International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 12 | Dec 2020 www.irjet.net p-ISSN: 2395-0072

Association Rule Mining with the Assistance of Apriori Algorithm

Harvendra Kumar [1], Rakesh Kumar [2]

[1]*Assistant Professor, Dept. of Computer Science Engineering, Surajmal Laxmi Devi Sawarthia Educational Trusts Group of Institutions, Kichha, Uttarakhand, India*
[2]*Assistant Professor, Dept. of Computer Science Engineering, Rajshree Institute of Management and Technology Bareilly, Uttar Pradesh, India*

Abstract - *Since this is an electronic era, so data are rapidly growing in this twenty-first century. Collected data is in the form of raw materials. So there is a need to analyze the data. Data analysis is a vital role, and it is not a simple task. Data analysis and finding a result from the large databases is a big problem. A lot of tools and techniques are available to finds the result and patterns. Data Mining is an emerging topic for research nowadays. When we collect the data from different sources, this may be in multiple forms. This data is stored in a commonplace that place is called the data warehouse (DW).*

There are many areas in which data used to penetrate market strategy. The retail sector is one of them [1]. In retail, customer's data are used to find the mindset /psychology of customers to increase the sales or to retain the sales and also predict new strategy to penetrate the market. Data is not beneficial to increase sales only, but it is also to compete for the market challenges. The association rule mining (ARM) is one such technique that is used to solve the problem. ARM is used to find the relationship or pattern among the itemset. This paper presents a related definition of data mining, association rule mining, and artificial neural network. An artificial neural network broadly accepts in many fields like data mining, finance, medicine, engineering, etc.

***Key Words*: Data mining, Association rule mining, Apriori algorithm, and frequent itemsets.**

1. INTRODUCTION

Data Mining is known as fetching information from an enormous group of data. We can say that data mining is the procedure to obtain knowledge from data. The data fetched so can be worn for many of the operations: Analysis of market, Fraud Detection, Customer Retention, Production Control Science Exploration. Today, for competing for the business, we need a data analysis model that analyzes the data and tells us how to that challenge accept. Data may be in different forms such as data warehouses, database system files, relational databases, transactional databases, etc. Data mining (DM) [2, 3] is a technique in which we describe the process of extracting the values from the database. The procedure of ruling unobserved information or unidentified information in a database is termed as investigative data analysis or data-driven discovery. The name of knowledge discovery and data mining in databases both are exchangeable. This collection of user content helps to analyze the user needs, and also to help other online web users to know about the content on the web. Sentiment analysis slant is a characteristic dialect handling assignment that mines data from different content structures, for example, tweets, news, and reviews, and arrange them based on their extremity as positive, negative, or impartial.

2. DATA MINING

Data mining algorithms are defined for the extraction of the information and patterns derived by the KDD process, whereas KDD is the method of discovering useful information and patterns in data. Some term which frequently used like data mining, KDD, and knowledge discovery process (KDP) is very confusing. Some researcher thinks, KDP is also known as KDD. KDD consist of a set of steps where DM is one of them. Therefore, DM is part of KDP. The first KDP model was introduced in 1990 by Fayyad et al., and after that, improvements and modifications are continuously going on. It refers to the process of analyzing data to determine the pattern and their relationships. Data mining is the core part of the KDD shown in figure 1. Each part of KDD is discussed below.

1.2 Data Cleaning

In the data gathering process, we collect a lot of unnecessary or useless data that are never used in research. Therefore a phenomenon or the procedure of removing noisy and inconsistent data from the gathered data is known as data cleaning.

1.2 Data Integration

After data cleaning, data is in a consistent form, and it may have multiple copies. Therefore various data sources are combined in a single print.

1.3 Data Selection

When data is in a single copy, then the appropriate data is retrieved from the database. The data are obtained from different sources like homogeneous or heterogeneous.

