International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-2, December, 2019

Classification of Pruning Methodologies for Model Development using Data Mining Techniques

Parashu Ram Pal, Pankaj Pathak, Vikash Yadav, Priyanka Ora

Abstract: ***Knowledge discovery process deals with two essential data mining techniques, association and classification. Classification produces a set of large number of associative classification rules for a given observation. Pruning removes unnecessary class association rules without losing classification accuracy. These processes are very significant but at the same time very challenging. The experimental results and limitations of existing class association rules mining techniques have shown that there is a requirement to consider more pruning parameters so that the size of classifier can be further optimized. Here through this paper we are presenting a survey various strategies for class association rule pruning and study their effects that enables us to extract efficient compact and high confidence class association rule set and we have also proposed a pruning methodology..***

Keywords: associative classification, data mining, knowledge discovery process, pruning.

I. INTRODUCTION

Associative Classification rule mining concept was first coined in the year 1997 [1] and [2]. The first classifier was named as CBA [3] in 1988. Later classifiers like CPAR [4] in 2003, MCAR [6] in 2005, etc. The first step finds the frequent item-sets and rules for class association. The threshold value is used to remove unwanted sets. Then strong rules are segregated. By using the confidence value the weak rules are pruned. Lastly, a small subset from all these rules is selected and a classifier is made. Many techniques have proposed various methods for optimization of rules to form a classifier [5]. Proposed associative classification techniques use several approaches to find, extract, save, arrange in ranks and prune the redundant rules. This paper aims to compare pruning methodologies that are used in different classifiers in order to find and develop an efficient classifier as an end result in the most efficient manner as possible.

Revised Manuscript Received on December 30, 2019.
* Correspondence Author
***Parashu Ram Pal,** Department of Information Technology, ABES Engineering College, Ghaziabad, India
Pankaj Pathak, Department of Information Technology, Symbiosis Institute of Telecom Management, Pune, India
Vikash Yadav, Department of Information Technology, ABES Engineering College, Ghaziabad, India
Priyanka Ora, Department of Computer Science, Medi-Caps University, Indore, India

II. ASSOCIATIVE CLASSIFICATION RULE MINING

Association rules are derived from association rule mining generated form transactional databases as per their co-occurrences in database. Let we have an item-set with a set of elements such as E = {E1, E2, …, En}, and a transaction set with a number of transactions such as S = {S1, S2, …, Sm}, such that Si∈S having a set of elements E' and E' ⊆ E. Support threshold and confidence threshold (used to determine the importance of the rules) are defined as below:

(i) The items that co-occur in S are known as support denoted by minimum-support, it is decided by the user. I is called an item-set, where I⊆ E, and ∀ai ∈ I co-occur in S., I is said to be frequent item-set if and only if the occurrence of I in S is greater than minimum-support.

(ii) To determine the strongness of an item-set I1 implies another item-set I2, where I1, I2 ⊆ E; and I1 ∩ I2 = {ϕ} is known as Confidence denoted by minimum-confidence. It is also decided by the user.

An association rule represented by I1 ⇒ I2 is correct if frequency for the co-occurrence of I1 and I2 is more than minimum-support, and the confidence of this association rule is more than minimum-confidence. The calculation of support is: (I1 ∪ I2) / (transactions in D_T). The ccalculation of confidence is: support (I1 ∪ I2) / support (I1). I1 ⇒ I2 can be analysed as "if I1 exists, it is likely that I2 will also exists". Associative classification rule mining has two steps: (i) finding all frequent item-sets. This is the most complex and time consuming task that generates all possible item-sets (combinations), 2^n for a set of n items and extracts only those item-sets having frequency at least minimum-support in a training database. (ii) Generating strong association rules. They are produced from frequent item-sets obtained in step (i). All rules having confidence not less than minimum-confidence are extracted. For instance we work with the training data which is shown in table 1. It has three attributes A (A1, A2, A3), B (B1, B2, B3), C (C1, C2, C3) and two class labels (L1, L2). The minimum-support = 30% and minimum-confidence = 70%. The Table 2 shows a classifier with the strong class association rules according to confidence they hold along with their support and confidence.

