International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT2390210

Data Mining Tools for Generate Item Set : Critical Review

S. V. Subramanyam
Professor, Department of Artificial Intelligence and Machine Learning, School of Engineering, Mallareddy University, Hyderabad, Telangana, India

ARTICLEINFO

Article History:

Accepted: 13 March 2023
Published: 18 March 2023

Publication Issue
Volume 10, Issue 2
March-April-2023

Page Number
176-183

ABSTRACT

Most algorithms used to identify large itemsets can be classified as either sequential or parallel. In most cases, it is assumed that the itemsets are identified and stored in lexicographic order (based on item name). This ordering provides a logical manner in which itemsets can be generated and counted. This is the normal approach with sequential algorithms. On the other hand, parallel algorithms focus on how to parallelize the task of finding large itemsets. Mining Associations is one of the techniques involved in the process mentioned in chapter 1 and among the data mining problems it might be the most studied ones. Discovering association rules is at the heart of data mining. Mining for association rules between items in large database of sales transactions has been recognized as an important area of database research. These rules can be effectively used to uncover unknown relationships, producing results that can provide a basis for forecasting and decision making. Today, research work on association rules is motivated by an extensive range of application areas, such as banking, manufacturing, health care, and telecommunications. It is also used for building statistical thesaurus from the text databases, finding web access patterns from web log files, and also discovering associated images from huge sized image databases.

Keywords : KDD, WWW, CAR, CHAID, AIS

I. INTRODUCTION

The role of data mining is simple and has been described as “extracting knowledge from large amounts of data” where data is stored in data warehouse , OLAP(On Line Analytical Process) , databases and other repositories of information[7]. Data mining has emerged as an important method to discover useful information , hidden patterns or rules from different types of datasets. Association rule mining is one of the dominating data mining technologies. Association rule mining is a process for finding associations or relations between data items or attributes in large datasets. Association rule is one of the most popular techniques and an important research issue in the area of data mining and knowledge discovery for many different purposes such as data analysis , decision support , patterns or

correlations discovery on different types of datasets. Association rule mining has been proven to be a successful technique for extracting useful information from large datasets. Various algorithms or models were developed many of which have been applied in various application domains that include telecommunication networks, market analysis, risk management, inventory control and many others [9].
Data mining is one of the step of KDD process. KDD process has several steps , which are performed to extract knowledge from data in the context of large databases such as data cleaning , data integration , data selection , data transformation , data mining , pattern evaluation , knowledge presentation [7].

1) Data Cleaning

The consistency of data is improved in this phase by eliminating the noise or unrelated data. It comprises of data clearing, for instance, handling missing values and elimination of noise or outliers. It will possibly involve using complex statistical techniques or a data mining algorithm.

2) Data Integration

The integration is one of the most significant features of data warehouse. Here, multiple data sources may be integrated. Data is given from multiple dissimilar sources into the data warehouse. As the data is fed, it is transformed, reformatted, summarized, and so forth. The result of integration is that once the data exists in the data warehouse, it has a particular physical corporate image. In all the integration architecture, there are several difficulties that come up when attempting to integrate data from different sources.

3) Data Selection

Once the data elements are chosen from several sources, it is essential to examine the value of the data. Data samples are accumulated from the sources and data profiling is carried out to recognize the issues of physical data quality. The data which are selected for an object are dependent on the patterns of significance. The data acquired from the sources will be required for three major purposes during the data mining process i.e. training the data mining model, testing it and applying it on the entire data.

4) Data Transformation

In this phase, the generation of enhanced data for data mining is organized and developed. Techniques include dimension reduction and attribute transformation. This phase is essential for the achievement of the entire KDD process, and it is typically very specific. On the other hand, even if the right transformation is not carried out initially, there is a chance to obtain an unexpected effect that makes it necessary to do the transformation in the next iteration. Consequently, the KDD process reflects upon itself and leads to an understanding of the transformation required.

5) Data mining

It is an indispensable process where intelligent techniques are applied with the intention of extracting interesting data patterns. There are two most important goals in data mining, namely, prediction and description. Prediction is referred to as supervised data mining, whereas descriptive data mining is referred to as unsupervised and visualization features of data mining. Several data mining approaches are based on inductive learning, where a model is built explicitly or implicitly by simplifying from an adequate number of training examples. The fundamental assumption of the inductive approach is that the trained model is relevant to future cases. The approach also considers the level of meta-learning for the particular set of available data.