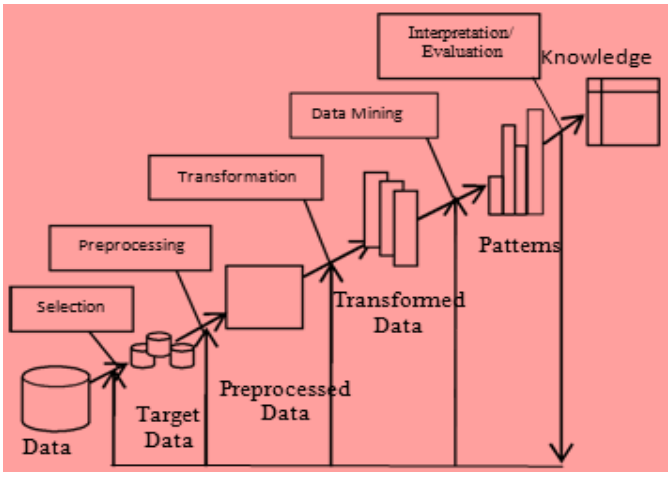

Fig - 1: The knowledge discovery process

1.4 Data Transformation

In this step, summary or aggregation operations are performed on the data for transforming or consolidating it into an appropriate form. Data transformation is nothing; it is common data architecture.

1.5 Data Mining

In this step, the intellectual method is applied for the extraction of data patterns or relationships amongst the items.

1.6 Pattern Evaluation

In this step, data patterns are evaluated.

1.7 Knowledge Presentation

At the last, after pattern evaluation, knowledge is representation.

Data mining is a component of KDD that has some methods and tasks. As shown in figure 2, data mining has two models and eight tasks. These methods are predictive data mining and descriptive data mining. The predictive data mining model makes the model of the system expressed by the given dataset or forecast the results based on the use of other historical data. A predictive data mining model is a process in which the model is described by a given historical dataset. This model is known as supervised learning. The descriptive data mining model provides information to describe what is taking place contained by the data without having a predetermined idea. This model explores or identifies a pattern or/and relationship in data. This model is as unsupervised learning. Descriptive data mining model that produces new non-trivial information based on the available dataset or it explores/ identifies patterns or relationships. DM is classified into eight tasks and two models that are either a predictive model or descriptive model, as shown in figure 2.

Classification: - In classification, data are mapped to predefined sets or classes. Classification and clustering both have groups or classes, but the classification has predefined classes or groups whereas clustering does not have any predefined classes or groups. For example, the identification of loan applicants as low, medium, and high credit risks. Therefore the classification task is categorized as supervised learning.

Regression: - Regression and classification both solve a similar type of problem, but there have some difference, the classification work on discontinuous data while regression works on continuous data. It maps a data function item to real-valued prediction variables [4]. The linear equation is an example of regression. Let y = mx +c is a linear equation which map a function f(x, y).

Time Series Analysis: - This is an ordered sequence of data points. The values are obtained that are usually uniformly distributed. Time series analysis is used where the plotline chart is needed. Examples are statistics, pattern recognition, signal processing, etc.

Prediction: - It is a method that employs data mining and probability to anticipate results. Predictive models have several predictors, i.e. they have several variables that are likely to influence future outcomes. An important example is weather forecasting, metrology, etc.

Clustering: - It is a technique for grouping of similar types of data. Cluster analysis refers to clusters that are the same (similar), but other sets (clusters) have different objects. The cluster technique and classification technique both are almost the same kinds of technique. The classification technique uses predefined classes while clustering doesn't use predefined classes. Groups or classes are obtained in clustering using similarity between records according to the characteristics in real data [5, 6]. The clustering method describes groups and objects in a dataset, while the classification technique is used in predefined classes to assign object. A book in a library is an example of clustering techniques. Outliers are those values that lie outside of any clusters and are found by the clusters technique.

Association Rule: - The objective of association mining is to discover (extract) correlations, causal relationships, frequent patterns, associations among the set of items in transaction DB, relational DB, or other data repositories. Association rules (AR) are often used in several areas like telecommunication networks, market and risk management, inventory control, cross-marketing, classification, catalog design, clustering, etc. Agrawal et al. was developed as a binary dependency between items. This problem generally occurred in business applications where some items depend on other items. The idea about ARM was commenced by a well-known researcher, R Agarwal et al. in 1993 [7]. An ARM is a technique of unsupervised learning. Association rules built on historical data show how one item affects another item. An ARM is used to determine the association or

relationship between data items. The field of an ARM is vast such as retail sectors, market basket analysis (MBA), agriculture, health sector, weather forecasting, image processing, education, banking, more. There are two measures in ARM: (i) Support (s) and (ii) Confidence (c).

