HYBRID ENSEMBLE LEARNING APPROACH FOR GENERATION OF CLASSIFICATION RULES

HAN LIU, ALEXANDER GEGOV, MIHAELA COCEA

School of Computing, University of Portsmouth, Portsmouth, PO1 3HE, United Kingdom
E-MAIL: han.liu@port.ac.uk, alexander.gegov@port.ac.uk, mihaela.cocea@port.ac.uk

Abstract:

Due to the daily increase in the size of data, machine learning has become a popular approach for intelligent processing of data. In particular, machine learning algorithms are used to discover meaningful knowledge or build predictive models from data. For example, inductive learning algorithms involve generation of rules which can be in the form of either a decision tree or if-then rules. However, most of learning algorithms suffer from overfitting of training data. In other words, these learning algorithms can build models that perform extremely well on training data but poorly on other data. The overfitting problem is originating from both learning algorithms and data. In this context, the nature of machine learning problem can be referred to as bias and variance. The former is originating from learning algorithms whereas the latter is originating from data. Therefore, reduction of overfitting can be achieved through scaling up algorithms on one side or scaling down data on the other side. Both bias and variance can be reduced through use of ensemble learning approaches. This paper introduces particular ways to address the issues on overfitting of rule based classifiers through both scaling up algorithms and scaling down data in the context of ensemble learning.

Keywords:

Data mining; Machine learning; Ensemble learning; Inductive learning; If-then rules; Rule based classification;

1. Introduction

Machine learning has become a powerful approach for intelligent data processing due to the daily increase in data size. In practice, machine learning algorithms are popularly used for discovery of meaningful knowledge and building of predictive models through learning from data. Inductive learning is a special type of learning methods and involves generation of rules. In general, rule generation can be divided into two categories: divide and conquer [1] and separate and conquer [2]. The former generates rules in the form of decision trees whereas the later generates if-then rules directly from training instances. In practice, rule generation is popularly involved in classification tasks, i.e. classifying data instances into a particular category by using the rules generated from training data.

As reported in [3, 4], rule based classifiers are usually not stable and are sensitive to the change of data sample. This is because most rule learning algorithms suffer from overfitting of training data. In other words, rule sets generated by these algorithms result in high level of accuracy on training data, but low level of accuracy on test data. The overfitting problem is due to bias and variance. As introduced in [4], bias means errors originating from learning algorithms and variance means errors originating from data. In this context, overfitting can be reduced through scaling up algorithms and scaling down data [5]. The former way is to reduce bias on algorithms side whereas the latter way is to reduce variance on data side. From this point of view, it is necessary to reduce overfitting through both ways mentioned above. Ensemble learning can achieve this goal, which will be introduced in Section 2 in more depth.

The rest of this paper is organized as follows: Section 2 introduces the concepts, popular methods and recent advancement of ensemble learning; Section 3 introduces a newly developed hybrid ensemble learning approach. An experimental study is reported in Section 4 and results are also discussed. Section 5 summarizes the contribution of this paper and highlights further directions of this research area.

2. Related work

Section 1 introduces the background of machine learning particularly on rule learning algorithms. The nature of learning problem, which is referred to as bias and variance, is also pointed out. Ensemble learning is stressed as an effective approach that addresses the issues relating to overfitting. This section introduce particular ways to reduce overfitting through scaling up algorithms or scaling down data. In particular, bagging and random forests, which are two popular methods of ensemble learning, are critically reviewed to identify their advantages and disadvantages. Another approach, which is recently developed and referred

to as collaborative and competitive random decision rules (CCRDR), is discussed with regard to its strength in filling the gaps that exist in bagging and random forests as well as its unresolved weaknesses.

2.1. Ensemble learning concepts

Ensemble learning is usually adopted to improve the overall accuracy in prediction. As mentioned in [4], ensemble learning can be done in parallel or sequentially. In the former way, there is no collaboration involved in training stage and only the predictions by different models are combined for final prediction making. In the latter way, the first iteration involves learning new concepts and the following iterations all involve correcting the latest learned concepts.

