Performance Evaluation of Classification-Based Association Rule Mining for Knowledge Discovery

Minakshi Kalra

Department of Computer Science,
Government College, Bahadurgarh, Haryana, India
minakshikalra2006@gmail.com

***Abstract*:** *Association rule mining is used to find the interesting association or correlation relationships among a large set of data items. This paper proposes classification-based association rule mining (CARM) using 'lift measure'. The proposed approach is evaluated on five well-known UCI machine learning datasets. R language is used to implement the proposed technique. The proposed approach is compared with the Naïve Bayes, Zero-R, and C 4.5 techniques. The effects of support and confidence are also investigated on different datasets. The proposed approach exhibits superiority over the existing classification algorithms.*

Keywords*:** ***Association, Apriori, Classification, Correlation, Lift measure

(Article history: Received: 1st November 2020 and accepted 17th June 2021)

I. INTRODUCTION

The analysis of datasets is to be done to organize them in a more meaningful way. This makes us understand the data in a useful way. This can be achieved with the help of data mining that includes techniques from Machine learning, pattern recognition, statistics, and databases [1]. The finding of association rules includes the study of the rate of items occurring together in the databases [2]. In order to detect the frequent itemsets, one makes use of a threshold called support whereas confidence, another threshold, is used to locate the association rules. It has been established that with the help of Predictive Apriori one can mine a high-quality set of association rules [3]. The predictive accuracy plays an important role to ensure the proximity of accurate rules to the top.

Fig. 1 shows the life cycle of data mining [1]. Six phases of data mining are considered. These are data understanding, preparation, modelling, evaluation, deployment and research understanding. The well-known data mining techniques are classification, association rule mining, and clustering (see Fig. 2). The classification techniques are widely used to solve the real-life problems. The class labels are required for classification. However, clustering techniques do not need any class labels. These techniques utilize the dissimilarity measures to determine the group based on the similarity among features. Association rule mining techniques are used to find association between data items. This approach can be used to solve the classification problem. The well-known association rule mining algorithm is Apriori algorithm. Apriori utilizes the two measures namely support and confidence for finding the strong association between features. However, these measures fail to filter out uninterestingness association among features. To handle this issue, 'lift measure' is incorporated in association rule-based classification technique. The purpose of this measure is to mine a high-quality rule set that is as small as possible. The classifiers are used for evaluation as the results are more satisfactory in terms of correctly classified instances.

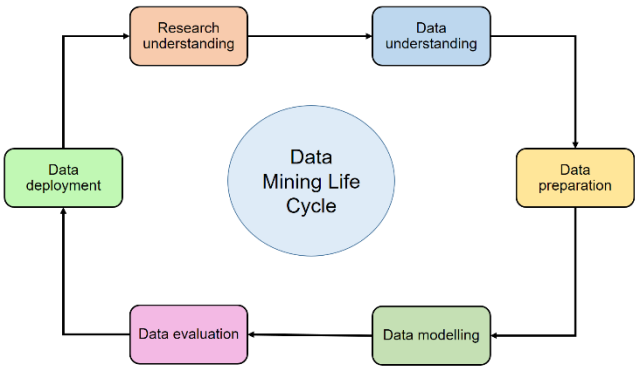

Fig.1. Life cycle of Data Mining [1]

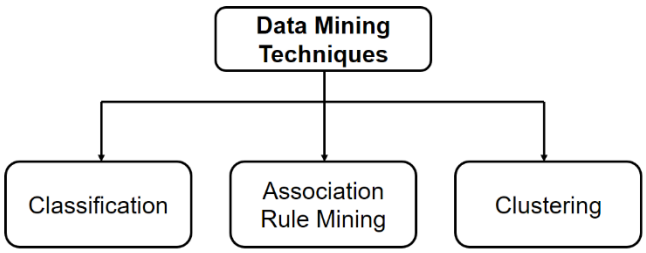

Fig. 2. Classification of data mining techniques

In this paper, the association rule-based classification technique is proposed. The proposed technique utilizes the concept of 'lift measure'. The proposed technique has been validated over the different datasets. The performance of

proposed technique has been compared with existing classification techniques in terms of accuracy. The effects of support and confidence measures are also analyzed. The remaining structure of this paper is as follows: The basic concepts of the Apriori algorithm are covered in Section 2. Section 3 describes the proposed classification algorithm. Experimental results and discussions are mentioned in Section 4. Section 5 draws conclusions.

