Collaborative Rule Generation: An Ensemble Learning Approach

Han Liu, Alexander Gegov and Mihaela Cocea
School of Computing, University of Portsmouth, Buckingham Building, Lion Terrace, Portsmouth, PO1 3HE, United Kingdom.

Abstract. Due to the vast and rapid increase in data, data mining has become an increasingly important tool for the purpose of knowledge discovery in order to prevent the presence of rich data but poor knowledge. Data mining tasks can be undertaken in two ways, namely, manual walkthrough of data and use of machine learning approaches. Due to the presence of big data, machine learning has thus become a powerful tool to do data mining in intelligent ways. A popular approach of machine learning is inductive learning, which can be used to generate a rule set (a set of rules) using a particular algorithm. Inductive learning can involve a single base algorithm learning from a single data set following a standard learning approach. In this approach, the learning algorithm can generate a single rule set such as decision trees. On the other hand, the inductive learning can also involve a single base algorithm learning from multiple data sets following an ensemble learning approach. In this approach, the learning algorithm can generate multiple rule sets such as random forests. The latter approach is usually designed to reduce overfitting of models that usually arises when the former approach is adopted. In this context, the ensemble learning approach usually enables the improvement of the overall accuracy in prediction. The aim of this paper is to introduce a new approach of ensemble learning called Collaborative Rule Generation. In the new approach, the inductive learning involves multiple base algorithms learning from a single data set to generate a single rule set, which aims to enable each rule to have a higher quality. This paper also includes an experimental study validating the Collaborative Rule Generation approach and discusses the results in both quantitative and qualitative ways.

Keywords: Data Mining, Machine Learning, Ensemble Learning, Rule Based Systems, Rule Based Classification, If-Then Rules

1. Introduction

The daily increase in the size of data has motivated people to discover knowledge from databases [1] in order to prevent the presence of rich data but poor knowledge [2]. In other words, there would be potentially a large amount of knowledge that could be extracted from data. In this context, data mining has been seen as an important tool for knowledge discovery [3]. Data mining can be done by subject experts through manual analysis of data or by using machine learning approaches through empirical analysis. Due to the presence of big data, it is necessary to employ data mining in more intelligent ways. From this point of view, machine learning has become a more popular approach that can support intelligent data mining for knowledge discovery. On the other hand, machine learning can also be used for predictive modelling. In particular, for predictive modelling, the learning approaches are used in a black box manner and emphasize only the mapping from inputs to outputs. In contrast, for knowledge discovery, machine learning approaches are used in a white box manner and allows the interpretation of the reasons for the mapping between inputs and outputs.

Inductive learning is a popular approach of machine learning, which typically aims to generate a rule set that can be used as a rule based system for prediction making. Rule based systems can be used for both knowledge discovery and predictive modelling. For the purpose of knowledge discovery, Higgins justified that rule based knowledge representation is highly interpretable, allowing people to read and understand the knowledge extracted from models, due to the advantage that rules specify relationships between attributes, which can provide explanations with regard to a decision of an expert system [4]. Therefore, Higgins

motivated the use of rule based knowledge representation. For the purpose of predictive modelling, inductive learning is also seen as a popular approach [5]. As mentioned above, the inductive learning can involve the generation of a set of rules, which can be achieved following two rule generation approaches namely, 'divide and conquer' and 'separate and conquer'. The former approach generates rules in the form of decision trees and thus are also known as 'Top-Down Induction of Decision Trees' (TDIDT) [6]. The latter approach generates if-then rules directly from training instances and is also known as 'covering approach' [7]. Both approaches follow the way that only generates a single rule set by learning from a single data set. However, these two approaches are likely to generate such rule sets that overfit training data [3, 8].

As mentioned above, rule sets generated by rule based methods are likely to overfit training data. The development of ensemble learning approaches has thus been motivated to improve the overall accuracy in prediction. Ensemble learning can be done in parallel or sequentially. In the former way, there are no collaborations among different learning algorithms and only their predictions are combined together for final prediction making [8]. In this context, the final prediction is typically made by voting in classification and by averaging in regression. In the latter way of ensemble learning, the first algorithm learns a model from data and then the second algorithm learns to correct the former one and so on [8]. In other words, the model built by the first algorithm is further corrected by the following algorithms sequentially.

For rule based methods, ensemble learning can be adopted in the way that a base algorithm is used to learn from a number of samples, each of which results from the original training data through random sampling with replacement. In this case, there are *n* rule sets generated where *n* is the number of samples. In testing stage, each of the *n* rule sets makes an independent prediction on an unseen instance and their predictions are then combined to make the final prediction. The *n* rule sets mentioned above can make up an ensemble rule based system for prediction purpose as mentioned in the literature [9]. A typical example of such systems is Random Forest which consists of a number of decision trees [10] and is usually helpful for decision tree learning algorithms to generate more accurate rule sets [8]. However, this type of ensemble learning approaches that is applied to rule based learning algorithms does not involve collaboration in the training stage. In general, predictive accuracy can be improved in two ways namely, scaling up algorithms and scaling down data [11]. The former way is to reduce the bias originating from learning algorithms and the latter way is to reduce the variance originating from training data [8]. From this point of view, it is necessary for an ensemble learning approach to be designed to improve accuracy not only on the data side but also on the algorithms side. Therefore, the type of ensemble learning approach mentioned above such as Random Forests would not comprehensively improve the overall accuracy in prediction tasks. This paper introduces a new approach of ensemble learning for rule generation that is designed to improve accuracy on algorithms side.