Both parallel and sequential learning can be achieved through scaling up algorithms or scaling down data. In parallel learning, scaling up algorithms is through combination of different algorithms, each of which generates a model on the same training set. The predictions by these models are combined for final prediction. Scaling down data is through use of a single algorithm for generation of different models on different samples of training data. In sequential learning, scaling up algorithms is through combination of different algorithms in the way that the first algorithm learns a model that is iteratively corrected by the latter algorithms. Scaling down data is in the way that a single algorithm is iteratively used to build models on different versions of training data. In particular, at each iteration, the training instances are weighted to different extents on the basis of the model quality estimated using validation data. Finally, these predictions by different models are combined to predict unseen instances.

For both parallel and sequential learning approaches, voting is involved in the testing stage when the independent predictions are combined to predict an unseen instance. Some popular methods of voting include equal voting, weighted voting and naïve Bayesian voting [4]. Equal voting is used for Bagging and Random Forests, which is introduced in Section 2.2 and 2.3. More details on weighted voting are introduced in Section 2.4.

2.2. Bagging

The term Bagging stands for bootstrap aggregating. It is a popular method developed by Breiman [6] and follows the parallel ensemble learning approach. Bagging involves sampling of data with replacement. In detail, the Bagging method is to take a sample with the size *n*, where *n* is the size of the training set, and to randomly select instances from the training set to be put into the sample set. This indicates that some instances in the training set may appear more than once in the sample set and some other instances may never appear in the sample set. On average, a sample is expected to contain 63.2% of the training instances [3, 4, 6]. In the training stage, the generation of classifiers, each of which results from a particular sample set mentioned above, are parallel to each other. In the testing stage, their independent predictions are combined to predict the final classification based on equal voting. As concluded in the literature [3, 4], Bagging is robust and does not lead to overfitting due to the increase of the number of generated models. Therefore, it is useful especially for those non-stable learning methods with high variance.

2.3. Random Forests

Random forests is another popular method [7] that can be seen as a special case of bagging. In particular, decision trees must be the base classifiers generated on each sample of training data. In addition, the attribute selection at each node of a decision tree is random to some extents. Otherwise, this ensemble learning method only belongs to Bagging. In this sense, at each node, there is a subset of attributes randomly chosen from the training set and the one which can provide the best split for the node is finally chosen [8]. In the training stage, the chosen algorithm of decision tree learning is used to generate classifiers independently on the samples of the training data. In the testing stage, the classifiers make the independent predictions that are combined to make the final prediction based on equal voting. As concluded in the literature [4], the random forests algorithm is robust because of the reduction of the variance for decision tree learning algorithms.

2.4. Collaborative and Competitive Random Decision Rules

The CCRDR approach has been developed in [9] in order to fill the gap that exists in Bagging and Random Forests. The basic idea of this approach is illustrated in Figure 1.

CCRDR stands for Collaborative and Competitive Random Decision Rules, which indicates that the ensemble learning framework involves both collaboration and competition. Therefore, the above approach is designed partially to overcome the limitation (of Bagging and Random Forests) that there is only a single base algorithm involved in the training stage, which cannot always generate robust hypothesis due to the absence of competition in this stage. In order to overcome the above limitation, the CCRDR approach is designed in a way that includes multiple base algorithms for training. On the basis of the design, there is thus competition among the models generated on the same

sample of training data. In other words, there are multiple learning algorithms applied to each sample of the training data, which results in the generation of multiple models on each sample. In this context, it becomes achievable to find better models to be involved in the testing stage and worse ones to be absent through competition among these models. The competition is based upon the weight (confidence) of each of the models by means of overall accuracy estimated using validation data. In the CCRDR framework, on each sample of training data, only the model with the highest weight (confidence) is eligible to be involved in the testing stage. The development of the CCRDR aims to enable that on each sample of training data the hypothesis generated becomes much stronger.

