A Comparative Study on Association Rule Mining and Its Preliminaries

Aditya Shukla[1], Pankaj Kumar Gond[2], Dr. Harvendra Kumar[3]

[1]B.Tech 4th Year, Dept. of Information Technology, ITM Gorakhpur, U.P, India

[2]B.Tech 4th Year, Dept. of Information Technology, ITM Gorakhpur, U.P, India

[3]Associate Professor, Dept. of Computer Science and Engineering, ITM Gorakhpur, U.P, India

harvendra.patel81@gmail.com

Abstract.

The process of extracting and identifying nuggets of findings from outsized amounts of records is known as "data mining (DM)". It consists of several approaches, such as clustering, data summarization, association mining, and classification. Association Rule Mining (ARM), in particular, aims to extract common patterns, interesting connections, associations, or structures that can be adjusted between sets of objects or other statistics. This technique plays a significant part in the route of refining strong rules to demonstrate the stable association between several itemset present in the database. With combined rule mines, different types of techniques and measures have been designed, but it is important to know which way is the best to extract appropriate association rules. Therefore, in this document, we assess the procedures used in an ARM to test the ability to extract high-dimensional data. The paper contains the following sections: section first is the introduction, section second is literature survey, section third is preliminary concepts, section fourth is DM and DM tasks, section fifth is association rule mining, and the last section is the conclusion.

Keywords Data Mining, Frequent Itemset, Data mining Model, Association rule mining

1. INTRODUCTION

DM is the methodology of finding sensible, new patterns related to trends by filtering large amounts of archived data using Pattern Recognition (PR) techniques and mathematical techniques. According to researchers, the two important DM models are predictive and descriptive. The predictive uses a variety of available databases to predict unknown results, while the descriptive focus on finding patterns that define data. Each model for further classification has four functions, as represented in figure 1. Out of these eight functions, the ARM is the most commonly used DM function for studying trends or patterns in a dataset. It also provides rules which help in understanding customers' behavior. Nevertheless, these document sprits around a relative study of different measures of ARM. The measures are support, lift, confidence, and conviction.

This document presents a relative study of the principle and methods used in each measure. The representations provided in this document provide a further understanding of the effectiveness of the measures. Tests are performed on these scales and the results are seen in the provision of time to act and use memory in addition to the patterns produced by it.

2. LITERATURE SURVEY

K. Solanki Surbhi and T. Patel Jalpa [1], write about the trouble faced by individuals while using the frequent pattern mining strategies. This Mining strategy work by examining the database several times, which eventually results in higher process costs. Not only this, this strategy leads to the production of candidate itemsets, which ultimately requires more memory and becomes more sophisticated in handling when the database is outsized. So, to reduce the downside of this problem of candidate $s$et generation, a tree-based strategy came for mining periodic patterns. However, the tree-based strategy produces numerous conditional fp trees. So, in progression to vanquish this problem of producing numerous based fp-trees, fp is DAG for normal pattern mining is enhanced, whereas fp-tree is constructed as a DAG. In addition, to find effectual mining, fuzzy deviation could be applied to the assayable database to provide the optimal patterns.

P. Amaranatha Reddy and MHM Krishna Prasad [2], on how to obtain ARs for the differing types of data-objects and requisitions involved in them. Some of the ways of data sets reasoned in this document are Boolean, Quantitative, weighted, time-series, stream data, infrequent, Diversified, and fuzzy item-sets. So the idea of differentiating ARs amongst correlated items like milk and banana will not be considered enterprise comprehension but differentiating unknown ARs between distinct items such as liquor and diapers will be beneficent in enterprises expansion and finding such an unknown category of ARs requires in-depth information about the data.

Mrs. Geeta S. Navale and Drs. Suresh N. Mali [3] proposed a variety of methods to hide association rules on the database and to develop support and confidence measures. The intention of this paper is to conceal the critical association rules of the DM on the following conditions: no production of false rules, no loss of information, Modification Degree and robustness against intentional or unintentional attacks. Except for the three conditions above, the proposed method will be compatible with a measurable database. Therefore, the proposed method of data encryption of the operating system will be evaluated and verified with regard to the various parameters as set out in the conditions and it is necessary to assess the satisfaction of the above-mentioned conditions.