The rest of this paper is organized as follows. Section 2 introduces the background of ensemble learning and some popular approaches such as Bagging, Boosting and Random Forests. Section 3 introduces a new ensemble learning framework, called Collaborative Rule Generation (CRG), and justifies how the CRG approach involves collaboration among different learning algorithms in the training stage. Section 4 describes the setup of an experimental study and presents the empirical results on the performance of CRG. Section 5 evaluates the CRG approach in both quantitative and qualitative ways and analyses its advantages and disadvantages. Section 6 summarises the contribution of this paper and specifies future work towards further improvements in this research area.

2. Related Work

Section 1 introduced the background on rule based learning methods and ensemble learning approaches. A gap that exists in ensemble rule based methods such as Random Forests has been mentioned briefly. This section introduces in more depth the ensemble learning with respects to concepts, methods and evaluation. This section also justifies in what way the ensemble learning approaches can be improved comprehensively towards the fulfilment of high level of accuracy in prediction tasks.

2.1. Ensemble Learning Concepts

Ensemble learning is usually adopted to improve overall accuracy. In detail, this purpose can be achieved through scaling up algorithms or scaling down data. As mentioned in Section 1, ensemble learning can be done both in parallel and sequentially. The parallel ensemble learning approach can be achieved by combining different learning algorithms,

each of which generates a model independently on the same training set. In this way, the predictions of the models generated by these algorithms are combined to predict unseen instances. This way belongs to scaling up algorithms because different algorithms are combined in order to generate a stronger hypothesis. In addition, the parallel ensemble learning approach can also be achieved by using a single base learning algorithm to generate models independently on different sample sets of training instances. In this context, the sample set of training instances can be provided by horizontally selecting the instances with replacement or vertically selecting the attributes without replacement. This way belongs to scaling down data because the training data is preprocessed to reduce the variance that exists on the basis of the attribute-values.

In sequential ensemble learning approach, accuracy can also be improved through scaling up algorithms or scaling down data. In the former way, different algorithms are combined in the way that the first algorithm learns to generate a model and then the second algorithm learns to correct the model and so on. In this way, there is nothing changed to training data. In the latter way, in contrast, the same algorithm is used iteratively on different versions of the training data. In each iteration, there is a model generated and the model is then evaluated using the validation data. According to the estimated quality of the model, the training instances are weighted to different extents and then used for the next iteration. In the testing stage, these models generated at different iterations make predictions independently and their predictions are then combined to predict unseen instances.

For both parallel and sequential ensemble learning approaches, voting is involved in the testing stage when the independent predictions are combined to make the final prediction on an unseen instance. Some popular methods of voting include equal voting, weighted voting and naïve Bayesian voting [8]. The following subsections will introduce in more detail the popular methods of ensemble learning and voting which are applied with the combination of decision tree based learning algorithms.

2.2. Bagging, Boosting and Random Forests

The term Bagging stands for bootstrap aggregating. It is a popular method developed by Breiman [12] 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, 8, 12]. 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, the combination of their independent predictions is made to predict the final classification based on equal voting. As concluded in the literatures [3, 8], Bagging is robust and does not lead to overfitting due to the increase of the number of generated models, and thus is useful especially for those non-stable learning methods with high variance such as neural networks, decision trees and rule based methods.

The term Boosting stands for Adaboost. It is also a popular method [13] and follows the sequential ensemble learning approach. In other words, the generation of a single classifier depends on the experience gained from its former classifier [14]. Each single classifier is assigned a weight depending on its accuracy measured by using validation data. The stopping criteria are satisfied while the error is equal to 0 or greater than 0.5 [14]. In the testing stage, each single classifier makes an independent prediction as similar to Bagging but the final prediction is made based on weighted voting among these independent predictions. As concluded in the literature [8], Boosting frequently outperforms Bagging, and can also be applied with those stable learning algorithms with low variance in addition to unstable ones, in contrast to Bagging. However, Boosting may generate an ensemble learner that overfits training data. In this case, the performance of the ensemble learner is worse than that of a single learner.

Random forests is another popular method [10] that is similar to Bagging but the difference is that decision trees must be the base classifiers generated and the attribute selection at each node is random. 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 [14]. As mentioned in Section 1, random forests have decision tree learning algorithms as the bases. 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 classifers make the independent predictions that are combined to make the final prediction based on equal voting. As concluded in the literature [8], the random forests algorithm is robust because of the

reduction of the variance for decision tree learning algorithms. However, the random forests algorithm makes it difficult to interpret the combined predictions, especially when the number of decision trees generated is more than 100, and thus leads to the incomprehensability of the predictions made by the decision trees. The same problem also happens with Bagging and Boosting.

2.3. Collaborative and Competitive Random Decision Rules (CCRDR)

The CCRDR approach of ensemble learning has been developed in [9], in order to fill the gap that exists in Random Forests and other similar methods as mentioned in Section 2.2. The basic idea of this approach is illustrated in Fig 1.

