (e.g., decision trees) may be necessary to forecast future values. For instance, prediction analysis technique can be used in sale to predict profit for the future if we consider sale is an independent variable, profit could be a dependent variable. Then based on the historical sale and profit data, we can draw a fitted regression curve that is used for profit prediction.

TYPES OF ASSOCIATION RULES:

Different types of association rules based on

- Types of values handled
- Boolean association rules
- Quantitative association rules
- Levels of abstraction involved
- Single-level association rules
- Multi-level association rules
- Dimensions of data involved
- Single-dimensional association rules
- Multidimensional association rules

APPLICATION OF DATA MINING

- Data Mining in Agriculture
- Surveillance / Mass surveillance
- National Security Agency
- Quantitative structure-activity relationship
- Customer analytics
- Police-enforced ANPR in the UK
- Stellar wind (code name)
- Educational Data Mining

V. CONCLUSION

There are several data mining techniques are already in use, but the use of association rule give the reducing capability of data through support and confidence method with extract a relative data from group of data. Through a use of support and confidence a communication cost and combination cost is calculated. Now a day's association Rule is widely used in commercial business.

VI. FUTURE SCOPE

In Future, It may be happed that the knowledge is not provided by the database, or any techniques used to retrieve a minimized data, The input itself taken by the technique. Monitoring the use of user or requested data by the system, it predicts and retrieves whatever the data is needed by the user. Here the kind of data wants to retrieve it predicted by techniques and executed in background.

REFERENCE:

[1] Heikki, Mannila. *1996. Data mining: machine learning, statistics, and databases*, *IEEE*

[2] Brachman, R., and Anand, T. *The process of knowledge discovery in databases: A human-centered Approach*. Fayyad, U., Piatetsky-Shapiro, G., Amith, Smyth, P., and Uthurusamy, R. (eds.), Advances in Knowledge Discovery and Data Mining, MIT Press, *Cambridge, 1996.*

[3] Neelamadhab Padhy, Dr. ragnyaban Mishra and Rasmita Panigrahi, *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

[4] Lee W. and Stolfo S, *"Data Mining Approaches for Intrusion detection", Computer Science Department Columbia University.*

[5] Harshna, NavneetKaur, *"Survey paper on Data Mining techniques of Intrusion Detection",* International Journal of Science, Engineering and Technology Research (IJSETR) Volume 2, Issue 4, April 2013.

[6] Theodoros Lappasand Konstantinos Pelechrinis, *"Data Mining Techniques for (Network) Intrusion Detection Systems"*.

[7] *Introduction to Data Mining and Knowledge Discovery*, Third Edition ISBN: 1-892095-02-5, Two Crows Corporation, 10500 Falls Road, Potomac, MD 20854 (U.S.A.), 1999.

[8] Larose, D. T., *"Discovering Knowledge in Data: An Introduction to Data Mining"*, ISBN 0-471-66657-2, ohn Wiley & Sons, Inc, 2005.

[9] Botia, J. A., Garijo, M. y Velasco, J. R., Skarmeta, A. F., *"A Generic Data mining System basic design and implementation guidelines"*, A Technical Project Report of CYCYT Project of Spanish Government.1998. WebSite: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.53.1935

[10] Kalyani M Raval,*"Data Mining Techniques"* ISSN: 2277 128X, Ijarcsse, Volume 2, Issue 10, October 2012.