II. Background

In this section, the basic description of Apriori algorithm is given below:

A. Apriori Algorithm

Apriori algorithm [4], the one amongst important algorithms, mines the frequent itemsets of association rules. The design of the algorithm has two sub tasks:
"(1) Find all the itemsets with support greater than minimum support called as frequent item.
(2) Based on the sets found in (1) all the Association Rules (ARs) are generated and for each frequent itemset A, all the subsets *a* of A is found if ratio of support (A)/support (a) is greater than or equal to min.confidence, to generate the association rules A-a. This algorithm has two sub-processes which are Apriori-gen () and subset (). Apriori-gen () produces a candidate followed by the use of Apriori property to delete candidates of non-frequent subsets [4]". Consequent upon generation of all the candidates, the database is scanned and for each transaction the Subset () is used to identify all the candidate subsets. Then all candidates meet the minimum support from frequent itemset. An illustration of Apriori Algorithm is as follows:
"Consider a database D with nine affairs i.e., |D| = 9. Minimum support threshold value (min.support) = 2/9 = 22% and minimum support count is 2. This is with the assumption that items are stored by the order of dictionary. Itemsets are I1, I2, I3, I4, I5. Scan the database to initialize the source data and form candidate 1-itemsets with all the items of the database for the total to find 1-itemsets. The candidate 1-itemsets C1 i.e., {{I1}, {I2}, {I3}, {I4}, {I5}} consists of each itemset. For each itemset, scan the database and calculate its support count. It will be added to the frequent 1-itemset to determine frequent 1-itemset L1 [4, 7]. To find frequent 2-itemset L2 we should connect L1 to generate aggregation C2 of candidate2-itemsets. C2={{I1,I2},{I1,I3},{I1,I4},{I1,I5},{I2,I4},{I2,I5},{I3,I4}, {I3,I5},{I4,I5}}

Next is to generate association rules from frequent itemsets. The main steps of Apriori algorithm are as follows:

Algorithm 1: Apriori Algorithm [4]

1. Find all frequent itemsets.
2. Get frequent items.
3. Items whose occurrence in database is greater than or equal to the min.support threshold.
4. Get frequent itemsets.
5. Generate candidates from frequent items.
6. Prune the results to find the frequent itemsets.
7. Generate strong association rules from frequent itemsets.
8. Rules that satisfy the min.support and min.confidence threshold."

Fig. 3 depicts general steps followed in association rule-based classification technique. Most Associative classifiers follow this framework [6]. The frequent itemsets are discovered in Step 1. In second step after having identified all the frequent itemsets for each of those which have min.confidence threshold, a rule of the form X→c is generated. The c denotes a class, among all the classes associated with itemset X, as having largest frequency. The ranking and pruning of rules are done in Step 3. The number of rules generated from AR mining is quite large. Hence, rule pruning is required. To avoid the problem of over fitting, proper rule pruning method is employed. It is imperative to go for ranking of rules especially when the test instance has more than one potentially applicable rules [8].

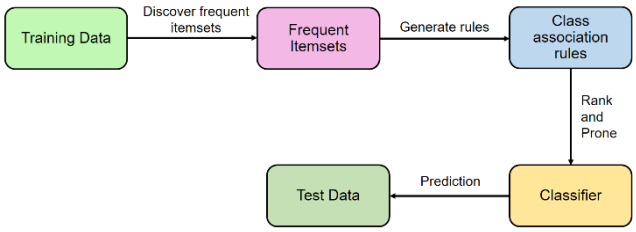

Fig. 3 Classification using Apriori [10]

III. Proposed Classification based Association Rule Mining

As we know that the association rule-based classification technique suffers from the uninterestingness rule generation problem. To resolve this problem, 'lift measure' is utilized in classification technique. The proposed classification algorithm also utilizes association rules. The proposed technique integrates both association rule mining and classification. Algorithm 2 depicts the proposed classification approach.

The mathematical formulation of lift measure (*L (A, B)*) is given below:

$$L(A,B) = \frac{P(AUB)}{P(A)P(B)} \quad (1)$$

where *P(AUB)* represents the occurrence of itemset *A* and *B* together.

Significance of lift measure in the proposed approach

The lift measure uses the concepts of both confidence and expected confidence. It extracts the number of transactions that include consequent divided by the number of transactions. The formulation of lift measure in terms of confidence is given below:

$$Lift = \frac{Confidence}{Expected\ Confidence}$$

For example, a dataset has 100,000 transactions. As many as 2,000 transactions include items X and Y and 800 transactions include item Z. There are 5000 transactions that have only Z item. The association rule "If X and Y are present in transaction, then Z is also present in the same transaction". Confidence of the specified rule is 800/2000=40%. The excepted confidence is 5000/100,000=5%. Lift is 40%/5%=8%. This value provides the information about increase in probability of consequent part if the antecedent part is given.