Table 1. Training Database

TID	A	B	C	Class
i.	A2	B2	C1	L1
ii.	A1	B2	C2	L2
iii.	A1	B3	C3	L2
iv.	A3	B1	C2	L1
v.	A1	B1	C3	L2
vi.	A2	B3	C1	L1
vii.	A3	B3	C2	L1
viii.	A1	B1	C1	L1
ix.	A2	B3	C1	L1
x.	A1	B1	C1	L2

Table 2. Strong Associative Classification Rule Set

Associative Classification Rule		Support	Confidence
Antecedent	Consequent		
A2	L1	03-Oct	03-Mar
B3	L1	03-Oct	03-Mar
A2,C1	L1	03-Oct	03-Mar
A1	L2	04-Oct	04-May
C1	L1	04-Oct	04-May

III. DIFFERENT PRUNING TECHNIQUES FOR ASSOCIATIVE CLASSIFICATION

Pruning means removal of unwanted/unuseful elements. In associative classification, pruning is used to remove infrequent item-sets, infrequent and weak class association rules. Pruning is also used to decide that which rule from the strong associative classification rule set will be included in the final classifier based on its confidence and coverage capacity. Pruning can be applied into three levels in the overall process of associative classification.

(a) Early pruning: early pruning is used not to generate irrelevant candidate sets, remove infrequent item-sets and infrequent associative classification rules. The support threshold value is used at this level.

(b) Intermediate pruning: intermediate pruning is used to remove weak class association rules and extract only strong class association rules. The confidence threshold value is used at this stage to accomplish the task.

(c) Late pruning: late pruning is used to extract only a selected subset of strong class association rules to form the final associative classifier. The strong ness and coverage capacity of the class association rules is used to do it.

Associative Classification easily removes noise and achieves higher accuracy. It a complete rule set than traditional classification techniques [14]. Lasso regulation is used to mine the association rules and to pruning of data. In lasso regulation, variable selection approach is used to introduce a new approach to tackle the problem of rule pruning and summarization. For massive high dimensional data, association rules are used to find the relationships between association and attributes. As the dimension set of data get higher, the data get sparse and results in number of association rules and it's difficult to understand them. By using new approach i.e. CARs (Class-Association Rules) which is based on Lasso regulation, prune the least interesting association rules which will enhance the numbers and the quality of association rules that are obtained to get good results than CBA [15]. Associative classification (AC) is an approach that utilizes the technique of association rule discovery to learn classifier. Associative classification can be categorized into two ways; one is eager associative classification and second is lazy learning associative classification [16]. The His-GC classifier is used. First phase is subset generation and second phase is subset evaluation. In the first phase, all the possible combination of item set is generated based on the testing dataset from the training dataset. Second phase is involved with the calculation of posteriori probability for each of the item set generated from the first phase. Highest probability class is assigned to the test tuple [17]. LLAC (Lazy Learning in Associative Classification): Test data is taken as input and generates the subset for each combination of class value. Support s and confidence c is calculated for each generated subsets. The class label for the new unseen test tuple is assigned as per the highest support and confidence values. HiSC i.e. Highest Subset Confidence algorithm is used to predict the class label of unseen test instance. LLAC outperforms in compare to traditional classification systems [17]. LACI (Lazy Associative Classification using Information Gain): In the previous works, all the possible subsets are generated. Generating all the subsets takes huge computation time. To address this challenge, LACI is proposed. To reduce the numbers of generated subsets, Information gain is calculated for each and every attribute from the training dataset and the highest information gain attribute is chosen to generate the subsets. [18] Apriori-type algorithm is very difficult to predict the defects by implementing it on misbalancing dataset. To improve this, class rules are used to separate the class label, which is the vital type of one of the association rules. It also tells the relationship between the attributes and categories in which dataset is further divided. Empirical comparison with four datasets, is the techniques used to implement the problem which superior than the other classification techniques. The research shows the results that are on the basis of SDP model on class-association [18].

A. Early Pruning Techniques

In early pruning, we use support threshold value to delete the infrequent item-sets. Following methods are used to reduce efforts involved in finding frequent item-sets.

❖ Handling Mutually Exclusive Items

This technique exploits the concept that values of an attribute used to be mutually exclusive means one attribute in an instance contains only one value of attribute. To put this into effect, those attributes having more than one value must be restricted by the candidate generator. This technique firstly accelerates the second pass and secondly, on subsequent passes, candidate generation for candidates pruned by this technique do not required for subset–support based pruning to be explicitly checked

❖ Exploiting the Equivalence of Supports

Usually the item-sets used to have their support near to either support of one of its subsets or support of one of its supersets. To prevent these equivalent rules from being generated or incurring any overhead, the exact-equivalence strategy remove from the set of frequent item-sets any set having a subset with equivalent support before forming the next set of candidates.

