Journal of Informatics and Mathematical Sciences
Vol. 9, No. 4, pp. 1047–1051, 2017
ISSN 0975-5748 (online); 0974-875X (print)
Published by RGN Publications
DOI: 10.26713/jims.v9i4.1006

http://www.rgnpublications.com

Research Article

Effects of Pruning on Accuracy in Associative Classification

Mohammad Abrar*and Alex Tze Hiang Sim

Faculty of Computing, Universiti Teknologi Malaysia, Johor, Malaysia
***Corresponding author:** muhammadabrar78@gmail.com

Abstract. A number of techniques are presented in the literature for pruning in both decision tree as well as rules based classifiers. The pruning is used for two purposes; namely, Improve performance, and improve accuracy. As the pruning is reducing the set of rules as well as the size of the tree, the probability of improvement in performance is, therefore high. While on the other side, the pruning may eliminate the interesting information which can lead to reducing the accuracy. In this research, the effects of pruning on the accuracy are studied in detail. The experiments were carried out on the same techniques with and without using pruning strategies and the results of both types are compared. The analysis of the five algorithms over fourteen datasets showed that the unwise selection of pruning strategy could reduce the accuracy.

Keywords. Rule pruning; Associative classification; Classification; Association rules mining; Data mining

MSC. 91-03

Received: July 31, 2016 **Accepted:** September 12, 2016

1. Introduction

Recent studies in data mining revealed that *Associative Classification* (AC) of data mining approach builds competitive classification classifiers with reference to accuracy when compared to classic classification methods including decision tree and rule-based. Nevertheless, AC algorithms suffer from a number of known defects as the generation of a vast number of rules which makes it hard for end-users to maintain and understand its outcome and the possible over-fitting issue caused by the confidence-based rule evaluation used by AC.

1.1 General Methodology of AC

AC have significant differences with AR. The AC is working with CAR which is a particular case of AR where the consequent allows only class label while in AR the consequent my have any disjoint subset of attributes. The goals of both techniques are different where one finds an association among attributes while others find rules that maximize the accuracy of the classifier Thus, in AR class attribute is not required while in CAR it is essential. Therefore the development of AC is different from generating AR. The AC can be divided into following three phases.

(1) Generate all frequent CAR using any Association Rule Miner.

(2) Sort the rules based on predefined criteria and prune the not interesting rules to build a classifier.

(3) Classify the test data and evaluate the classifier.

In AC the most expensive phase in CAR generation. It is using a priori or its variant to generate the rule. ARM needs to generate an exponential number of rules that qualify the minimum threshold. The generation of an exponential number of rules in computationally costly due to a number of passes through the dataset [1–5]. After generation of all frequent rules, the AC rank and prune the rules in phase 2. Different ranking schemes are used in ranking which is discussed in detail. Once the rules are ordered, the pruning is used to eliminate the redundant and non-interesting rules. The remaining subset of rank rules makes the classifier. Once the subset of rank rules is identified for the classifier, then this model is used to predict the class label for test data where the class label is unknown. The classifier is then evaluated based on the accurate prediction of the test data items.

2. Literature Review

The concept of classification comes from statistics which include, specifically: prediction of a categorical variable, while the other is called regression [6]. In Data Mining and Machine Learning, classification is used in the same spirit - to predict the unknown variable and its application ranging from filling missing values to the prediction and diagnosis of complex diseases, speech and handwriting recognition to image analysis. To adequately cover the topics, this section is divided into subsections. First, there is the Associative Classifier which highlights the general framework of AC and the previous work followed by Rules Pruning and Rule Ranking used in AC so far.

Different types of pruning strategies have been used in classification. These procedures are taken from various fields such as decision trees, and statistics. The prominent strategies include χ^2 [7], Pessimistic Error Estimation (PEE) [8] which are taken from statistics.

These techniques can be used in either phase of classification, whether it is a rule discovery phase or classification phase. The former is called pre-pruning while the latter is known as post-pruning. The commonly used technique for post pruning is Database coverage [9] and Lazy [10] while Pearson Correlation Coefficient is used in pre-pruning. Besides these two staged pruning technique, the basic pruning is applied before the generation of rules in term of minimum support and confidence. This pruning is used in Associative Classifiers only while

the other strategies are used in both Associative classifier and other Rules-based and Tree-based algorithms.