L Greeshma and Dr. G Pradeepini [4] focused on developing the latest Apriori-*b*ased algorithm, which satisfies positive aspects of constrained *i*temset based mining such as anti-*m*onotonicity. The problem of ARM is to retrieve relevant itemsets for which it presents a new constraint, called relation-based constraints, applicable to relevant data. In the CIM algorithm, it helps us to recognize the main components of a candidates' key itemsets and generate frequent itemsets, which satisfy the anti-*m*onotonicity properties, which means small coverage and cardinal size limited to a particular dataset.

Table 1: A comparative study of algorithms used in DM:

S. NO.	Algorithm Name	Application	Advantages	Disadvantages	Year
1.	AIS	Not frequently used, but when used is used for small problems.	1. Better than SETM. 2. Easy to use	1. Candidate sets generated on the fly. 2. Size of candidate set large.	1994
2.	SETM	Not frequently used.	1. Separates generation from counting.	1. Very large execution time and the size of candidate set is large.	1994
3.	Apriori	Best for closed item sets.	1. Less candidate sets. 2. Generates candidate sets from only those items that were found large.	1. Takes a lot of memory.	2003
4.	AprioriTID	Used for smaller problems.	1. Better than SETM, Apriori for small databases. 2. Fast & Time saving	--	2013
5.	Apriori Hybrid	Used where Apriori and AprioriTID used.	1. Better than both Apriori and AprioriTID.	--	2013
6.	Eclat	Best used for free item sets.	1. Less memory usage. 2. Lower minimum support.	1. Apriori wins in cases where candidate sets are more.	2004
7.	Recursive Elimination	Effectively select the most relevant itemset.	1. Better than Apriori in all cases.	1. Less than éclat in all cases.	2005
8.	FP Growth	Used in cases of large problems as it doesn't require generation of candidate sets.	1. Only 2 passes requires. 2. No candidate set generation required.	1. Using tree structure creates complexity.	2003

P. Naresh and Dr.R. Suguna [5] explain a relative study of four ARM algorithms, namely $Apriori$, $FPGrowth$, LCM and FIN. These algorithms are discovered in terms of their purpose, the way they instigate recurring itemsets, and their appearance on the organization's data. The performances of all these algorithms are evaluated by means of time. The algorithms presented in the DM configuration, namely "Sequential Pattern Mining Framework" are used to make an investigational assessment of the algorithms. These investigational results on these algorithms disclosed that the amount of data has a great significance on the implementation time and at last these led to the conclusion that: the FIN algorithm displays the least implementation time while $FPGrowth$ shows minimum memory utilization.

3. PRELIMINARIES

1) Frequent Set: T is a transactional database and σ is the "minimum support threshold" specified by the user or domain experts. An itemset is frequent if they satisfy the min support threshold,

$$s(M)_{T \geq \sigma}$$

2) Maximal Frequent Set: In order to be a "$maximal\ frequent\ set$", a frequent set must be recurring and no superordinate of it must be recurring.

3) Support(s): In a database D, s is the proportions of agreements that comprise both M and N itemsets. In s of an ARs $M \longrightarrow N$, there is,

$$supp(M \longrightarrow N) = supp(M \cup N) = P(M \cup N)$$

4) Confidence(c): In database D, this is the percentage of agreements that contain itemsets M and N. 'c' is calculated by considering the conditional probability as well as the itemset support. Confidence can be calculated using the equation,

$$conf\ (M \longrightarrow N) = P(N/M) = supp\ (M\ U\ N)\ /\ supp\ (M)$$

Here, $supp(M \cup N)$ indicates the number of agreements considering both sets of items M and N, and $supp(M)$ indicates the number of agreements considering just set M.

5) Lift: It is used to analyze the frequency M and N together, if both are precisely in different. The lift of rule $M \longrightarrow N$ is defined as,

$$lift\ (M \longrightarrow N) = \frac{confidence}{expected\ confidence} = \frac{conf(M \longrightarrow N)}{supp(N)}$$

6) Conviction: Conviction analyses the implication stability of the rule from statistical independent. Conviction is defined as,

$$conv\ (M \longrightarrow N) = \frac{1 - supp(N)}{1 - conf(M \longrightarrow N)} = \frac{P(M) * P(\overline{N})}{P(M \cup \overline{N})}$$

Where $P(\overline{N})$ is the probability that N doesn't appear in an agreement.
It compares the probability that M appears without N if they were dependent on the actual frequency of the appearance of M without N [6, 7].

4. DM AND DM TASKS

DM can often be categorized into two categories based on what a particular project is trying to accomplish. Those two categories are descriptive model and predictive model. There are a number of DM functions, such as ARM, Time Series Analysis, prediction, Neural Network, etc. Either of these functions falls under the predictive model or descriptive model. The DM system can perform one or more of the above functions as part of the DM.