❖ Other Techniques

We can find some other techniques that may optimize the candidate rule sets. The algorithm [11] reduces the candidate item-sets by removing those item-sets that are containing more number of items than available number of attributes. The algorithms including [11] do not generates candidate item-sets, and uses other data structure FP-tree to generate the frequent item-sets. We can search some more techniques that can be used to prune the infrequent item-sets more efficiently.

B. Intermediate Pruning Techniques

Intermediate pruning uses confidence threshold value to remove the weak class association rules. But following more methods may optimize the size of class association rules without reducing its classification strength:

❖ Removing Redundant Class Association Rules

The idea behind this technique is to exploit the fact that if a rule R meets the confidence threshold value, then any rule containing R and having confidence less than R will apply to only covered instances. Such rules are termed as redundant class association rules. Redundant rule pruning has been reported in [5]. Its works as follows: Let $R \rightarrow C$ be a general rule, any rule $R' \rightarrow C$ such as $R \subseteq R'$ and $R' \rightarrow C$ has lower confidence in compare to $R \rightarrow C$, will be redundant and they are removed from associative classification rule set. It significantly reduces the size of the class association rule set and minimizes rules redundancy. The algorithms, including [5, 7] have used the pruning of redundant class association rules. It perform pruning immediately as a rule is inserted into the data structure called CR-tree.

❖ Handling Conflicting Class Association Rules

Conflicting class association rules refers to such rules that have similar LHS item-sets but predicting different classes in RHS. Let's take given rules such as $R \rightarrow C1$ and $R \rightarrow C2$ [7]. Proposed a pruning technique considers these conflicting rules and removes them. The algorithm [8] considers such rules as useful knowledge and combines them in a single rules naming multi-label class association rule i.e. $R \rightarrow C1 \vee C2$.

❖ Correlation Testing Between Rule Body & its Class

This concept is taken from statistics that finds the correlation between rule body and its predicting class to determine whether they are correlated or not. The chi-square testing is used to for this purpose. If the discovered class association rule is negatively correlated, it is pruned. If the rule body is positively correlated to its class, it is stored in class association rule set. The algorithm [4] performs the chi-square testing in its rule discovery step to retain or remove the class association rules.

❖ Backward Class Association Rule Pruning

Pruning in decision tree involves pre-pruning and post-pruning. The post-pruning, known as backward pruning is frequently used by decision tree algorithms like C4.5 [9]. First a decision tree is constructed and, then it is decided whether each node and its descendants is replaced or not by a single leaf. The decision is made on the basis of the estimated error using pessimistic error estimation [10] of a node and comparing it with its potential replacement leaf. This backward pruning can also be used in class association rule pruning. The algorithms including [3] have used it to effectively remove the number of extracted class association rules.

C. Late Pruning Techniques

Late pruning is involved with the final step of classifier formation. There are a large number of late pruning techniques available that have been used by different researchers. Some of these techniques are listed out here along-with their scope and importance:

❖ Matching Longest Associative Classification Rules

In this technique we select those class association rules having longest left hand side that matches a particular case. The longest match method is based on the conclusion that the class association rules with longest left hand side will contain more accurate and richer information for the prediction of a class. We know that longest match is more specific and accurate, but the problem with this case is that the support and confidence of the class association rule decreases exponentially as the size of its left hand size increases.

❖ Database Coverage

Database coverage is a very popular pruning technique in associative classification. The algorithms including [3] and [4] have successfully exploited this pruning technique to minimize the size of class association rule set. This method works as follows: first all the rules of class association rule set are sorted in descending order using their confidence. Then each class association rule is tested against the training dataset instances. If a rule correctly classifies some instances in the training dataset, all instances covered by the rule are removed from the training dataset and the rule is marked as candidate rule. If a rule does not correctly covers any instance in training dataset then it is removed from the class association rule set. Finally we get the class association rule set, having candidate rules only.

❖ Lazy Pruning

Lazy pruning aims to remove only those rules from class association rule set that incorrectly classify the training data set instances. In lazy pruning each rule of class association rule set is tested against training dataset instances and we delete those rules that either incorrectly classifies at least one training data set instance or they do not covers even single instance of the training data set. Here we do not delete the covered instances from the training data set as it is done in database coverage method. Lazy pruning considers all class association rules classifying the instances of training data set, whereas the database coverage method considers only single rule classifying an instance. In other words in lazy pruning an instance is classified by several class association rules but in database coverage method an instance is covered by only single class association rule. The experimental results have shown that lazy pruning produces large number of potential class association rules and therefore consumes more memory space in compare to other techniques.