6) Pattern Evaluation

In order to recognize the interesting patterns representing knowledge depending on some interestingness measures are evaluated. The evaluation and interpretation of the mined patterns

(rules, reliability) concerning the goals defined in the initial phase are carried out. This phase concentrates on the comprehensibility and effectiveness of the induced model. In this phase, the discovered knowledge is also documented for further purpose. The last phase is the usage and overall feedback on the patterns and the discovery results obtained by the data mining.

I. RELATED WORK

Jaishree Singh, Dr. J.S. Sodhi[1], has explained Classical Apriori algorithm generates large number of candidate sets if database is large. And due to large number of records in database results in much more I/O cost. In this project, we proposed an optimized method for Apriori algorithm which reduces the size of database. In our proposed method, we introduced an attribute Size_Of_Transaction (SOT), containing number of items in individual transaction in database.

Chang-Hung Lee, Ming-Syan Chen[2], the discovery of association relationships among a huge database has been known to be useful in selective marketing, decision analysis, and business management. A popular area of applications is the market basket analysis, which studies the buying behaviors of customers by searching for sets of items that are frequently purchased together (or in sequence).

Farah Hanna AL-Zawaidah, Yosef Hasan Jbara[3], in this paper we present a novel association rule mining approach that can efficiently discover the association rules in large databases. The proposed approach is derived from the conventional Apriori approach with features added to improve data mining performance. We have performed extensive experiments and compared the performance of our algorithm with existing algorithms found in the literature. Experimental results show that our approach outperforms other approaches and show that our approach can quickly discover frequent itemsets and effectively mine potential association rules

. **Du Ping, Gao Yongping**[5], In this paper, they have explained an enhance algorithm associating which is based on the user interest and the importance of itemsets is put forward by the paper, incorporate item that user is interested in into the itemsets as a seed item, then scan the database, incorporate all other items which are in the same transaction into item sets, Construct user interest itemsets, reduce unnecessary itemsets; through the design of the support functions algorithm not only considered the frequency of itemsets, but also consider different importance between different itemsets.

The new algorithm reduces the storage space, improves the efficiency and accuracy of the algorithm.

Zhuang Chen, Shibang Cai, Qiulin Song and Chonglai Zhu[23], In this paper, they have analyzed the basic ideas and the shortcomings of Apriori algorithm, studies the current major improvement strategies of it. In order to solve the low performance and efficiency of the algorithm caused by its generating lots of candidate sets and scanning the transaction database repeatedly, it studies the pruning optimization and transaction reduction strategies, and on this basis, the improved Apriori algorithm based on pruning optimization and transaction reduction is put forward. According to the performance comparison in the simulation experiment, by using the improved algorithm, the number of frequent item sets is much less and the running time is significantly shortened as well as the performance is enhanced then finally the algorithm is improved.

II. RESEARCH METHODOLOGY

Techniques of Data Mining

The process of mining is often controlled by the requirements of the users. The user may be a business analyst or may be a marketing manager. Different users have different need of information. Depending on the requirements we can use different data mining techniques. The different types of data mining functionalities and the patterns they discover are described below:

Association

Association is one of the best recognized data mining approaches. In association, a pattern is discovered depending on an association of a particular item on other items in the same transaction. For instance, the association technique is implemented in market basket analysis to recognize the products that the customers often purchase together. Depending on this data businesses can have equivalent marketing campaign to advertise more products to create an increase in income. Association and correlation is typically to discover frequent item set findings among large data sets. This type of finding assists businesses to make certain verdicts, like catalogue design, cross marketing and customer shopping behavior analysis. Association rule algorithms are required to be capable of generating rules with confidence values of less than one. On the other hand, the number of possible association rules for a given dataset is normally very large and a high proportion of the rules are generally of little value. The major types of association rule are multi-level association rule, quantitative association rule, redundant association rule and negative association rules.