3. Pruning ad its Effects

The Associative Classification uses *Association Rules* (AR) to develop the classifiers. These ARs has the issue with the exponential number of rules generation where they are not necessarily all interesting. Therefore, there is a need for selecting the association rules which are necessary for the classifier. There are a number of reasons for reducing the list of rules: specifically, a) redundancy in rules, b) Classifier time, and c) harmful rules that degrade the ratio of correct classification.

Therefore, different types of pruning strategies are used in the AC. The primary sources of these techniques are Rules based, Decision Tree, or Statistics, i.e. the Pessimistic Error Estimation is taken from Decision Tree while chi-square testing and coefficient correlation are from statistics. Similarly, the database coverage and lazy rules are based on the association rules.

The rule pruning plays twofold advantages; a) reducing the classifier size to improve the performance and b) by eliminating the harmful rules to increase the classification accuracy rate. Thus, the classification size performs an important role in the accuracy and efficiency; the larger classifier may improve the accuracy, but reduce the speed while the reduced classifier can help in improving the efficiency but can decrease the accuracy. Therefore, a wise trade-off between the accuracy and effectiveness is needed that can be achieved by a wise pruning strategy.

4. Results and Discussion

We randomly selected 14 different datasets from more commonly cited literature, including [2–5]. All datasets are available on-line at UCI Machine Learning Repository [11].

Five algorithms are selected for experiments where two algorithms belong to Rules-based, and three belongs to Tree-based classifiers. The rule-based classifiers include CBA and PART while the Tree-based classifiers include J48, J48Graft, and CART. All experiments were conducted on core i7 3.60GHz with 16GB of memory. The results were generated by WEKA 3.7.13. In order to make the regeneration of result easier, all experiments were conducted with the default parameters setting. The numeric datasets were discretized using unsupervised discretizer of Weka.

Two sets of results were generated during experimentation; one with pruning strategy and the second without pruning strategy. The results are presented in table 0. The table shows that in almost all techniques, the results of the algorithm without pruning strategy are better than that of with pruning strategy.

The results indicate that the PART and the J48 graft non-pruning strategies are performing better that the pruning strategies where eight out of fourteen datasets performed better as compared to its counterpart. The second and third are the CART and J48 with seven and six datasets respectively. While on the average all algorithms without pruning strategies are better that its counterparts. The overall win/loss/tie comparison is shown in Table 2.

Table 1. Accuracy comparison of pruned and unpruned classification techniques

Dataset	PART Prune	PART Unpruned	CBA Prune	CBA Unpruned	J48 Prune	J48 Unpruned	J48 Gr Prune	J48 Gr Unpruned	CART Prune	CART Unpruned
balance-scale	**76.72**	75.68	**45.76**	45.76	65.12	**68.73**	66.00	**70.25**	**77.68**	**77.68**
balloons	**100.00**	**100.00**	50.00	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
car	**95.86**	95.81	70.02	**77.78**	92.39	**94.01**	92.39	**94.01**	**97.22**	97.14
data_banknote	97.49	**98.83**	56.38	**67.86**	98.36	**98.65**	98.36	**98.69**	98.76	**98.80**
hayes-roth	75.27	**79.15**	**37.91**	37.91	**71.59**	70.08	**71.59**	70.08	**81.04**	80.99
iris	**95.33**	93.00	**69.67**	64.67	**96.00**	94.33	**96.33**	94.67	**96.00**	95.33
iris2D	93.00	**95.00**	**69.67**	64.67	**96.00**	**96.00**	**96.33**	**96.33**	**96.00**	95.67
page-blocks	**93.61**	**93.61**	**90.06**	**90.06**	93.14	**93.61**	92.97	**93.48**	93.67	**93.82**
pima_diabetes	**72.85**	69.54	**67.19**	**67.19**	**73.77**	70.58	**73.90**	71.29	**73.71**	71.56
tae	48.67	**54.60**	**34.42**	**34.42**	50.00	**57.56**	50.33	**58.23**	52.65	**55.29**
tic-tac-toe	94.36	**97.13**	**65.34**	**65.34**	84.91	**85.96**	85.33	**86.59**	93.21	**93.94**
weather.nominal	65.00	**67.50**	**70.00**	50.00	50.00	**62.50**	50.00	**62.50**	45.00	**65.00**
weather.numeric	57.50	**60.00**	**70.00**	67.50	**57.50**	37.50	**57.50**	37.50	62.50	**75.00**
breast-cancer	94.28	**95.21**	**89.77**	**89.77**	**94.28**	93.64	94.14	**94.57**	**93.71**	92.85
Average	82.853	**83.933**	63.299	**65.924**	80.219	**80.225**	80.369	**80.585**	82.939	**85.219**