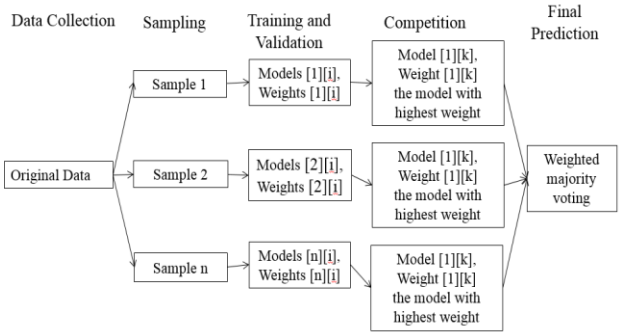

Figure 1. Procedures of CCRDR Ensemble Learning [9]

On the other hand, as mentioned in Section 2.1, voting is usually involved in the testing stage for ensemble learning approaches such as Bagging and Random Forests. As introduced in [4], voting can be based on different criteria such as equal voting and weighted voting. Bagging and Random Forests adopt equal voting for final prediction in the testing stage. However, in classification tasks, weighted voting is usually preferred to equal voting. This is because of the possibility that some classifiers are highly reliable whereas the others are less reliable. For example, there are three base classifiers: A, B and C. A predicts the classification X with the weight 0.8, and B and C predict the classification Y with the weights 0.55 and 0.2 respectively so the final classification is X if using weighted voting (weight for X: 0.8> 0.55+0.2=0.75) but is Y if using equal voting (frequency for Y: 2>1).

For weighted voting, the weight can also be determined in accordance with different criteria. Several possible ways to determine the weight have been discussed in [9]. These criteria include overall accuracy, precision and recall, all of which are popularly used for evaluating the quality of a classifier. These criteria are also compared experimentally in terms of their effectiveness for evaluating the reliability of a classifier in ensemble classification tasks. The experimental results reported in [9] indicate that precision is more effective than the other two in estimation of classifier reliability. The reasons can be explained by the following:

- Overall accuracy is to show the performance of a classifier in predicting classes on average. In other words, high overall accuracy does not necessarily indicate that the classifier can accurately predict each individual class. It is possible that the classifier performs well on some classes but poorly on the others.
- Recall is less reliable than precision. For example, there are 5 positive instances out of 20 in a test set and a classifier correctly predicts the 5 instances as positive but incorrectly predicts other 5 instances as positive as well. In this case, the recall/true positive rate is 100% as all of the five positive instances are correctly classified. However, the precision on positive class is only 50%. The above case indicates that high recall may result from low frequency of a particular class.

The CCRDR framework still has gaps that need to be filled. In particular, different learning algorithms involved in the training stage for modelling could help generate better models from each sample of training data. However, these learning algorithms do not collaborate with each other. For separate and conquer rule learning, collaboration can be achieved in the way that different rule learning algorithms are combined to generate a single rule set on each sample of training data. More details on this are justified in Section 3.

3. Hybrid ensemble rule based classification

As mentioned in Section 2.4, separate and conquer rule learning can be enhanced through the way that different algorithms collaborate to generate a single rule set on a given data set. Therefore, the CCRDR framework still has gaps to be filled. The modified framework is illustrated in Figure 2. The new framework is referred to as hybrid ensemble rule based classification, due to the involvement of data sampling and algorithms collaborations for reduction of bias and variance respectively.

3.1 Key features

The modified framework for ensemble rule based classification differs from the CCRDR illustrated in Figure 1 in two aspects.

The first aspect is on the number of generated rule based classifiers learning from each sample of training data, i.e. there is only one classifier generated from each sample data.

The second aspect is on the removal of the competition stage illustrated in Figure 1. In other words, after the

completion of training and validation, all the classifiers are not involved in competition with regard to their reliabilities but are combined straightaway to predict unseen instances.