Sequence Discovery: - It's a method that searches to distinguish comparable examples and ordinary events in transaction data over a business period. This task is particularly for examining sequential data to find out the sequential pattern. The best example of this task is business. In a business, a customer purchase plan A, in what circumstances he/she purchase plan B.

Summarization: - It is called generalization and it maps data items into subsets with associated simple descriptions.

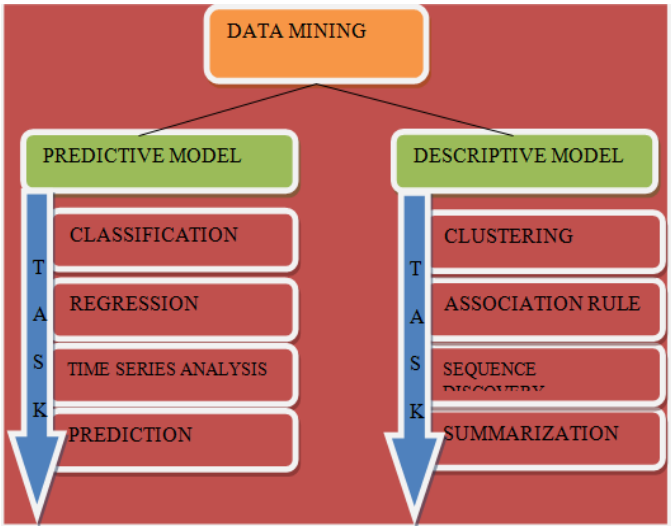

Fig - 2: Data mining model and its task

The data mining technique is used for identifying valid and meaningful patterns [8] that produce useful information and knowledge. It has become a requirement for various fields such as business, education, agriculture, health, telecommunication, etc. Data mining can have the answer to the following questions such as:

- What will the client buy? What product sells together?
- How can a company foresee which customers are at risk for roiling?
- Where has been the marketplace, and where is it going?
- How can an organization determine the success of a marketing campaign?
- What are the best techniques for analyzing unstructured data?

Data Mining Process

DM is a method of searching for different models, summaries, and derived values from a given gathering of data. DM obtains unseen predictive information and helps companies that store the most vital information (facts) in their DW (data warehouse). It allows businesses to create proactive DM tools to forecast future trends and knowledge-driven decisions. DM tools can respond to business questions that traditionally have been used extensively to fix them. DM can be promising and comparatively technology. The DM describes the method of obtaining hidden information by analyzing immense data amounts. Many companies (firms) are taking benefits of DM like manufacturing, marketing, chemicals, aerospace, etc. A DM process should be consistent and repeated by a business person. The steps of the DM process are shown below in 1.13. The data mining process consists of some steps, which are shown below in figure 3 (Database creation, exploring the database, creating a data mining model, DM model deployment).

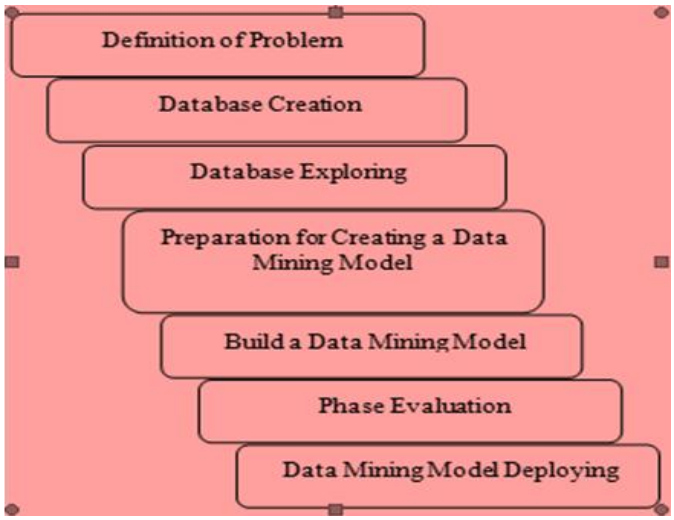

Fig- 3: Phases of the data mining process

3. ASSOCIATION RULE

The idea of association rule mining (ARM) was introduced by a well-known researcher R. Agarwal et al., in 1993 [5]. An ARM is an unsupervised learning method. The association rule made on historical data shows the relationship between how items influence the other items. The ARM is used to find the relationship or behavior between the data items. This algorithm worked in a bottom-up manner, i.e. breadth-first search iteratively. The area of an ARM is vast, such as it can be used in retail sectors, market basket analysis (MBA), agriculture, health sector, weather forecasting, image processing, education, banking, etc. In these areas, there is a big possibility to find the relationship between the items. Support, Confidence, Lift, and Conviction is the measures used in ARM.