As a data mining function, clustering can be used as a standalone tool to gain insight into the distribution of data, to observe the characteristics of each cluster, and to focus ona particular set of clusters for further analysis.

Requirements of clustering in data mining:

- **Scalability** – We need highly scalable clustering algorithms to deal with large databases.
- **Ability to deal with different kinds of attributes** – Algorithms should be capable to be applied on any kind of data such as interval-based (numerical) data, categorical, and binary data.
- **Discovery of clusters with attribute shape** – The clustering algorithm should be capable of detecting clusters of arbitrary shape. They should not be bounded to only distance measures that tend to find spherical cluster of small sizes.
- **High dimensionality** – The clustering algorithm should not only be able to handle low-dimensional data but also the high dimensional space.
- **Ability to deal with noisy data** – Databases contain noisy, missing or erroneous data. Some algorithms are sensitive to such data and may lead to poor quality clusters.
- **Interpretability** – The clustering results should be interpretable, comprehensible, and usable.

Example for understanding Association Rules

The Table 2.1 depicts an example Transaction database andTable 2.2 shows that {1, 2, 3} and {1, 2, 5} are frequent 3-itemsets.

The non empty subsets of {1,2,3} are{1},{2},{3},{1,2}, {1,3} and {2,3}.

The association rules generated are:

{1, 2} -> {3}	confidence=2/4 = 50%
{1, 3}->{2}	confidence=2/2 = 100%
{2, 3} ->{1}	confidence=2/3 = 66%
{1} -> {2, 3}	confidence=2/4 = 50%
{2} -> {1, 3}	confidence=2/6 = 33%
{3} -> {1, 2}	confidence=2/3 = 66%

If minimum confidence is equal to 66% then the following rules are strong rules: {1, 3} ->{2}, {2, 3}->{1}, {3}-> {1, 2}

APRIORI ALGORITHM

Apriori Algorithm is a classic technique used in association analysis to discover or mine association rules. Given the minimum support and minimum confidence threshold, it attempts to find all the association rules whose rule support values are greater than or equal to the minimum rule support threshold and confidence values that are greater than or equal to the minimum confidence threshold.

A computer program analyzes a large set of purchase records (transactions) to find the sets of item that are frequently purchased together. Such sets are called frequent item sets. The definition of "frequent" is based on a user-provided frequency and is called the necessary support. Once the frequent item sets are known, a separate process can use these associations to help suggest additional items that a current customer might purchase.

Aproiri algorithm for frequent itemset generation

1. $k = 1$
2. $F_k = \{i | i \in I \Lambda \sigma(\{i\}) \geq N \times minsup\}$ $\{Findallfrequent1 - itemset\}$
3. $repeat$
4. $k = k + 1$
5. $C_k = apriori - gen(F_{k-1})$ $\{Generatecandidateitemset\}$
6. $foreachtransactiont \in Tdo$
7. $C_t = subset(C_k, t)$ $\{Identifyallcandidatesthatbelongtot\}$
8. $foreachcandidateitemset$ $c \in C_tdo$
9. $\sigma(c) = \sigma(c) + 1$ $\{Incrementsupportcount\}$
10. $endfor$
11. $endfor$
12. $F_k = \{c | c \in C_k \wedge \sigma(c) \geq N \times minsup\}$ $\{Extractthefrequentk - itemset\}$
13. $untilF_k = \emptyset$
14. $Result = U_k F_k$

$Procedureapriori\_gen\ (F_{k-1})$

1. $foreachitemsetsf_1 \in F_{k-1}do$
2. $foreachitemsetsf_2 \in F_{k-1}do$
3. $if(f_1[1] = f_2[1]) \wedge (f_1[2] = f_2[2]) \ldots \wedge (f_1[k-2] = f_2[k-2]) \wedge (f_1[k-1] = f_2[k-1])then$
4. $c = f_1 \otimes f_2$ $\{joinstep: generatecandidates\}$
5. *for each (k-1)-subsets s of c do*
6. $if(s \notin f_{k-1})$ $then$
7. $deletec\ \{prunestep: removecandidate\}$
8. $else$
9. $addctoC_k$
10. $endfor$
11. $returnC_k$
12. $endif$
13. $endfor$
14. $endfor$

Procedure subset (C_k, t)