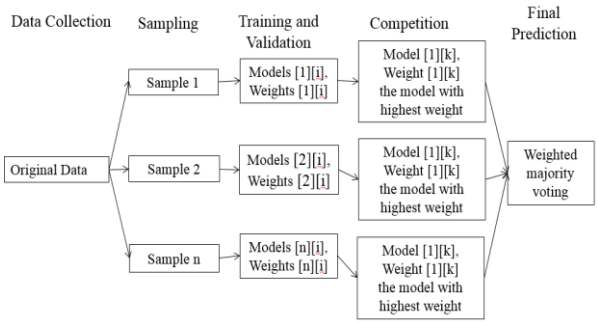

Fig.1. Procedures of CCRDR Ensemble Learning [9]

CCRDR stands for Collaborative and Competitive Random Decision Rules, which indicates that the ensemble learning framework includes both cooperative learning and competitive learning. Therefore, the above approach is designed partially to overcome the limitation that there is only a single base algorithm involved in the training stage, which cannot always generate robust rule sets 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 rule sets 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 rule sets on each sample. In this context, it becomes achievable to find better rule sets to be involved in the testing stage and to eliminate the worse ones through competition among these rule sets. The competition is based upon the weight (confidence) of each of the rule sets by means of overall accuracy estimated using validation data. In the CCRDR framework, only the rule set 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 data the hypothesis constructed becomes much stronger.

On the other hand, the CCRDR framework is also designed to address the issues relating to voting. As mentioned in Section 2.1, voting can be based on different criteria such as equal voting and weighted voting. From the three popular methods introduced in Section 2.2, Bagging and Random Forests are based on equal voting whereas Boosting is based on weighted voting. In classification tasks, the weighted voting is preferred to the equal voting. This is due to 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). Liu & Gegov also discusses several possible ways [9] for determining the weight for weighted voting. These ways are based on overall accuracy, precision and recall respectively, all of which are popularly used for evaluating the quality of a classifier. These ways are also compared experimentally in terms of their effectiveness for evaluating the reliability of a classifier in ensemble learning tasks. The experimental results indicate that precision is more effective than the other two. In addition, the following justifications also strongly support the above indication from the experimental results:

- The strategy based on overall accuracy is not reliable enough in determining the weight, especially for unbalanced data sets. This is due to the high possibility that a classifier performs better on predicting positive instances but worse on negative instances if it is a two class classification task. The similar cases can also happen in multi-class classification tasks. Therefore, it is necessary to adopt the individual accuracy (recall) for a single classification as the weight [9].
- Precision would be more reliable than recall in determining the weight of a classifier. 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%. This is because the classifier predicts 10 instances as positive and only five of them are correct. This case indicates the possibility that high recall could result from coincidence due to low frequency of a particular classification. Therefore, precision is more reliable in determining the weight of a classifier on a particular prediction from this point of view [9].

The CCRDR framework still has limitations. One of them is due to the absence of collaboration among different learning algorithms in the training stage. Therefore, this framework only follows the parallel ensemble learning approach and has no connection to sequential ensemble learning. Section 3 introduces a new ensemble learning approach that involves such collaboration in the training stage to overcome the above limitation that arises with the CCRDR framework.

3. Collaborative Rule Generation Approach

As mentioned in Section 2, the Random Forests and other similar ensemble rule based methods focus on parallel learning, which means that the building of each rule set is totally parallel to the others without collaborations in the training stage and only their predictions in the testing stage are combined for the final classification. However, the ensemble learning can also be achieved with collaborations in the training stage. In this way, it would potentially help to improve the quality of each single rule generated in the training stage. Therefore, this section introduces a new framework of ensemble learning with collaboration among different rule based methods involved in the training stage. The procedure is illustrated by Fig.2.

Iteration 1:
three rules generated: $rule_{11}$, $rule_{12}$ and $rule_{13}$
the one with the highest quality: $rule_{11}$
The one added into the rule set: $rule_{11}$
Iteration 2:
three rules generated: $rule_{21}$, $rule_{22}$ and $rule_{23}$
the one with the highest quality: $rule_{22}$
The one added into the rule set: $rule_{22}$
…….
…….
Finally, the generated rule set comprises: {$rule_{11}$, $rule_{22}$ ……..}.

Fig.2. Collaborative rule generation procedure

3.1. Key Features

The essence of this approach above is based on the procedure of the 'separate and conquer' rule generation. In particular, there is a single rule generated in each of the iterations. The approach introduced above has all chosen rule based methods involved in the iteration to generate a rule; each of the rule based methods may also be assisted by some pruning methods depending on the setup of experiments; in the next step, all of the rules are compared with respect to their quality; finally, only the rule with the highest quality is selected and added into the rule set. This process is repeated until all of instances have been deleted from the training set as specified in the 'separate and conquer' approach. This way of rule generation ensures that in each of the iterations the rule generated has a quality as high as possible. This is because of the possibility that some of the rules are of higher quality but the others are of lower quality if there is only one rule based method involved in the training stage. The main difference to the CCRDR introduced in Section 2 is with respect to competition between rule based methods. CCRDR involves a competition between rule based methods per rule set. In other words, the competition is made after each of chosen methods has generated a rule set, in order to compare the quality of a whole rule set. In contrast, the CRG approach involves such a competition per rule generated. In other words, the competition is made once each of the methods has generated a rule in order to compare the quality of a single rule.

3.2. Justification

As mentioned in Section 2, Bagging, Boosting and Random Forests all have the disadvantage of incomprehensibility of the predictions made by different models. The same disadvantage also arises with the CCRDR approach introduced in Section 2.3. This is a serious drawback that arises with most existing ensemble learning approaches for data mining tasks. As mentioned in Section 1, data mining is aimed at knowledge discovery. Therefore, it is necessary for the models to allow explicit interpretation of the way the prediction is made. The CRG approach would be able to fill the gap to some extent as it only generates a single rule set that is used for prediction. In addition, rule based models are highly interpretable as mentioned in Section 1; consequently, the CRG approach would fit well the

purpose of knowledge discovery especially on interpretability.