Besides this, Lift measure also helps determine the type of correlation between itemsets (positively, negatively, and independent) between antecedent and consequent parts of association rule instead of using support and confidence. The use of support and confidence is applicable to single dimension association rule, but this is not applicable at the multidimensional level. To resolve this problem, lift measure is used to apply in multidimensional level.
Based upon this concept, lift measure is incorporated in the proposed classification algorithm.

Algorithm 2: Proposed Approach

1. Initialize the rule set to empty
2. **For each** class label **do**
3. **While** termination criterion is not satisfied
4. Set rule = Apriori (Dataset, support, confidence class label)
5. Eliminate the instances using lift measures from dataset.
6. **End While**
7. Add new rule into rule set.
8. **End For**
9. Classify the dataset using the rule set

IV. Experimental Results and Discussions

In this section, the performance of the proposed approach is validated over UCI machine learning datasets.

A. Dataset used

The datasets used are from the UCI Machine Learning Repository [9]. Table I shows the properties of UCI machine learning datasets. The class attribute is always nominal. Some of these contain missing values.

TABLE I. UCI DATASETS USED AND THEIR PROPERTIES

Dataset	Instance	Numeric	Nominal
Lens	24	0	5
Supermarket	4627	0	217
Soyabean	683	0	36
Weather	14	0	5
Vote	435	8	17
Diabetes	768	8	1
Glass	214	9	1

B. Experimentation 1: Performance of proposed approach

The results obtained through proposed approach are given in Tables II-VII. The effect of confidence on proposed approach is shown in Tables II, IV and VI. For Lens, Supermarket, and Weather datasets, it can be seen that the number of rules is same (i.e., 10) for fixed value of support measure whereas, the number of instances increases with increase in the value of confidence measure.

The effect of support on the proposed algorithm is shown in Tables III, V and VII. For Lens, Supermarket, and Weather datasets, it is observed that the number of instances increases with increase in support measure for fixed value of confidence measure whereas, the number of rules decreases with increase in the value of support measure.

TABLE II. RESULTS OBTAINED ON LENSES DATASET ACCORDING TO CONFIDENCE

Support	Confidence	No. of instances	No of rules
0.2	0.3	5	10
0.2	0.5	6	10
0.2	0.7	7	10
0.2	0.9	8	10

TABLE III. RESULTS OBTAINED ON LENSES DATASET ACCORDING TO SUPPORT

Support	Confidence	No. of instances	No of rules
0.2	0.9	5	10
0.3	0.9	7	1
0.4	0.9	10	1
0.5	0.9	12	1

TABLE IV. RESULTS OBTAINED ON SUPERMARKET DATASET ACCORDING TO CONFIDENCE

Support	Confidence	No. of instances	No of rules
0.45	0.5	2082	10
0.45	0.6	2082	10
0.45	0.7	2151	10
0.45	0.8	2388	10

TABLE V. RESULTS OBTAINED ON SUPERMARKET DATASET ACCORDING TO SUPPORT

Support	Confidence	No. of instances	No of rules
0.3	0.7	1851	10
0.4	0.7	1853	10
0.5	0.7	2121	3
0.6	0.7	2314	0

TABLE VI. RESULTS OBTAINED ON WEATHER DATASET ACCORDING TO CONFIDENCE

Support	Confidence	No. of instances	No of rules
0.2	0.6	4	10
0.2	0.7	4	10
0.2	0.8	5	10
0.2	0.9	6	10

TABLE VII. RESULTS OBTAINED ON WEATHER DATASET ACCORDING TO SUPPORT

Support	Confidence	No. of instances	No of rules
0.1	0.8	3	10
0.2	0.8	4	10
0.3	0.8	5	5
0.4	0.8	6	1

C. *Experimentation 2: Comparison of classification algorithms*

The performance of the proposed classification algorithm is compared with the three existing techniques namely Naïve Bayes classifier (NB), C4.5, and Zero-R. Table VIII shows the accuracy obtained from classification algorithms. The classification accuracies obtained from NB, Zero-R, and C4.5 are 71.46%, 56.12%, and 63.41%, respectively. The average accuracy obtained from the proposed algorithm is 76.61 %, which is higher than other classification algorithms at confidence threshold of 50%. The accuracy obtained from the proposed approach is approximately 13% more than average of accuracies of other three classification algorithms. Fig. 4 shows the comparative analysis of classification techniques in terms of average accuracy.