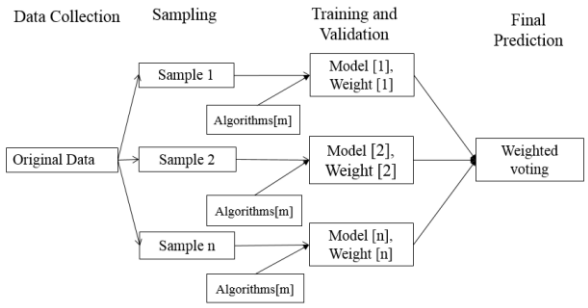

Figure 2. Hybrid ensemble rule based classification framework

The two aspects for modification made to the CCRDR framework are due to the change in the strategy for combination of different learning algorithms. In the modified framework, the learning algorithms are combined to generate only a single classifier on each sample of training data. Therefore, there is no competition necessary.

The combination of learning algorithms involves collaborations in a Macro vision but competitions in a Micro vision. This is because these algorithms are combined to generate a rule set (a set of rules), but each of these rules is actually generated through competitions among rules generated by different algorithms at each iteration of rule learning. In other words, at each iteration, each of the chosen algorithms generates a rule that is compared with the other rules generated by other algorithms in terms of rule quality. The competitions mentioned above aim to have each single rule with a quality as high as possible for each rule set generated from a sample of training data. The quality of a single rule can be estimated by some statistical measures, such as J-measure [10] as illustrated in formula (1).

$$J(Y, X = x) = P(x) \cdot (P(y \mid x) \cdot (\frac{P(y \mid x)}{P(y)}) + (1 - P(y \mid x)) \cdot (\frac{1 - P(y \mid x)}{1 - P(y)})) \quad (1)$$

The notation *P(x)* and *P(y)* are read as the coverages of rule antecedent (left hand side) and rule consequent (right hand side) respectively. In addition, the notation *P(y |x)* is read as posterior probability that the rule consequent occurs given the antecedent as the condition. Detailed overviews of rule quality measures can be found in [11, 12].

3.2 Justification

The hybrid ensemble rule based classification framework can fill the gaps that exist in Random Forests and CCRDR to a large extent.

In comparison with Random Forests, the new framework illustrated in Figure 2 employs algorithms that follow separate and conquer rule learning. This rule learning approach has a higher flexibility than divide and conquer rule learning.

Divide and conquer approach involves generating a set of rules, all of which are strongly connected with each other to fit in a tree structure. In this context, changing even one rule is likely to destroy the whole tree as a result of that the rules cannot fit in a tree any more. This is because each rule is represented as a branch of the tree that has some common parts with other branches. On the other hand, each of the branches is grown parallel to all the others.

In contrast, separate and conquer approach involves generating a set of modular rules. These rules are not connected each other in terms of rule representation. On the other hand, the rules are generated sequentially, i.e. the completion of the generation for one rule is followed by the start of the generation for another rule.

On the basis of above comparison, separate and conquer approach enables collaborations involved in the training stage of ensemble rule learning whereas divide and conquer approach cannot. This is because the former approach enables the evaluation of each single rule once the rule is generated due to the sequential rule generation, whereas the latter approach cannot achieve it due to parallel rule generation. Therefore, the hybrid ensemble rule based classification framework can fill the gap that exists in Random Forest.

In comparison with the CCRDR framework, the improvement is on the quality of the classifiers, each of which is generated from a particular sample of training data. This is because of the possibility originating from CCRDR that some of the rules are of higher quality but the others are of lower quality, while there is only one rule learning algorithm involved in the training stage. CCRDR involves a competition between algorithms per rule set. In other words, the competition is made after each of the algorithms has generated a rule set, in order to compare the quality of a whole rule set. In contrast, the hybrid ensemble approach involves such a competition per rule generated. In other words, the competition is made once each of the algorithms has generated a rule in order to compare the quality of each single rule generated by a particular algorithm.

Overall, in the context of ensemble rule based classification, the new framework illustrated in Figure 2 shows a greater robustness and flexibility than the other two.

4. Experimental setup and results

The hybrid ensemble rule based classification framework is validated in an experimental study and is compared with Random Forests and CCRDR, in terms of classification accuracy. The experiments are conducted using 10 data sets retrieved from the UCI repository [13]. The characteristics of these data sets are shown in Table 1. The results are also discussed in both quantitative and qualitative terms.