With regard to accuracy, Bagging, Boosting and Random Forests all aim to improve it on the data side (scaling down data). However, there is nothing done on the algorithms side (scaling up algorithms) for improving accuracy. As mentioned in Section 1, it is necessary to deal with the issues on both algorithms and data sides in order to comprehensively improve the accuracy. The CCRDR can fulfil the need to a large extent. As justified in [9], each algorithm may have a different level of suitability to different data sets. On the same data set, different algorithms may also demonstrate different levels of performance. From this point of view, the CCRDR approach is designed in the way that after the training data is scaled down by drawing different samples, a group of learning algorithms are combined to generate a model on each of the samples. In this context, the CCRDR approach does not only scale down the data but also scale up the algorithms. However, as mentioned in Section 2.3, this approach does not involve any collaborations among different algorithms in the training stage. For rule based learning algorithms, it is very likely to generate a rule set that has some rules of a high quality but also others of a low quality. In other words, it is difficult to guarantee that every single rule generated by a particular algorithm is of high quality. In this sense, the CCRDR approach is only able to select the rule sets, each of which is generated on a particular sample set and has the highest quality on average compared with the others generated on the same sample set. In the testing stage, a rule set usually makes a prediction using a single rule that fires. If the single rule is of a low quality, it is very likely to make an incorrect prediction although most of the other rules are of high quality. On the other hand, for data mining tasks, each of the rules is used to provide knowledge insight for domain experts. Therefore, the reliability of each single rule is particularly significant. On the basis of the above description, the CRG approach would be useful and effective to help the CCRDR fill the gap relating to the quality of each single rule and thus also complements the other three popular methods mentioned in this section.

4. Experimental Setup and Results

This section introduces the empirical validation of the CRG approach described in Section 3. The setup of the experimental study is specified in detail and the results are presented.

The aim of this experimental study is to validate that the combination of different rule based learning algorithms usually improves the overall accuracy compared with the use of a single base algorithm. On the other hand, this experimental study also includes the validation on the quality of rule sets generated. This is in order to show that the combination of different learning algorithms usually improves the quality of each single rule on average in comparison with the use of a single base algorithm.

In this experimental study, there are 20 data sets (see Table1) chosen from the UCI repository [15] for the validation of the approach mentioned above. The selection of these data sets is in accordance with the dimensionality and number of instances under the consideration of computational constraints. In general, ensemble learning methods are computationally more expensive than those base learning algorithms. On the basis of the above consideration, all of the chosen data sets are relatively small. It can be seen from the Table 1 that the dimensionality is lower than 100 and the number of instances is less than 10,000 for all of the data sets. In addition, the chosen data sets cover both discrete and continuous attributes in order to validate that the proposed approach of ensemble learning can effectively deal with both types of attributes.

On algorithms side, two learning algorithms namely, Prism [16] and Information Entropy Based Rule Generation (IEBRG) [17] are chosen to be the base algorithms for the CRG framework. In general, this framework can include any algorithms, which follow the separate and conquer rule learning approach, to be combined for the generation of a rule set. In this experimental study, there are only two algorithms chosen 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 learning algorithms, if no parallelisation is adopted. The reason that the Prism algorithm is chosen is due to the advantage that this algorithm can typlically overcome some limitations of decision tree based learning algorithms to a large extent, such as the replicated subtree problem [16]. The IEBRG algorithm is also chosen because it complements the Prism algorithm with regard to some of its disadvantages. In fact, the aim of the CRG approach is to enable that combined algorithms complement each other. In other words, the disadvantages of one algorithm could be overcome by the advantages of another algorithm. Therefore, it

would be appropriate to choose algorithms that have different advantages and disadvantages and that are complementary to each other.

On the other hand, as mentioned in Section 3, the CRG approach involves measuring the quality of each single rule generated. In this context, the approach needs to employ at least one of the measures of rule quality to judge which one of the generated rules is of the highest quality. In this experimental study, the four measures, namely confidence, J-measure, lift and leverage, are chosen due to their significance and popularity in real applications [18].

Table.1. Data Sets

Name	Attribute Types	#Attributes	#Instances	#Classes
anneal	discrete, continuous	38	798	6
credit-g	discrete, continuous	20	1000	2
diabetes	discrete, continuous	20	768	2
heart-stalog	continuous	13	270	2
ionosphere	continuous	34	351	2
iris	continuous	4	150	3
kr-vs-kp	discrete	36	3196	2
lymph	discrete, continuous	19	148	4
segment	continuous	19	2310	7
zoo	discrete, continuous	18	101	7
wine	continuous	13	178	3
breast-cancer	discrete	9	286	2
car	discrete	6	1728	4
breast-w	continuous	10	699	2
credit-a	discrete, continuous	15	690	2
heart-c	discrete, continuous	76	920	4
heart-h	discrete, continuous	76	920	4
hepatitis	discrete, continuous	20	155	2
mushroom	discrete	22	8124	2
vote	discrete	16	435	2

Table 2 provides more details about these measures. In this table, the notation *P(x)* and *P(y)* are read as the coverages of rule antecendant (left hand side) and rule consequent (right hand side) respectively. In addition, the notation *P(x, y)* is read as the coverage of a single rule (for both antecendant and consequent of the rule). Justification of these objective measures is out of the scope of this paper. A more detailed overview of measures of rule quality (including the four ones listed in the Table 2) can be found in [18, 23]. Under the above setup, for the measure of classification accuracy, 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 five times and the mean and standard deviation of the accuracies are calculated for comparative validation. As mentioned earlier, ensemble learning approaches are usually computationally more expensive. Therefore, cross validation is not adopted in this study.