❖ Laplace Accuracy

The Laplace accuracy is used by the associative classification algorithm [4]. It is mainly used in class association rule mining to calculate the expected error of the rules. It calculates expected accuracy for each class association rule before the rule is applied for classification of test instances. The CPAR algorithm has shown that exploitation of Laplace accuracy has produced better results in compare to CBA.

IV. PROPOSED METHODOLOGY

Our proposed methodology aims to reduce the number of rules as well as to study the impact of pruning on accuracy. Database coverage method which provides heuristics to select the rule subset from set of rules. This method has one shortfall that in some cases like when there is no rule to classify it considers largest frequency class for remaining unclassified instances. We can hybrid database coverage pruning with the rule induction for maximum coverage of dataset. The rule which has inducted followed by rule evaluation step through the rank. The proposed method includes rule induction, evaluation of the rules and classifying the test data. Evaluation of the rule helps to seek out whether the rule is able to cover the large part of dataset or not. While evaluating the rule the rank of the rule is constantly revised to reflect the coverage of the rule on test examples. Proposed method tries to acquire as many as possible instances of dataset within the rule and hence less number of rules derived. So there is no majority voting class concept is used for unclassified instances.

V. EXPERIMENTAL EVALUATION

We have compared the effect of three pruning techniques with the number of rules derived by them. These are CBA [3] (pessimistic error and database coverage), MCAR [6] database coverage and lazy pruning [12]. The experiments are done on the fourteen datasets available on UCI M/L data repository [13]. Table 3 gives the number of rules derived from different pruning techniques.

Table 3: Set of Rules Derived by Different Pruning Techniques

S. No.	Name of Data Set	Pessimistic Error & Database Coverage	Proposed Database Coverage	Lazy Pruning
1.	Breast	47	67	22183
2.	Glass	29	39	11061
3.	Heart	43	80	40069
4.	Iris	5	15	190
5.	Labor	17	16	7967
6.	Lymph	35	52	86917
7.	Pima	40	93	9842
8.	Tic-tac	28	28	41823
9.	Wine	11	51	40775
10.	Zoo	5	9	380921

From the above table it is obvious that huge number of classification rule are generated by the lazy pruning technique. The reasons is that lazy pruning involves the method in which a large number of those class association rules (as spare rules) are stored that do not covers even any objects. The proposed database coverage methods overcomes with this problem and removes these spare rules that reduces the size of the associative classifiers. The CBA [3] algorithm generates the associative classifiers of reasonable size in compare to the lazy pruning methods [12].

VI. CONCLUSION

Associative classification is a significant technique of knowledge discovery in data mining field. Pruning techniques are the most important part of the process of constructing and effective classifier with high accuracy standard. Effective class association rule mining yields a classifier that reduces error possibility and increases the accuracy rate and can be deployed for use in big data analytics and data science. The paper discusses the different pruning methods that have been proposed since the inception of the class associative rule mining technique and compares them with the latest ones. Comparison has been made on the results obtained by different associative classification algorithms employing a particular pruning technique. The results show that the database coverage with pessimistic error and database coverage pruning have produced better results in comparison with the lazy pruning methods. They generate compact classifiers that are easy to understand, implement and use for classification of new data items.