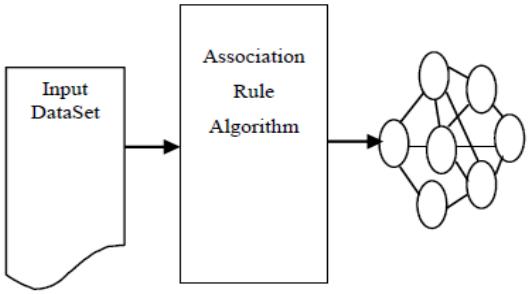

Fig - 4: Unsupervised learning process

$A \rightarrow B$ is an association rule where A (antecedent) and B (consequent) are independent items i.e. $A \cap B = \Phi$ and A & B are itemsets. $A \rightarrow B$, means customers purchase an itemset A, then probably they purchase an itemset B also. If A= {Tea, Sugar} and B= {bread, milk}. The association rule indicates that the people who bought an itemset A also probably purchase an itemset B. Support(s) and confidence (c) are two main parameters of the association rule. Domain expert defines the min_support and min_confidence threshold. An item is classified for association rule if the support of the itemset is greater than or equal to its threshold value. Support finds out how often a rule applies to a given dataset, whereas confidence finds out how frequently items (objects) in B look in transactions that hold A. Support (s) and Confidence (c) are two measures [9]. The support defined as a ratio of $A \cup B$ to the total number of transaction i.e.

Support (s):- It is defined as the percentage of records that contains $A \cup B$ to the total number of transactions (N).

$Support\ (A \rightarrow B) =$
$occurrence\ of\ A\ and\ B /$
$Total\ number\ of\ the\ transaction,$
and mathematically defined as

$$Support\,(A \rightarrow B) = P(A \cup B) = \frac{n(A \cup B)}{N}$$

Where $A \in I, B \in I,$ and $A \cap B = \Phi$.

Confidence (c):- Confidence for association rule $A \rightarrow B$ is the ratio $A \cup B$ to the total number of records that contain A. It is also called a conditional probability.

$\text{Confidence}\ (A \rightarrow B) = \text{Total occurance of A and B /}$
$\text{Total occurance of A,}$
and mathematically it is defined as

$$Confidence\ (A \rightarrow B) = P\left(\frac{A}{B}\right) = \frac{n(A \cup B)}{n(A)}$$

Where $A \in I, B \in I,$ and $A \cap B = \Phi$.

Minimum Support: - It's a user-defined value that helps to remove the non-frequent item from the DB.

Itemset: - A group (collection) of unique items. {Milk, Bread, Tea} is a group (itemsets) that has three distinct items.

K- Itemset: - An itemsets containing K items.

Frequent Itemsets: -Those itemsets are frequent itemsets that have support greater than or equal to the minimum support threshold (domain expert) and the rest itemsets are infrequent.

4. Apriori Algorithm

It's a popular algorithm to discover frequent itemsets of ARM. This system uses a bottom-up search approach. In this approach, k itemsets are used to calculate the next k+1 itemsets. In the initial level, the algorithm searches the whole database and determines frequent 1-itemsets. And the result is symbolized by L1. Support of each frequent itemsets is either greater than or equal to the minimum support. Frequent 1-itemsets are used to discover frequent 2-itemsets. And this is shown by L2. Frequent 3-itemsets (i.e. L3) are found by using the L2, and more until new frequent k-itemsets are found. For improving the efficiency of frequent itemsets of the level-wise generation, an essential property is called the Apriori property. The Apriori property is used to reduce search space [10, 11]. Since the Apriori algorithm is simple to learn. However, another algorithm has proposed that was more memory efficient and fast. This algorithm is known as FPGrowth (frequent pattern growth). Apriori algorithms become useless when efficiency is required, then more efficient algorithms FPGrowth must be used. Apriori algorithm is two steps approach, (i) Join step (ii) Prune step.