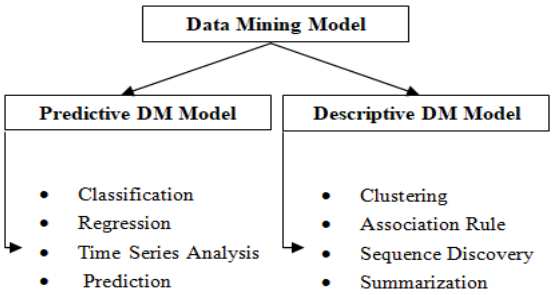

Fig 1: DM Model and its Tasks

4.1 Descriptive Model: The descriptive model defines a domain that represents a way that can be described or find the relationship among data. It can be used for many purposes. It may include ethical, structural, and other definitions that establish reasonable relationships about the system, such as its component relationships, the interactions between its components, and the distribution of its ethical properties to structural elements. Descriptive models are not usually constructed in a way that directly supports imitation, animation or performance, but can be considered compliance with grammatical rules, and logical relationships can be considered for them. This model basically relies upon an unsupervised learning approach. Some of the major tasks of descriptive models are as follows: Association Rule, Clustering, Sequence Discovery, and Summarization.

4.2 Predictive Model: The Predictive-model is a mathematical method that is generally used by DM technologies to define *m*odeling prediction as to the process of predicting the subsequent time behavior of an analysis of factual and recent data. A predictive model makes theories based on what has previously occurred and what is currently occurring. If new data shows that the current situation has changed, the likelihood of the outcome must be recalculated. This model relies upon the supervised learning approach. Some major tasks of predictive models are regression, classification, prediction, and Time Series Analysis.

5. ASSOCIATION RULE MINING

ARM [3] is an event for determining organizations, patterns, and relationships between sets of objects on dataset. The law of association is form $M \longrightarrow N$ [support, confidence]. The support and confidence are two measures that are used for assurance of the rule. AR is said to be strong if it satisfies both the minsupp(minimal support) and the minconf(minimal confidence) that is defined by the user. These ARs are facile to establish due to the *s*mall database but become more complex as the database transforms. Some general values and concepts are required in order to better understand DM in large data sets. A set contains n items and is called an n-itemset. So set {A, B} is a set of 2- itemset. Based on the frequency of itemsets, the number of active functions is calculated. So now, let us assume the value of minsupp=0.25 and minconf=0.10.
The dilemma of mining ARs can be fragmented into two sub-dilemmas:

1) Find all sets of itemsets whose support is greater thanσ. These itemsets are known as $frequent\ itemsets$.
2) Use these $frequent\ itemsets$ to develop the desired rules. The conventional ideology is that if, say p, q, r, s are $frequent\ itemsets$, then we can decide on the rule $pq \rightarrow rs$ that holds by checking the following inequality

$$\frac{\mathrm{s}(\{\mathrm{p}, \mathrm{q}, \mathrm{r}, \mathrm{s}\})}{\mathrm{n}} \geq \sigma$$

Where n is the total number of transactions and σ is a minsupp.

Measures for Association Rule Mining

Table2 presents Transactional Super Market Data

TID	Items
T_1	A_1, A_2, A_3
T_2	A_1, A_2, A_4
T_3	A_1, A_3, A_5
T_4	A_2, A_4, A_6
T_5	A_3, A_6, A_7
T_6	A_1, A_2, A_3, A_4
T_7	A_2, A_3, A_4
T_8	A_3, A_7
T_9	A_2, A_3, A_4, A_8
T_{10}	A_6, A_7

Here: Ten transactions and eight items in a transactional dataset are shown in table 1.

5.1 $\boldsymbol{Support(s)}$: $Support$ is measured as the amount of logs that contain $M \cup N$ from the entire logs in the database. The proportion for each item is altercation by one, when so ever the item is crossover in dissimilar transaction in a database across the period of scanning.
$Support$ is used to measure the quantity or frequency of an itemset in a database. This measure gives an idea of how common itemset is in all activities. Support can be measured as:

$$Supp(M \longrightarrow N) = P(MN) = Freq\frac{(M, N)}{n}$$

Where: Freq (M, N) = Transaction containing M and N, and n = Total number of transitions.
It helps us to identify the rules that need to be considered for further analysis or not. If support of the rule is greater than minsupp, then find the confidence of the rule.

$$Support\ (A_2, A_4 \longrightarrow A_3) = 3/10 = 0.3$$

Thus the support value of items A_2, A_4 & A_3 is 0.3.