REFERENCES

1. A., Azmi M. and Bernado. 2016. "Class Association Rules Pruning using Regularization ." *In Proceeding of International Conference on Computer System and Applications.* IEEE.
2. Agarwal R., Imielinski T. and Swami A. 1993. "Mining Association Rules between Sets of Items in Large Databases." *In Proceedings of International Conference on Management of Data.* Washington DC. 207-216.
3. Bayardo R. 1997. "Brute Force mining of high confidence classification rules." *In proceedings of an International conference on Knowledge Discovery and Data Mining.* Newport Beach, CA, United States. 123-126.
4. Coenen F., and Leng P. 2004. "An Evaluation of Approaches to Classification Rule Selection ." *In Proceedings of International Conference on Data Mining.* Brighton, United Kingdom: IEEE. 359-362.
5. Hiang, Mohammad S. A. and Tze. 2017. "Effects of Pruning on Accuracy in Associative Classification." *In Journal of Informatics and Mathematical Sciences, Vol. 9, No. 4.*
6. J., Quinlan. 1993. "C4.5: Programs for Machine Learning." San Mateo, CA: Morgan Kaufmann.
7. J., Vishwakarma N. and Agrawal. 2013. "Comparative Analysis of Different Techniques in Classification based on Association Rules." *In Proceeding of International Conference on Computational Intelligence and Comuting Research.* IEEE.
8. Liu B., Hsu W. and Ma Y. 1998. "Integrating Classification and Association Rule Mining." *In Proceedings of International Conference on Knowledge Discovery and Data Mining.* New York. 80-86.
9. P., Baralis E. and Torino. 2002. "A Lazy Approach to Pruning Classification Rules." *In Proceeding of International Conference on Data Mining.* IEEE.
10. P., Merz C. and Murphy. n.d. "UCI Repository of Machine Learning Databases." Irvine CA,: University of California.
11. Pal P. R., and Jain R. C. 2010. "CAAC: Combinatorial Approach of Associative Classification." *International Journal of Networking and Applications Vol. 2, No. 1.* 470-474.
12. S., Tamrakar P. and Ibrahim. 2018. "A Review of Lazy Learning Associative Classifications ." *In International Journal of Pure and Applied Mathematics, Vol. 119, No 15.*
13. Tao F., Murtagh F., and Farid M. 2003. "Weighted Association Rule Mining using Weighted Support and Significance Framework." *In proceedings of 9th ACM Conference on Knowledge Discovery and Data Mining.* Washington DC. 661-666.

14. Thabtah F., Cowling P. and Peng Y. 2005. "MCAR: Multi-class Classification based on Association Rule Approach." *In Proceedings of International Conference on Computer System and Applications.* Cairo, Egypt: IEEE. 1-7.
15. Thabtah F., Cowling P. and Peng Y. 2004. "MMAC: A new Multi-class Multi-label Associative Classification Approach." *In Proceedings of International Conference on Data Mining.* Brighton, United Kingdom. 217-224.
16. Y., Han J. Pei and Yin. 2000. "Mining Frequent Patterns without Candidate Generation." *In Proceedings of International Conference on ACM SIGMOD.* 1-12.
17. Yin X., and Han J. 2003. "Classification based on Predictive Association Rules." *In Proceedings of International Conference of Data Mining.*
18. Yuanxum Shao, Bin Liu Guoqi Li and Shihai Wand. 2017. "Software Defect Prediciton based on Class Association Rules." *In Proceeding of International Conference on Reliability System Engineering.* IEEE. 1-7.

AUTHORS PROFILE

Dr. Parashu Ram Pal, obtained Ph.D. in Computer Science. He is working as a Professor in Department of Information Technology, ABES Engineering College, Ghaziabad, India. He has published three books and more than 40 Research Papers in various International, National Journals & Conferences. His area of interests are Data Mining, Computer Architecture, Computer Graphics and Operations Research. He is devoted to Education, Research & Development for more than twenty years and always try to create a proper environment for imparting quality education with the spirit of service to the humanity. He believes in motivating the colleagues and students to achieve excellence in the field of education and research**.**

Dr. Pankaj Pathak obtains Masters and Ph.D. in 2005, 2014 respectively. He is working as an Assistant Professor in Symbiosis Institute of Telecom Management. His area of interests are Data Mining, AI, and Smart Technologies. He has Published Several Research papers in the area of Data Mining, IOT security and Speech Recognition Technology.in the field of education and research**.**

Dr. Vikash Yadav received his Ph.D. (Computer Science & Engineering) degree from Dr. A.P.J Abdul Kalam University (Formerly U. P. Technical University) Lucknow, (U.P. India) in 2017. He is currently working as an Assistant Professor in the Department of Computer Science & Engineering, ABES Engineering College, Ghaziabad, India and has more the 7 years of Teaching/Research experience and published more than 30 research papers in various National/International Conferences/Journals. He is also a reviewer of various SCI/SCIE/Scopus indexed journals. His area of interest includes Data Structure, Data Mining, Image Processing and Big Data Analytics.

Dr. Priyanka Ora completed her Masters and Ph.D. in Computer Science. She is working as an Assistant Professor in Department of Computer Science, Medi-Caps University, Indore, India. She has more than six years of academic experience. She published published more than 10 Research Papers in various International, National Journals & Conferences. Her area of interests are cloud computing, cyber security, internet of things.