Table.2. Measures of rule quality

Name	Formula	References
confidence	$Conf = \frac{P(x,y)}{P(x)}$	[19]
J-measure	$J(Y, X = x) = P(x) \cdot j(Y, X = x)$ $j(Y, X = x) = P(y \mid x) \cdot \log(\frac{P(y \mid x)}{P(y)}) + (1 - P(y \mid x)) \cdot \log(\frac{1 - P(y \mid x)}{1 - P(y)})$	[20]
lift	$Lift = \frac{P(x,y)}{P(x) \cdot P(y)}$	[21]
leverage	$Leverage = P(x,y) - P(x) \cdot P(y)$	[22]

Table.3. Accuracy

Dataset	Prism	IEBRG	CRG (Conf)	CRG (J-measure)	CRG (Lift)	CRG (Leverage)
anneal	0.68±0.070	0.90±0.022	0.90±0.025	0.90±0.022	**0.91±0.023**	**0.91±0.023**
credit-g	0.69±0.022	0.67±0.038	**0.72±0.011**	**0.72±0.011**	0.71±0.011	0.72±0.022
diabetes	0.71±0.018	0.70±0.043	0.73 ±0.032	0.73±0.013	0.72±0.023	**0.75±0.019**
heartStatlog	0.64±0.051	0.66±0.056	0.71±0.026	0.74±0.036	**0.75±0.026**	0.74±0.029
ionosphere	0.87±0.024	0.81±0.029	0.84±0.023	0.86±0.043	0.86±0.011	**0.88±0.027**
iris	0.72±0.160	0.93±0.029	0.92±0.073	0.94±0.027	0.95±0.026	**0.95±0.017**
kr-vs-kp	0.77±0.111	0.83±0.074	**0.95±0.033**	0.93±0.013	0.92±0.019	0.92±0.008
lymph	0.68±0.067	0.70±0.069	0.75±0.038	0.75±0.019	0.75±0.033	**0.78±0.050**
segment	0.55±0.091	0.68±0.061	0.80±0.031	0.80±0.059	**0.81±0.046**	0.77±0.035
zoo	0.62±0.070	0.79±0.057	0.87±0.038	0.85±0.033	0.84±0.031	**0.88±0.071**
wine	0.80±0.072	0.91±0.021	0.92±0.019	0.94±0.013	0.94±0.018	**0.95±0.018**
breastCancer	0.67±0.047	0.69±0.043	0.71±0.018	0.71±0.013	**0.73±0.033**	0.72±0.033
car	0.71±0.012	0.76±0.018	0.76±0.030	0.75±0.012	**0.77±0.017**	0.76±0.015
breast-w	0.90±0.046	**0.95±0.020**	0.92±0.017	0.94±0.011	0.95±0.021	0.93±0.011
credit-a	0.69±0.061	0.69±0.091	0.77±0.021	0.79±0.041	0.80±0.057	**0.80±0.034**
heart-c	0.70±0.057	0.69±0.048	0.76±0.023	0.74±0.020	**0.76±0.029**	0.75±0.017
heart-h	0.76±0.036	0.74±0.033	0.83±0.038	0.81±0.023	0.82±0.016	**0.83±0.032**
hepatitis	0.82±0.052	0.82±0.061	0.83±0.015	0.85±0.026	0.84±0.023	**0.86±0.045**
mushroom	0.97±0.008	0.98±0.005	0.97±0.011	0.97±0.011	0.97±0.013	**0.98±0.013**
vote	0.88±0.008	0.90±0.051	0.93±0.013	**0.96±0.015**	0.95±0.021	0.95±0.018

On the other hand, the whole data set is used for training and then the same data set is used to evaluate each single rule generated with regard to the quality of the rule. The experimental results are presented in Tables 3 to 7. Table 3 shows the comparison among CRG with different measures of rule quality, Prism and IEBRG in terms of classification accuracy. This part of validation is with regard to the performance of the CRG in machine learning tasks for the purpose of predictive modelling. On the other hand, Tables 4 to 7 show on average the quality of each single rule generated by using different learning approaches. This part of validation is with regard to the performance of the CRG in data mining tasks for the purpose of knowledge discovery. All of these empirical results will be discussed in more detail in Section 5 both quantitatively and qualitatively.

5. Evaluation

As mentioned in Section 4, the validation is divided into two parts. The first part is to measure the classification accuracy for the CRG approach. This is in order to estimate the performance of this approach in machine learning tasks. The second part of the validation is to measure the quality of each single rule generated. This is in order to estimate the reliability of each single rule for knowledge usage. This section analyses the results presented in Section 4 to evaluate the performance of the CRG approach for both machine learning and data mining purposes.

With regard to classification accuracy, Table 3 shows that the CRG approach, which has different variants, outperforms both Prism and IEBRG in all cases, except for the case of the *breast-w* data set, on which the above approach with lift performs the same as Prism on average but a bit worse on standard deviation. This indicates that the combination of different learning algorithms usually improves the overall accuracy as expected. In some cases (on *ks-vs-kp*, *segment* and *credit-a* data sets), the CRG approach even outperforms both of the two base algorithms to a large extent. This phenomenon can support the argument that two algorithms can be complementary to each other, especially on the basis that they have different advantages and disadvantages and that they are combined in an effective way. On the other hand, the results show that this approach has a bias on the chosen measure of rule

quality. It can be seen from Table 3 on the data sets, *anneal*, *ionosphere*, *iris*, *car*, *breast-w* and *mushroom*, that at least one of the measures of rule quality fails to help outperform both of the two base learning algorithms namely, Prism and IEBRG. This phenomenon is also due partially to the variance on data side but it is still critical to appropriately choose the measure of rule quality to reduce the bias on the algorithms side.