- **Join step: -**In this step, the algorithm joins two frequent Lk-1 itemsets to create the next frequent Lk itemsets.
- **Prune step: -**In the prune step, the algorithm removes unnecessary itemsets. Prune items that do not support minimum support [12, 13].

Apriori property: - A frequent itemset must be frequent for all nonempty subsets. If {ABC} is frequent itemset, then ({A}, {B}, {C}, {AB}, {AC}, and {BC}) will be frequent itemsets.

A transactional database (D) having nine transactions (T) ={T1,T2,T3,T4,T5,T6,T7,T8,T9} and five itemsets (I). Table 1.1 shows how many items are in each transaction. Suppose the min_support count is 2, then apply the Apriori algorithm on table 1.1 and find the frequent itemsets.

Table 1.1

S. No	TID	List of items
1	T1	I1, I2, I5
2	T2	I2, I4
3	T3	I2, I3
4	T4	I1, I2, I4
5	T5	I1, I3
6	T6	I2, I3
7	T7	I1, I3
8	T8	I1, I2, I3, I5
9	T9	I1, I2, I3

We found the two largest frequent itemsets that are {{*I*1,*I*2,*I*3},{*I*1,*I*2,*I*5}} by the Apriori method.

Transactional table 1.1 has two largest frequent itemsets $\{\{I1, I2, I3\}, \{I1, I2, I5\}\}$. Let say $X1 = \{I1, I2, I3\}$ *and* $X2 = \{I1, I2, I5\}$. After getting largest frequent itemsets from dataset, then we will find association rule.

Association rules for X1 are

$I1 \rightarrow I2, I3$ Confidence = 2/4 = 50%

$I2 \rightarrow I1, I3$ Confidence = 2/4 = 50%

$I3 \rightarrow I1, I2$ Confidence = 2/4 = 50%

$I1, I2 \rightarrow I3$ Confidence = 2/6 = 33%

$I1, I3 \rightarrow I2$ Confidence = 2/7 = 29%

$I2, I3 \rightarrow I1$ Confidence = 2/6 = 33%

Association rules for X2 are

$I1 \rightarrow I2, I5$ Confidence = 2/4 = 50%

$I2 \rightarrow I1, I5$ Confidence = 2/2 = 100%

$I5 \rightarrow I1, I2$ Confidence = 2/2 = 100%

$I2, I5 \rightarrow I1$ Confidence = 2/6 = 33%

$I1, I5 \rightarrow I2$ Confidence = 2/7 = 29%

$I1, I2 \rightarrow I5$ Confidence = 2/2 = 100%

If the min_confidence threshold is 60%, then X1 largest frequent itemset does not qualify any strong association rules. X2 largest frequent itemset qualify second, third, and six strong association rules.