5.2 **Confidence(c):** Confidence is defined as the [9] amount of the numbers of transactions that contains $M \cup N$ to the entire logs that contain M, where, if the fraction exceeds the kick-off of $confidence$, an ARs $M \longrightarrow N$ can be obtained. Confidence is an indication of how often the rule is true.
Confidence ($M \longrightarrow N$) in relation to the set of functions n, the part of the functions containing M and N.

$$conf\ (M \longrightarrow N) = \frac{supp(M\ U\ N)}{supp\ (M)}$$

Confidence explains how N is frequently occur when already buying M. This describes the link between two things. For example, if a person buys jam there is a good chance to buy bread. It is calculated as part of the number of operations where both M and N occur to support the M object.

$$Supp(A_2, A_4 \longrightarrow A_3) = 3/10 = 0.3$$
$$Supp(A_2, A_4) = 5/10 = 0.5$$
$$conf\ (A_2, A_4 \longrightarrow A_3) = 0.3/0.5 = 0.6$$

An association rule $M \longrightarrow N$ will be strong if, $conf(M \longrightarrow N) \geq minconf)$ and here $conf(A_2,\ A_4 \longrightarrow A_3)$ is 0.6 which is greater than the minconf, therefore the rule can be reasoned as a strong rule because it met with minsupp and minconf conditions. But we need to check it further than support and confidence alone cannot be sufficient to find a strong rule. In the above association rule$(A_2,\ A_4 \longrightarrow A_3)$, support of the consequent $(s(A_3) = \frac{7}{10} = 0.7)$ is greater than the confidence of the rule (0.6). This is not feasible. Therefore, this may be a misleading rule. Misleading rules can be generated from irrelevant datasets. Therefore, additional steps are needed to avoid misleading rules.

So to solve this problem of misleading rules another two measures can be used i.e. lift and conviction.

5.3 **Lift:** The Lift [10] is defined as one of the measures of ARM which define how far the inter-dependence in between M and N. The measure lift is not sensitive to rule i.e. $(lift(M \longrightarrow N) = lift(N \longrightarrow M))$. A Lift could be formulated as:

$$lift\ (\mathrm{M} \longrightarrow \mathrm{N}) = \frac{conf(M \longrightarrow N)}{supp(N)} = \frac{supp(MUN)}{\mathrm{supp}(M) * supp(N)}$$

So,

$$lift(A_2, A_4 \longrightarrow A_3) = \frac{0.3}{0.5 * 0.7} = 0.35 < 1$$

The fraction of the observance $support$ and the predicted $support$ if M and N are free for each other is known as lift. It has three possible values:

- If Lift = 1, the probability of occurrence and outcome are independent of each other.
- If Lift >1, then the itemsets are dependent on each other.
- If Lift < l, tells us that one thing replaces other things, which means one thing has a negative effect on another thing.

An ARs $M \longrightarrow N$ is engaging if it is $strong$ and $lift(M \longrightarrow N) > 1$.

Here as $lift(M \longrightarrow N) < 1$, thus the rule is not valid to consider it as a strong rule.

5.4 **Conviction:** The conviction is defined as one of the measures of ARM which undertake to analyze the magnitude of execution of the rule, to evaluate the conviction. [8, 11].

Unlike Lift, Conviction is tactful to rule direction i.e. $(conv\ (M \longrightarrow N) \neq conv\ (M \longrightarrow N))$.

A Large conviction value shows that the obtained result is largely relying on the predecessor. The conviction can be formulated as:

$$conv(M \longrightarrow Y) = \frac{1-supp(N)}{1-conf(M \longrightarrow N)} = (1 - supp\ (N))/(1 - conf(M \longrightarrow N))$$

It correlates the likeliness that M exists without N when they are relying on the factual frequentness of the existence of M without N. In that scenario, it is similar to $lift$. However, $conviction$ have monotonousness in $confidence$ and $lift$.

So,

$$conv(A_2, A_4 \longrightarrow A_3) = \frac{1 - 0.7}{1 - 0.6} = 0.75$$

$$conv(A_3 \longrightarrow A_2, A_4) = \frac{1 - 0.5}{1 - 0.42} = 0.86$$

Here, as the value of $conv(A_2, A_4 \longrightarrow A_3)$ has a value less than the value of $conv(A_3 \longrightarrow A_2,\ A_4)$, therefore the $conv(A_3 \longrightarrow A_2, A_4)$ can be considered as a strong rule.