1. $forallcandidatess \in C_kdo$
2. $iftcontainss$
3. $subset = subset + \{s\}$
4. *end for*

DATA MINING TOOLS

Data mining , a set of techniques used for the purpose of obtaining information from the data. Statistical analysis of data using a combination of techniques and artificial intelligence algorithms and data quality information in the disclosure of confidential information, a process of transformation. In this context, SPSS Clementine, Excel, SPSS, SAS, Angoss, KXEN, SQL Server, MATLAB, commercial and **RapidMiner** (YALE), **Weka**, **R**, **C4.5**, **Orange**, **KNIME** developed several programs, including open source.

i) Open Source Programs Data Mining

Data mining applications is necessary to use a computer program to do. In this context, most software is developed. In this section, the Open Source Data Mining Programs and **RapidMiner** (YALE), **Weka** and **R** programs mentioned.

i) RapidMiner (YALE)

By scientists from Yale University in the United States was developed using Java language. RapidMiner (previously: Rapid-I, YALE) is a mature, Java-based, general DM tool currently in development by the company RapidMiner, Germany. Previous versions (v. 5 or lower) were open source.

RapidMiner also offers the option of application wizards that construct the process automatically based on the required project goals (e.g. direct marketing, churn analysis, sentiment analysis).

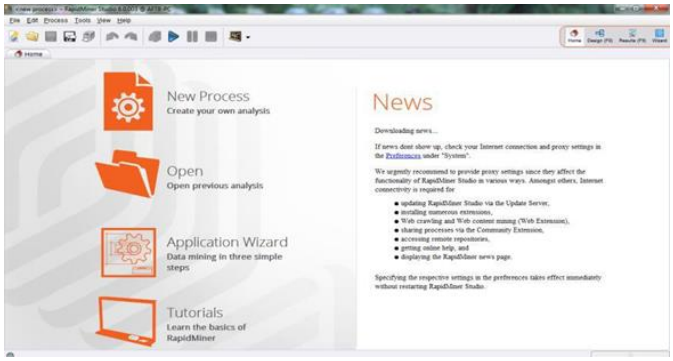

Figure 1 : RapidMiner Application Menu

WEKA

Weka is a Java-based, open-source DM platform developed at the University of Waikato, New Zealand. The software is free under GNU GPL 3 for noncommercial purposes. Weka has had mostly stable popularity over the years, which is mainly due to its user friendliness and the availability of a large number of implemented DM algorithms. It is still not as popular as RapidMiner or R, both in business and academic circles, mostly because of some slow and more resource demanding implementations of DM algorithms. Although it is not a single tool of choice in DM, Weka is still quite powerful and versatile, and has a large community support.

WEKA Application Menu

III. IMPLEMENTATION AND PERFORMANCE EVALUATION

While implementing Apriori algorithm such as spend a large overhead to deal with large candidate set, many repeat comparison of itemset in join step and repeatedly scanning the transaction database requires a lot of I/O load etc. Hence to analyze the algorithm in depth, we have used WEKA tool, which is build for various kind of data mining algorithms and in respected research area.

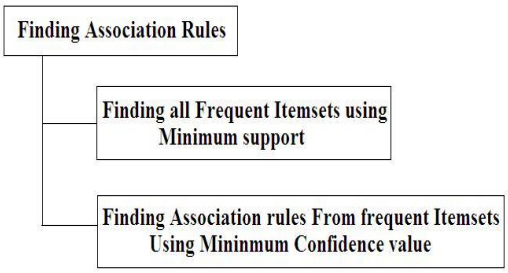

Figure 4: Finding the association Rule

i) Implementation of Apriori Algorithm: To perform the Apriori algorithm, the best open source data mining tool is Weka, which is developed at the University of Waikato, New Zealand, first we retrieve the dataset that is already exist in weka tool, by which we could perform the algorithms and analyze the objectives.