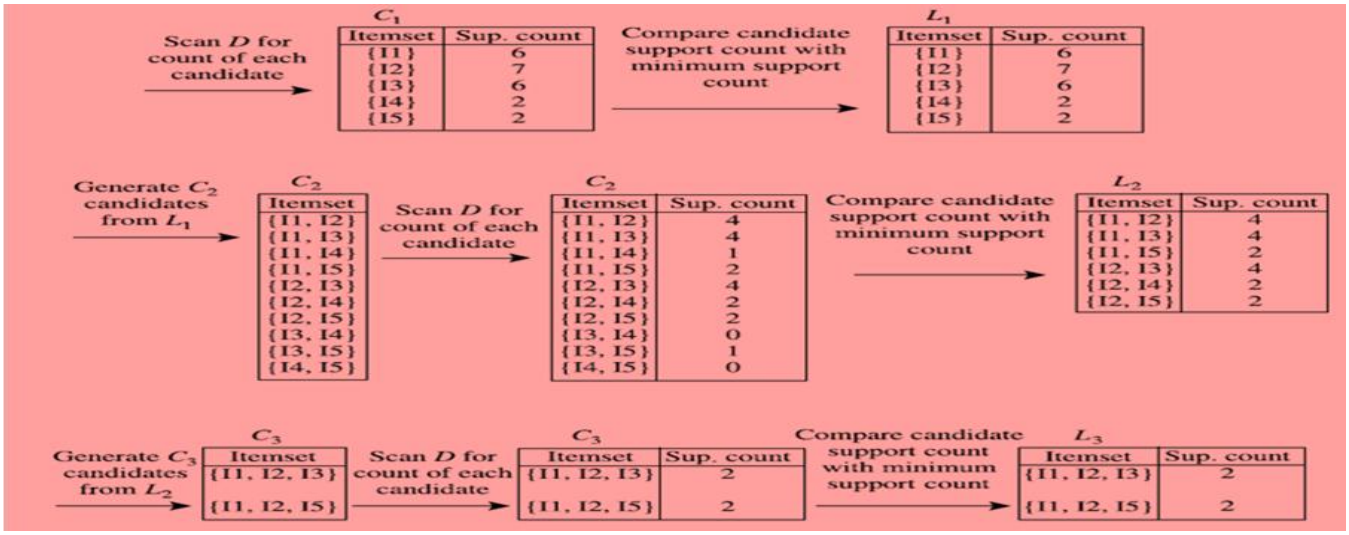

Fig - 5: Process of generating frequent itemsets

5. Conclusion

Apriori algorithm is efficient and effective for small datasets. It is not better for large databases because it scans the whole database again and again, so it takes more execution time. Therefore the complexity of this algorithm is high. If we can overcome this drawback (scanning database again and again), then performance can be increase. Other algorithms are available for data mining that solves this problem, but they have another problem. We have extracted frequent itemset and association rules in this paper.

REFERENCES

[1] Prachi Agarwal, "Benefits and issues surrounding data mining and its application in the retail industry", International journal of scientific and research publications, Volume 4, Issue 7, ISSN 2250-3153, July 2014.

[2] Pramod Prasad, Dr. Latesh Malik, "Using association rule mining for extracting product sales patterns in retail store transactions", International journal of computer science and engineering, Vol 3, ISSN 0975-3397, 2011.

[3] Bai Rama V Dr., et al., "Data mining: Techniques customer relationship management in banking and retail industries", International journal of innovative research in computer and communication engineering, Volume 2, Issue 1, Jan 2014.

[4] Wesam S. Bhaya, "Review of data preprocessing techniques in data mining", Journal of engineering and applied sciences, pp 4102-4107, Sep 2017.

[5] K. Ranjini, and Dr. N. Ranjalingam, "Performance evaluation of hierarchical Clustering Algorithms", International journal of advanced networking and applications, pp 1006-1011, May 2011.

[6] Kwame Boakye Agyapong, and et al., "Overview of data mining models (Descriptive and Predictive)", International journal of software & hardware research in engineering, Volume 4 Issue 5 May, pp 53-60, May 2016.

[7] R Agrawal, et al., "Mining association rules between sets of items in large databases", in proceeding of the ACM SIGMOD, international conference on management of data, pp 207-216, May 1993.

[8] Morana M. C. Prof, Nazarb Abdul Prof., "A roadmap to build data warehousing for retail industry", Online international interdisciplinary research journal, ISSN 2246-9698, Vol. II, Issue I, Jan –Feb 2012.

[9] Kenneth Lai and Narciso Cerpa, "Support vs. confidence in association rule algorithms", Conference of the ICHIO (The Chilean Operations Research Society), At Curico Chile in pp 01-14, Oct. 2001.

[10] Ashish Shah, "Association rule mining with modified Apriori algorithm using top down approach", 2016 2nd International conference on applied and theoretical computing and communication technology (iCATccT), Bangalore, 2016, pp. 747-752.

[11] Run_Ming Yu et al., "An efficient frequent patterns mining algorithm based on map reduce framework", International conference on software intelligence technologies and applications & international conference on frontiers of internet of things, 2014.

[12] Nidhi Makarand and dr. Snehil Dahima, "Market basket analysis using apriori algorithm in R language", Journal of trend in scientific research and development, pp 2628-2633, May 2018.

[13] R R Shelke et al., "Data mining for supper market sale analysis using association rule", Journal of trend in scientific research and development, 179-183, June 2017.