Table 3: Quality measures and range of feasible values:

Name	Equation	Feasible values
Support	P_{MN}	[0,1]
Confidence	$\frac{P_{MN}}{P_M}$	[0,1]
Lift	$\frac{P_{MN}}{P_M * P_N}$	[0,1]
Conviction	$\frac{P_M * P_{\bar{N}}}{P_{M\bar{N}}}$	$\left[\frac{1}{n}, \frac{n}{4}\right]$

6. CONCLUSION

In this paper, there is a preliminary of the DM and a detailed discussion on ARM. ARM is facing the problem of finding the most efficient and strong rules which are suitable for any dataset due to the presence of numerous rules. In most literature, the fascinating steps of governance in ARM algorithms are based on $support$ and $confidence$. Depending on the types of application, different measures could be used to compute the interestingrules. As described in the above sections of the paper, the main issues are to trust the method for support is low forecasting capability and similar support issues. Whereas, the previous work retains presuming solutions to these issues, as an add-on the rate of promotion or sentencing and the use of an unusual support barrier, where it remains without guidance in defining support. Without such an order, users may set the wrong support limit and suffer from combinatorial explosion or loss of new cognitive arrangement. Our plan for constructing the solution for this issue is to exclude clients from determining a support limit. Therefore, we need to refine the rules in sort to obtain strong rules that must satisfy the following parameters.
Association rules will be strong if they satisfy the following conditions: (i) $supp(M \longrightarrow N) \geq minsupp,$
(ii) $conf(M \longrightarrow N) \geq minconf$ (iii) $lift(M \longrightarrow N) > 1$ and higher $conv(M \longrightarrow N)$ value.

7. REFERENCES

1. Surbhi K. Solanki and Jalpa T. Patel, “A Survey on Association Rule Mining”, Fifth International Conference on Advanced Computing & Communication Technologies, Volume 5, pp. 212-216, 2015.
2. P. Amaranatha Reddy and MHM Krishna Prasad, “Challenges to find Association Rules over various types of data items: a Survey”, International Conference on Computing, Communication and Automation, pp. 180-184, 2017.
3. Mrs. Geeta S. Navale and Drs. Suresh N. Mali, “A Survey on Sensitive Association Rules Hiding Methods”, Third International Conference on Computing, Communication, Control And Automation (ICCUBEA), IEEE, 2017.
4. L Greeshma and Dr. G Pradeepini, “Unique Constraint Frequent Item Set Mining”, 6th International Conference on Advanced Computing, IEEE, pp. 68-72, 2016.
5. P.Naresh and Dr.R.Suguna, "Association Rule Mining Algorithms on Large and Small Datasets: A Comparative Study", Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 587-592, 2019.
6. Hemant Kumar Soni, “Multi-objective Association Rule Mining using Evolutionary Algorithm”, IJARCSSE, Volume 7, Issue 5, May 2017.
7. H. K. Soni et al., “Frequent Pattern Generation Algorithms for Association Rule Mining: Strength and Challenges”, IEEE International Conference on Electrical, Electronics and Optimization Techniques (ICEEOT), pp. 3744-3747, 2016.
8. Dinesh J. Prajapati et al., “Interesting association rule mining with consistent and inconsistent rule detection from big sales data in distributed environment”, Future Computing and Informatics Journal 2, pp. 19-30, 2017.
9. J. M. Luna et al., “Optimization of quality measures in association rule mining: an empirical study”, International Journal of Computational Intelligence Systems, Volume 12, pp. 59-78, 2018.
10. H. K. Soni et al., “Association Rule Mining: A data profiling and prospective approach”, International Journal of Current Engineering and Scientific Research, Volume 3, pp. 57-60, 2016.
11. Memoona Khanum and Tahira Mahboob, “A Survey on Unsupervised Machine Learning Algorithms for automation, Classification and Maintenance”, International Journal of Computer Applications, volume 119-No.13, June 2015.
12. Sikha Bagui and Probal Chandra Dhar, "Positive and negative association rule mining in Hadoop’s MapReduce environment", Journal of Big Data, pp.1-16, 2019.
13. Lichun Li et al., “Privacy-Preserving-Outsourced Association Rule Mining on Vertically Partitioned Databases”, IEEE Transactions on Information Forensics AND Security, pp. 1-15, 2016.
14. Mandeep Mittala et al., “Loss profit estimation using association rule mining with clustering”, Management Science Letters, pp. 167–174, 2015.