Table.4. Rule quality by confidence

Dataset	Prism	IEBRG	CRG
anneal	0.75±0.009	0.78 ± 0.053	**0.91 ± 0.009**
credit-g	0.78±0.029	0.78 ± 0.028	**0.79 ± 0.020**
diabetes	0.69±0.075	0.81 ± 0.318	**0.82 ± 0.026**
heart-statlog	0.79±0.018	0.79 ± 0.022	**0.84 ± 0.017**
ionosphere	0.76±0.055	0.89 ± 0.013	**0.92 ± 0.008**
iris	0.70±0.092	**0.70 ± 0.076**	**0.70 ± 0.076**
kr-vs-kp	0.42±0.419	0.69 ± 0.023	**0.77 ± 0.027**
lymph	0.68±0.092	0.83 ± 0.017	**0.87 ± 0.025**
segment	0.36±0.058	0.57 ± 0.103	**0.69 ± 0.121**
zoo	0.78±0.110	0.80 ± 0.085	**0.84 ± 0.099**
wine	0.81±0.031	0.73 ± 0.080	**0.82 ±0.059**
breast-cancer	0.68±0.074	0.72 ± 0.071	**0.76 ± 0.028**
car	0.59±0.147	0.20 ± 0.127	**0.72 ± 0.147**
breast-w	0.83±0.088	0.83 ± 0.109	**0.94 ± 0.007**
credit-a	0.75±0.015	**0.83 ± 0.020**	**0.83 ± 0.020**
heart-c	0.68±0.024	0.78 ± 0.017	**0.80 ± 0.022**
heart-h	0.64±0.084	0.82 ± 0.019	**0.88 ± 0.012**
hepatitis	**0.81±0.097**	**0.81 ± 0.097**	**0.81 ± 0.097**
mushroom	0.90±0.033	0.77 ± 0.094	**0.95 ± 0.023**
vote	**0.94±0.024**	0.84 ± 0.068	0.92 ± 0.041

Table.5. Rule quality by J-measure

Dataset	Prism	IEBRG	CRG
anneal	0.019±0.007	0.055 ± 0.007	**0.138 ± 0.021**
credit-g	0.011±2.92E-4	0.011 ± 2.92E-4	**0.071 ± 0.032**
diabetes	0.031±7.21E-4	0.034 ± 4.50E-4	**0.040 ± 5.20E-4**
heart-statlog	0.034±9.33E-4	0.041 ± 0.001	**0.073 ± 0.004**
ionosphere	0.059±0.009	**0.165 ± 0.014**	**0.165 ± 0.014**
iris	0.216±0.029	0.242 ± 0.050	**0.415 ± 0.026**
kr-vs-kp	0.017±0.148	0.039 ± 0.003	**0.091 ± 0.013**
lymph	0.042±0.001	**0.088 ± 0.002**	**0.168 ± 0.056**
segment	0.039±0.013	0.144 ± 0.225	**0.150 ± 0.038**
zoo	0.269±0.041	0.279 ± 0.036	**0.295 ± 0.041**
wine	0.254±0.021	0.225 ± 0.031	**0.266 ± 0.021**
breast-cancer	0.007±1.16E-4	0.013 ± 2.20E-4	**0.016 ± 2.05E-4**
car	0.015±0.002	0.016 ± 0.002	**0.018 ± 0.002**
breast-w	0.137±0.003	**0.260 ± 0.003**	**0.260 ± 0.003**
credit-a	0.039±0.004	0.081 ± 0.006	**0.090 ± 0.007**
heart-c	0.015±4.44E-4	0.049 ± 9.41E-4	**0.053 ± 0.001**
heart-h	0.013±5.01E-5	0.045 ± 7.72E-4	**0.051 ± 0.003**
hepatitis	0.046±5.39E-4	0.046 ± 5.39E-4	**0.057 ± 0.002**
mushroom	0.071±0.005	0.104 ± 0.008	**0.426 ± 0.147**
vote	0.133±0.010	0.126 ± 0.013	**0.153 ± 0.014**

Table.6. Rule quality by lift

Dataset	Prism	IEBRG	CRG
anneal	1.31 ± 0.072	3.41 ± 18.71	**14.56 ± 276.44**
credit-g	0.84 ± 0.206	**1.23 ± 0.037**	1.21± 0.080
diabetes	1.12 ± 0.132	**1.64 ± 0.388**	**1.64 ± 0.388**
heart-statlog	0.83 ± 0.052	**1.55 ± 0.100**	**1.55 ± 0.100**
ionosphere	1.42 ± 0.372	**1.99 ± 0.626**	**1.99 ± 0.626**
iris	1.26 ± 0.115	**2.11 ± 0.686**	**2.11 ± 0.686**
kr-vs-kp	0.98 ± 0.855	**1.40 ± 0.099**	**1.40 ± 0.099**
lymph	3.79 ± 43.326	1.80 ± 0.227	**10.22± 182.25**
segment	2.50 ± 2.833	**4.00 ± 5.056**	3.79 ± 5.458
zoo	8.03 ± 76.830	**8.25 ± 74.17**	7.35 ± 17.00
wine	**2.41 ± 0.480**	2.20 ± 0.818	**2.41 ± 0.480**
breast-cancer	1.13 ± 0.042	1.21 ± 0.138	**1.29 ± 0.131**
car	1.56 ± 1.254	0.58 ± 0.685	**2.09 ± 1.277**
breast-w	1.63 ± 0.677	1.94 ± 0.995	**2.24 ± 0.447**
credit-a	1.44 ± 0.057	**1.77 ± 0.144**	**1.77 ± 0.144**
heart-c	1.35 ± 0.074	**1.64 ± 0.097**	**1.64 ± 0.097**
heart-h	1.05 ± 0.165	**1.75 ± 0.257**	**1.75 ± 0.257**
hepatitis	1.17 ± 0.014	1.17 ± 0.014	**1.67 ± 0.355**
mushroom	**1.78 ± 0.125**	1.53 ± 0.390	1.53 ± 0.390
vote	**1.97 ± 0.214**	1.52 ± 0.180	**1.97 ± 0.214**