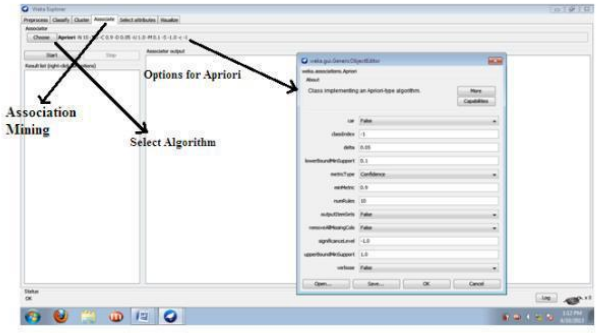

Figure 5: Apriori Algorithm with differ properties in Weka

ii **Implementation of Filter Association**

To perform the Filter Associator, we have to do same procedure as Apriori algorithm i.e. just select the Filter Associator in place of Apriori algorithm. After taking the value of support and confidence, the execution of Filter Associator is done by clicking the "Start" button and according to that it generates the best association rules.

iii) Performance Evaluation of Apriori algorithm and Filter Associator

After performing the execution of both algorithms: Apriori algorithm and Filter Associator in the weka tool, we found that Apriori algorithm takes more number of cycle performed and for specific value of support it also generates extra large itemsets compare to the Filter Associator.

IV. CONCLUSION

In this paper we have discussed various association rule algorithms and compared two algorithms: Apriori algorithm and Filter Associator. We have analyzed the frequent itemsets generation and number of cycle performed over the Apriori algorithm and Filter Associator in the context of association analysis. According to the comparison of above two algorithms on weka tool, we conclude that Filter Associator is efficient algorithm than Apriori algorithm based on above two factors (Number of cycle performed, large itemsets) because the Apriori algorithm generates more number of cycle performed and generate extra large itemsets which degrades the performance of algorithm.

Future Recommendations

Some of the future enhancements of the thesis are presented below:

- The work presented in the thesis can be extended for multi-level association rule mining.
- The work can be enhanced to generate multi-dimensional association rules.
- A tool for generating association rules can be developed. This tool can choose the approach for frequent itemsets mining according to the properties of the dataset to be mined.

V. REFERENCES

[1]. Hassan M. Najadat, Mohammed Al-Maolegi, Bassam Arkok, "An Improved Apriori Algorithm for Association Rules", International Research Journal of Computer Science and Application Vol. 1, No. 1, June 2013, PP: 01 – 08.

[2]. H.Toivonen, "Sampling large databases for association rules". In Proc. 2006 Int. Conf. Very Large Data Bases(VLDB'06),pages 134-145, Bombay, India, Sep.2006.

[3]. Hu Ji-ming, Xian Xue-feng. " The Research and Improvement of Apriori for association rules mining",

[4]. Computer Technology and Development 2006 16(4) pp. 99-104.

[5]. Jiao Yabing, "Research of an Improved Apriori Algorithm in Data Mining Association Rules", International

[6]. Journal of Computer and Communication Engineering, Vol. 2, No. 1, January 2013.

[7]. J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2000.

[8]. Jaishree Singh, Hari Ram, Dr. J.S. Sodhi, "Improving Efficiency of Apriori Algorithm Using Transaction Reduction ", International Journal of Scientific and Research Publications, Volume 3, Issue 1, January 2013.

[9]. Li Qingzhong , Wang Haiyang, Yan Zhongmin, "Efficient mining of association rules by reducing the number of passes over the database", Computer Science and Technology, 2008, pp. 182-188.

[10]. L. Klemetinen, H. Mannila, P. Ronkainen, et al. (1994) “Finding interesting rules from large sets of discovered association rules”. Third International Conference on Information and Knowledge Management pp. 401-407.Gaithersburg, USA.

[11]. Li Yang, Mustafa Sanver; Mining Short Association Rules with One Database Scan; Int'l Conf. on Information and Knowledge Engineering; June 2004.

[12]. Mohammed M. Mazid, A.B.M. Shawkat Ali and Kevin S. Tickle(2008), “Finding a Unique Association Rule Mining Algorithm Based on Data Characteristics”, 5th International Conference on Electrical and computer science deptt.

Cite this article as :

S V Subramanyam, "Data Mining Tools for Generate Item Set : Critical Review", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 9, Issue 2, pp.176-183, March-April-2023. Available at doi : https://doi.org/10.32628/CSEIT2390210
Journal URL : https://ijsrcseit.com/CSEIT2390210