Table 2. The win-loss comparison of the pruned and unpruned strategies

	PART	**CBA**	**J48**	**J48 Gr**	**CART**
WIN	8	3	7	8	6
LOSS	4	4	5	4	6
TIE	2	7	2	2	2
Win + Tie	10	10	9	10	8

5. Conclusion and Future Work

The extensive experiments were conducted during the research and results were generated where it is shown that pruning strategy can affect the accuracy negatively. Pruning is highly advantageous in removing redundant rules which always lead to improved performance while sometimes helps to improve accuracy as well. On the other hand, whenever it tries to remove the rules based on the interestingness, it mostly affects the accuracy negatively. The experiments also showed that no single pruning technique always results in better performance as well as high accuracy. Mostly selection of AC technique will be based on the nature of data.

Therefore, it is concluded that pruning can be used for improving performance in a general, while for accuracy improvement, careful selection should be made. There is no universal rule which can identify the pruning strategy for all situations.

Acknowledgment

This research is supported by research grant FRGS (vot no:4F431, 13H89). The authors also acknowledge the support previced by the Ministry of Higher Education (MOHE), Research Management Centre (RMC), and Universiti Teknologi Malaysia.

Competing Interests

The authors declare that they have no competing interests.

Authors' Contributions

All the authors contributed significantly in writing this article. The authors read and approved the final manuscript.

References

[1] W.H.B. Liu and Y. Ma, Integrating classification and association rule mining, in *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining*, pp. 80 – 86, 1998.

[2] W. Li, J. Han and J. Pei, Cmar: Accurate and efficient classification based on multiple class-association rules, in *Proceedings IEEE International Conference on Data Mining* (ICDM 2001), pp. 369 – 376, 2001.

[3] F. Thabtah, P. Cowling and Y. Peng, Mcar: multi-class classification based on association rule, in *The 3rd ACS/IEEE International Conference on Computer Systems and Applications*, p. 33, 2005.

[4] J. Han, Cpar: Classification based on predictive association rules, in *Proceedings of the Third SIAM International Conference on Data Mining*, Vol. 3, pp. 331 – 335, Siam, 2003.

[5] N. Abdelhamid, A. Ayesh, F. Thabtah, S. Ahmadi and W. Hadi, Mac: A multiclass associative classification algorithm, *Journal of Information & Knowledge Management* **11** (2), 2012.

[6] T. Hastie, R. Tibshirani, J. Jerome and H. Friedman, *The Elements of Statistical Learning*, Vol. 1, Springer, New York, 2001.

[7] G.W. Snedecor and W.G. Cochran, *Statistical Methods*, 8th edn., Ames, Iowa.

[8] J.R. Quinlan, *C4. 5: Programs for Machine Learning*, Vol. 1, Morgan Kaufmann, 1993.

[9] B. Liu, W. Hsu and Y. Ma, Integrating classification and association rule mining, in *Proceedings of the 4th*, pp. 80 – 86, 1998.

[10] E. Baralis, S. Chiusano and P. Garza, On support thresholds in associative classification, in *Proceedings of the 2004 ACM Symposium on Applied Computing*, pp. 553 – 558, ACM, 2004.

[11] K. Bache and M. Lichman, *UCI Machine Learning Repository*, 2013.