TABLE VIII. ACCURACY OBTAINED FROM CLASSIFICATION ALGORITHMS

	NB	Zero-R	C4.5	Proposed
Lenses	70.83	62.50	54.17	72.93
vote	90.11	61.38	61.38	88.86
glass	48.60	35.51	63.08	67.76
diabetes	76.30	65.10	75.0	76.90
Average accuracy	71.46	56.12	63.41	76.61

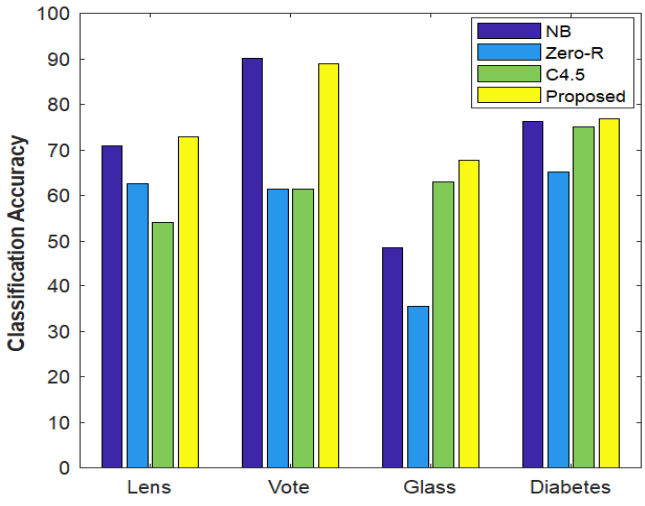

Fig. 4 Comparataive analysis of classification techniques

V. CONCLUSIONS

In this paper, the classification based on association rule mining is proposed. The concept of lift measure has been used to remove the uninterestingness rules. The proposed approach has been validated on UCI machine learning datasets. Experimental results reveal that the proposed approach outperform the existing techniques in terms of classification accuracy. The effect of both support and confidence has also been investigated. These measures play a vital role in the generation of rules and instances. For fixed value of support measure, the number of instances increases with increase in the value of confidence measure. However, the number of rules remains the same. And for fixed value of confidence measure, the number of rules decreases with increase in the value of support measure but the number of instances increases.

In the future study, the proposed approach may be hybridized with metaheuristic techniques for further improvement.

REFERENCES

[1] Agrawal, R., Shrikant, R., (1994)," Fast algorithms for mining association rule". In the proceeding of 20th international conference on VLDB, 1994, pp487-499.

[2] Xu, Y., Gong, J. and Zhou, S., (2005), "Mining association rules with new measure criteria". In the proceeding of international conference on Machine Learning and Cybernetics, 2005, pp. 2257-2260.

[3] Kumar V., and Kumar D., (2018). Binary Whale Optimization Algorithm and its Application to Unit Commitment Problem. *Neural Computing and Applications.*

[4] Liu, Y., (2010), "Study on Application of Apriori Algorithm in Data Mining". In the proceeding of 2nd International Conference on Computer Modeling and Simulation, 2010, pp .111-114.

[5] Sumithra, R., Paul, S., (2010), "Using distributed apriori association rule and classical apriori mining algorithms for grid-based knowledge discovery". In the proceeding of 2nd International Conference on Computing, Communication and Networking Technology, 2010, pp. 1-5.

[6] Prasad, P., Malik, L., (2011), "Using Association Rule Mining for Extracting Product Sales Patterns in Retail Store Transactions". In the International Journal on Computer Science and Engineering, 2011, pp. 2177-2182.

[7] Kumar, V., and Kumar, D. (2019). Automatic Clustering and Feature Selection using Gravitational Search Algorithm and its Application to Microarray Data Analysis. Neural Computing and Applications, Vol. 31 (8), pp. 3647-3663

[8] Shen, Y., Liu, J., Shrn, J., (2010), "The further development of Weka Base on Positive and Negative Association Rules". In the International Conference on Intelligent Computational Technology and Automation, 2010, pp. 811-814.

[9] http://www.ics.uci.edu/_mlearn/MLRepository.html.

[10] Mohanappriya K., Vaidhehi M. (2014) Constructing the Classification systems using Association rule mining for Predictive Analysis.

AUTHOR PROFILE

Minakshi Kalra

She is working as Assistant Professor in the Department of Computer Science at Government College, Bahadurgarh, Haryana, India with more than 13 years' experience. Her areas of research interest include image processing and data mining.