TABLE 1. DATA SETS

Name	Attribute Types	#Attributes	#Instances	#Classes
credit-a	mixed	15	690	2
credit-g	mixed	20	1000	2
vote	discrete	16	435	2
hepatitis	mixed	20	155	2
lung-cancer	discrete	32	57	3
lymph	mixed	19	148	4
breast-cancer	discrete	9	286	2
breast-w	continuous	10	699	2
labor	mixed	17	57	2
heart-h	mixed	76	920	4

NB: Mixed means containing both discrete and continuous attributes

The above data sets are chosen considering the computational constraints as ensemble learning approaches are usually more expensive than base learning approaches. In addition, the CCRDR and hybrid ensemble rule based classification frameworks both involve combinations of different learning algorithms and thus are more expensive than Random Forests in general. On the basis of above consideration, all the chosen data sets have lower dimensionality and smaller number of instances. In addition, these data sets contain both discrete and continuous attributes in order to validate that the newly developed framework can effectively deal with both types of attributes.

In the CCRDR and hybrid ensemble rule based classification frameworks, two rule learning algorithms, which are referred to as Prism [14] and Information Entropy Based Rule Generation (IEBRG) [15], are employed to generate base classifiers. In general, both frameworks can employ any types of rule learning algorithms, which follow separate and conquer approach and suit the training data, to be combined for generating base classifiers. In this experimental study, these are the only algorithms employed due to the consideration of computational constraints. The computational complexity of this kind of ensemble learning approaches is approximately *n* times the complexity of a single learning algorithm, where n is the number of base algorithms, if no parallelization is adopted.

On the other hand, these two algorithms are considered representative for two types of separate and conquer rule learning, namely forward rule generation and backward rule generation. IEBRG is seen as the representative for the former type. This is because IEBRG aims to specialize the left hand side of a rule until all instances covered by the rule belong to the same class. In contrast, Prism aims to first give a target class as the consequent of the rule, and then to search for causes as the condition that can derive this target class until the adequacy condition is found. Therefore, Prism is seen as the representative for backward rule generation.

For the hybrid ensemble rule based classification framework, it is required to employ those algorithms which can complement each other with respects to their advantages and disadvantages. Prism can typically overcome some limitations of decision tree learning algorithms to a large extent, such as replicated subtree problem [14] and thus is able to make the framework competitive to Random Forests. The IEBRG complements the Prism with regard to some of disadvantages of the latter algorithm as discussed in [15]. Therefore, choosing these two algorithms is appropriate to enable that the chosen algorithms are complementary to each other. In addition, the J-measure is used to evaluate rule quality for the partial fulfilment that each single rule is generated to have a quality as high as possible as mentioned in Section 3. This choice of metric is due to the fact that J-measure takes into account both simplicity and goodness of fit of a single rule as justified in [10].

The experiments are conducted by splitting a data set into a training set and a test set in the ratio of 70:30. For each data set, the experiment is done 10 times and the average of the accuracies is taken for comparative validation. As mentioned earlier, ensemble learning approaches are usually computationally more expensive. Therefore, cross validation is not adopted in this study. The results are presented in Table 2 as below.

TABLE 2. ACCURACY

Data set	Random forests	CCRDR	Hybrid
credit-a	85%	70%	**87%**
credit-g	72%	71%	**74%**
vote	97%	93%	**98%**
hepatitis	85%	84%	**92%**
lung-cancer	70%	86%	**93%**
lymph	86%	70%	**90%**
breast-cancer	65%	78%	**81%**
breast-w	**97%**	85%	91%
labor	88%	**90%**	88%
heart-h	83%	79%	**85%**

Table 2 shows that the hybrid ensemble rule based classification framework outperforms random forests and CCRDR in 8 out of 10 cases. On ‘breast-w’ and ‘labor’ data sets, the newly developed framework performs a bit worse than random forests and CCRDR.