Table.7. Rule quality by leverage

Dataset	Prism	IEBRG	CRG
anneal	0.021±2.74E-4	0.017±0.001	**0.060±9.82E-4**
credit-g	0.003±5.62E-5	**0.017±5.55E-4**	**0.017±5.55E-4**
diabetes	0.007±0.003	0.028±0.001	**0.048±6.87E-4**
heart-statlog	0.038±0.002	0.045±0.002	**0.048±9.66E-4**
ionosphere	0.032±0.002	**0.115±5.50E-5**	**0.115±5.50E-5**
iris	0.107±0.006	**0.136±0.006**	**0.136±0.006**
kr-vs-kp	0.024±9.31E-4	0.040±0.001	**0.041±0.001**
lymph	0.037±0.001	0.076±0.002	**0.077±0.002**
segment	0.018±0.001	0.063±0.002	**0.072±0.001**
zoo	0.110±0.007	0.117±0.006	**0.149±0.008**
wine	**0.147±0.002**	0.118±0.006	**0.147±0.002**
Breast-cancer	0.004±1.25E-4	**0.019±3.29E-4**	**0.019±3.29E-4**
car	0.009±7.15E-4	0.010±7.85E-4	**0.012±7.99E-4**
breast-w	0.053±0.009	0.083±0.016	**0.132±0.002**
credit-a	0.045±0.004	**0.065±0.004**	**0.065±0.004**
heart-c	0.024±7.53E-4	**0.064±0.002**	**0.064±0.002**
heart-h	0.004±3.03E-4	**0.043±4.23E-4**	**0.043±4.23E-4**
hepatitis	0.041±9.23E-4	0.041±9.23E-4	**0.043±4.95E-4**
mushroom	0.046±0.001	0.022±0.009	**0.061±0.004**
vote	0.079±0.003	**0.085±0.004**	**0.085±0.004**

With regard to rule quality, Tables 4 to 7 show that in most or even all of the cases, the CRG approach can generate a rule set that has a higher quality in comparison with both Prism and IEBRG algorithms. Even if the CRG approach fails to outperform both of the two base algorithms in a few cases, the results originating from the above approach is still very close to the best ones. This phenomenon indicates that the proposed approach of ensemble learning usually improves the quality of each single rule generated on average. Table 4 shows that the CRG approach achieves a significant improvement when comparing with the two base algorithms in all cases except for the case of the *vote* data set, on which the result is a bit worse than the best one from Prism. This is parially due to the disadvgantage that confidence may result in misleading for estimation of rule quality in some cases,

especially when the coverage of the rule consequent is higher than the rule confidence [18, 21]. J-measure is seen as the most appropriate measure for evaluating the quality of a single rule [8]. The argument to some extent can be supported by the results shown in Table 5 that in 17 out of 20 cases the CRG approach outperforms the two base algorithms in terms of rule quality. In addition, in the other three cases, the CRG still performs the same as one of the two base algorithms, whichever performs beter. Table 6 and Table 7 show that the CRG performs the best in 16 out of 20 cases and all cases respectively. It can be seen from both tables that there are about 50% of the cases on which the CRG and one of the two base algorithms perform equally well. This could be explained by the assumption that in each iteration one base algorithm generates a higher quality rule than the other one and, as a result, the CRG approach generates a rule set identical to the one generated by one of the base algorithms that performs better. The results show that the use of both lift and leverage can improve the quality of each single rule in general. The leverage is equal to the division of rule confidence by the coverage of rule consequent, which would complement rule confidence with regard to its disadvantage as mentioned earlier in this section. In other words, it can avoid any misleading from the use of rule confidence when the coverage of the rule consequent is higher than the confidence. However, the leverage may suffer from the low frequency of training instances [21, 24]. This disadvantage can be overcome by the use of lift while the difference between the two terms: $P(x, y)$ and $P(x) \cdot P(y)$ (See Table 2), is replaced by the division of these two terms.

Overall, the empirical results shown in Tables 3 to 7 indicate that the CRG approach is useful for improving the quality of each single rule generated and thus improving the overall accuracy. In machine learning tasks, the main concern of a rule based learning algorithm is typically about using a rule set as a whole to accurately predict on unseen instances. In this context, some rules that are in low quality may be rarely or even never used for prediction. In this case, although the accuracy may not be seriously affected, the improvement for the quality of each single rule is still necessary towards the improvement of overall accuracy, especially when a large set of test instances are used. On the other hand, the rules generated in data mining tasks aim for knowledge usage. From this point of view, the main concern would be about the reliability of each single rule when the rule is used to provide insights for a knowledge domain. This even makes it necessary to a larger extent to improve the quality of each single rule. Besides, for separate and conquer rule learning, the generation of each single rule would affect that of all subsequent rules. In other words, the quality of each single rule generated would lead to a chained impact on the generation of all subsequent rules. Therefore, it is important to ensure that each single rule generated has as high a quality as possible. On the basis of above description, the CRG approach introduced in Section 3 deserves further attention and development, especially on the reduction of bias originating from algorithms, such as the choice of rule quality measures.