The results indicate that it is necessary to take both

scaling up algorithms and scaling down data in order to comprehensively improve classification accuracy like the hybrid ensemble rule based classification framework. In this way, accuracy can be improved through reduction of both bias and variance. In contrast, random forests only involves scaling down data and nothing on scaling up algorithms. Therefore, random forests only enables the reduction of variance on data side but is biased on the decision tree learning algorithm chosen. CCRDR enables the reduction of both bias and variance. However, on algorithms side, the chosen algorithms do not collaborate with each other and thus the reduction of bias is not sufficient. This could be explained by the assumption that each algorithm may generate a rule set that has some rules of high quality but the others of low quality. In other words, it cannot ensure that each single rule is generated to have a high quality and thus may result in incorrect classifications by low quality rules.

On the basis of above discussion, the hybrid ensemble rule based classification framework is strongly motivated due to its flexibility in employing rule learning algorithms and rule quality measures, as well as its involvement that different rule learning algorithms collaborate to complement each other.

5. Conclusions

In this paper, the recent advancement of ensemble learning is outlined and some popular approaches are reviewed critically. This paper also introduces a new framework for ensemble learning, which is referred to as hybrid ensemble rule based classification and involves scaling up algorithms and scaling down data for reduction of bias and variance respectively. The new framework is also validated empirically by comparing its performance with Random Forests and CCRDR. The results indicate that the new framework is helpful to improve overall classification accuracy through reduction of both bias and variance. As this newly developed framework achieves a high flexibility in employing rule learning algorithms and rule quality measures, this framework will be investigated further with respect to the employment of such algorithms and measures. In this way, any bias originating from chosen algorithms or measures will be avoided to a larger extent and thus accuracy will be improved further.

Acknowledgements

This paper relates to the first author's PhD research which is funded by the university specified in the authors' affiliation.

References

[1] J. R. Quinlan, C4.5: programs for machine learning. Morgan Kaufman, San Francisco, CA, USA, 1993.

[2] J. Furnkranz, Separate-and-Conquer rule learning, *Artificial Intelligence Review*, Vol. 13, pp. 3-54, 1999.

[3] P. N. Tang, M. Steinbach and V. Kumar, Introduction to Data Mining. New Jersey: Pearson Education, 2006.

[4] I. Kononenko and M. Kukar, Machine Learning and Data Mining. Chichester, West Sussex: Horwood Publishing Limmited, 2007.

[5] D. Brain, Learning From Large Data: Bias, Variance, Sampling, and Learning Curves. PhD Thesis, Deakin University, 2003.

[6] L. Breiman, Bagging Predictors, *Machine Learning*, Vol. 2, No. 24, pp.123-140, 1996.

[7] L. Breiman, Random Forests. *Machine Learning*, Vol. 45, No. 1, pp. 5-32, 2001.

[8] J. Li and L. Wong, Rule-Based Data Mining Methods for Classification Problems in Biomedical Domains. In: *15th European Conference on Machine Learning and 8th European Conference on Principles and Practice for of Knowledge Discovery in Databases*, Pisa, 2004.

[9] H. Liu and A. Gegov, Collaborative Decision Making by Ensemble Rule Based Classification Systems. In: W. Pedrycz and S. M. Chen, Granular Computing and Decision Making, *Studies in Big Data*, Vol. 10, pp. 245-264, Springer, 2015.

[10] P. Smyth and R. M. Goodman, An Information Theoretic Approach to Rule Induction from Databases. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 4, pp. 301-316, 1992.

[11] P. N. Tan, V. Kumar and J. Srivastava, Selecting the right objective measure for association analysis. *Information Systems*, Vol. 29, pp. 293-313, 2004.

[12] L. Geng and H. J. Hamilton, Interestingness measures for data mining: A survey. *ACM Computing Surveys*, Vol. 38, No. 3, 2006.

[13] M. Lichman, UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, 2013.

[14] J. Cendrowska, PRISM: an algorithm for inducing modular rules. *International Journal of Man-Machine Studies*, Vol. 27, pp. 349-370, 1987.

[15] H. Liu, A. Gegov and F. Stahl, Unified framework for construction of rule based classification systems. In: W. Pedrycz and S. M. Chen, Information Granularity, Big Data and Computational Intelligence, *Studies in Big Data*, Vol. 8, pp. 209-230, Springer, 2015.