6. Conclusion

This paper introduced a new approach of ensemble learning, referred to as Collaborative Rule Generation (CRG). The CRG approach has been validated empirically and the experimental results show that the combination of different learning algorithms usually improves the classification accuracy and the quality of each single rule in comparison with the use of a single base algorithm. This paper also introduced the background of data mining and machine learning. In particular, the ensemble learning concepts were described and three popular methods namely, Bagging, Boosting and Random Forests, were reviewed. In accordance with the disadvantages of these methods, the necessity that the CCRDR approach had been developed was highlighted. However, as mentioned in Section 2.3, the CCRDR still had some limitations that made it insufficient to improve accuracy comprehensively. This motivated the development of the CRG approach as mentioned above. The CRG approach will be investigated further with respect to the way to effectively estimate the quality of each single rule towards the avoidance of any bias originating from the theoretical measures reviewed in the paper [18] and other literature. In addition, this approach will also be combined with the CCRDR framework to investigate to what extent the classification accuracy can be improved through both scaling up algorithms and scaling down data.

References

[1] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, "From data mining to knowledge discovery in databases," *AI Magazine, American Association for Artificial Intelligence,* pp. 37-54, 1996.

[2] F. Stahl and I. Jordanov, "An overview of use of neural networks for data mining tasks," *WIREs: Data Mining and Knowledge Discovery*, pp. 193-208, 2012.
[3] P.-N. Tan, M. Steinbach and V. Kumar, Introduction to Data Mining, New Jersey: Pearson Education, 2006.
[4] C. M. Higgins, "Classification and Approximation with Rule Based Networks," Pasadena, California, 1993.
[5] T. M. Mitchell, Machine Learning, McGraw Hill, 1997.
[6] R. Quinlan, C4.5: programs for machine learning, Morgan Kaufman, 1993.
[7] J. Furnkranz, "Separate-and-Conquer rule learning," *Artificial Intelligence Review*, vol. 13, pp. 3-54, 1999.
[8] I. Kononenko and M. Kukar, Machine Learning and Data Mining: Introduction to Principles and Algorithms, Chichester, West Sussex: Horwood Publishing Limmited, 2007.
[9] H. Liu and A. Gegov, "Collaborative Decision Making by Ensemble Rule Based Classification Systems," in *Granular Computing and Decision-Making: Interactive and Iterative Approaches*, vol. 10, W. Pedrycz and S. Chen, Eds., Springer, 2015, pp. 245-264.
[10] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.
[11] D. Brain, "Learning from Large Data, Bias, Variance, Sampling and Learning Curves," Geelong, Victoria, 2003.
[12] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, p. 123–140, 1996.
[13] Y. Freund and R. E. Schapire, "Experiments with a New Boosting Algorithm," in *Machine Learning: Proceedings of the Thirteenth International Conference (ICML '96)*, 1996.
[14] J. Li and L. Wong, "Rule-Based Data Mining Methods for Classification Problems in Biomedical Domains," in *A tutorial note for the 15th European Conference on Machine Learning (ECML) and the 8th European Conference on Principles and Practice for Knowledge Discovery in Databases (PKDD)*, Pisa, 2004.
[15] M. Lichman, "UCI Machine Learning Repository," University of California, School of Information and Computer Science, 2013. [Online]. Available: http://archive.ics.uci.edu/ml. [Accessed 12 May 2015].
[16] J. Cendrowska, "PRISM: an algorithm for inducing modular rules," *International Journal of Man-Machine Studies*, vol. 27, p. 349–370, 1987.
[17] H. Liu, A. Gegov and F. Stahl, "Unified Framework for Construction of Rule Based Classification Systems," in *Inforamtion Granularity, Big Data and Computational Intelligence*, vol. 8, W. Pedrycz and S. Chen, Eds., Springer, 2015, pp. 209-230.
[18] P.-N. Tan, V. Kumar and J. Srivastava, "Selecting the right objective measure for association analysis," *Information Systems*, vol. 29, p. 293–313, 2004.
[19] R. Agrawal, T. Imielinski and A. Swami, "Mining association rules between sets of items in large databases," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Washington D.C., 1993.
[20] P. Smyth and R. M. Goodman, "An Information Theoretic Approach to Rule Induction from Databases," *IEEE Transactions on Knowledge and Data Engineering*, vol. 4, no. 4, pp. 301-316, 1992.
[21] S. Brin, R. Motwani, J. D. Ullman and S. Tsur, "Dynamic itemset counting and implication rules for market basket data," in *Proceedings ACM SIGMOD International Conference on Management of Data*, Tucson, Arizona, USA, 1997.
[22] G. Piatetsky-Shapiro, "Discovery, analysis, and presentation of strong rules," in *Knowledge Discovery in Databases*, Cambridge, AAAI/MIT Press, 1991.
[23] L. Geng and H. J. Hamilton, "Interestingness measures for data mining: A survey," *ACM Computing Surveys*, vol. 38, no. 3, 2006.
[24] M. Hahsler, "A Probabilistic Comparison of Commonly Used Interest Measures for Association Rules," 13 February 2015. [Online]. Available: http://michael.hahsler.net/research/association_rules/measures.html. [Accessed 20 February 2015].