NEW FEATURE SELECTION MODEL-BASED ENSEMBLE RULE CLASSIFIERS METHOD FOR DATASET CLASSIFICATION

Mohammad Aizat bin Basir[1] and Faudziah binti Ahmad [2]

[1] Universiti Malaysia Terengganu (UMT)Terengganu, Malaysia
[2] Universiti Utara Malaysia(UUM) Kedah, Malaysia

ABSTRACT

Feature selection and classification task are an essential process in dealing with large data sets that comprise numerous number of input attributes. There are many search methods and classifiers that have been used to find the optimal number of attributes. The aim of this paper is to find the optimal set of attributes and improve the classification accuracy by adopting ensemble rule classifiers method. Research process involves 2 phases; finding the optimal set of attributes and ensemble classifiers method for classification task. Results are in terms of percentage of accuracy and number of selected attributes and rules generated. 6 datasets were used for the experiment. The final output is an optimal set of attributes with ensemble rule classifiers method. The experimental results conducted on public real dataset demonstrate that the ensemble rule classifiers methods consistently show improve classification accuracy on the selected dataset. Significant improvement in accuracy and optimal set of attribute selected is achieved by adopting ensemble rule classifiers method.

KEYWORDS

Feature Selection, Attribute, Ensemble, and Classification.

1. INTRODUCTION

Real world dataset usually consist a large number of attributes. It is very common some of those input attributes could be irrelevant and consequently give an impact to the design of a classification model. In situations where a rule has too many conditions, it becomes less interpretable. Based on this understanding, it becomes important to reduce the dimensionality (number of input attributes in the rule) of the rules in the rule set. In practical situations, it is recommended to remove the irrelevant and redundant dimensions for less processing time and labor cost. The amount of data is directly correlated with the number of samples collected and the number of attributes. A dataset with a large number of attributes is known as a dataset with high dimensionality [1]. The high dimensionality of datasets leads to the phenomenon known as the curse of dimensionality where computation time is an exponential function of the number of the dimensions. It is often the case that the model contains redundant rules and/or variables. When faced with difficulties resulting from the high dimension of a space, the ideal approach is to decrease this dimension, without losing the relevant information in the data. If there are a large number of rules and/or attributes in each rule, it becomes more and more vague for the user to understand and difficult to exercise and utilize. Rule redundancy and/or attribute complexity could overcome by reducing the number of attributes in a dataset and removing irrelevant or less significant roles. This can reduce the computation time, and storage space. Models with simpler and small number of rules are often easier to interpret.

DOI : 10.5121/ijaia.2017.8204

The main drawback of rule/attributes complexity reduction is the possibility of information loss. It is important to point out that two critical aspects of the attribute reduction problem are the degree of attribute optimality (in terms of subset size and corresponding dependency degree) and time required to achieve this attribute optimality. For example, existing methods such as Quick Reduct and Entropy-Based Reduction (EBR) methods find reduced in less time, but could not guarantee a minimal subset [1] –[3] whereas other hybrid methods which combine rough sets and swarm algorithm such as GenRSAR, AntRSAR, PSO-RSAR and BeeRSAR methods improve the performance but consume more time [1], [2].

In feature selection, also known as variable selection, attribute selection or variable subset selection is the process of selecting a subset of relevant features (attributes) for use in model construction. It is the process of choosing a subset of original features so that the feature space is optimally reduced to evaluation criterion. Feature selection can reduce both the data and the computational complexity. The raw data collected is usually large, so it is important to select a subset of data by creating feature vectors. Feature subset selection is the process of identifying and removing much of the redundant and irrelevant information possible.

However, the use of a subset of a feature set may disregard important information contained in other subsets. Consequently, classification performance is reduced. Therefore, this paper aims to find the optimal set of attributes and improve the classification accuracy by adopting the ensemble classifier method. Firstly, an optimal set of attribute subsets are extracted by applying various search method and a reduction algorithm to the original dataset. Then an optimal set of attributes further classified by adopting a classification ensemble approach. In the experiment, 6 various datasets were used. The experiment results showed that the performance of the ensemble classifier was improved the classification accuracy of the dataset. This paper is organized as follows: in Section II, related works are discussed. The proposed methodology is presented in Section III. In Section IV, the results and discussion are given. Finally, the conclusions presented in Section V.

2. RELATED WORKS

There many research in feature selection methods for constructing an ensemble of classifiers. The ensemble feature selection method is where a set of the classifiers, each of which solve the same original task, are joined in order to obtain a better combination global classifier, with more accurate and reliable estimates or decisions than can be obtained from using a single classifier. The aim of designing and using the ensemble method is to achieve a more accurate classification by combining many weak learners.

Previous studies show that methods like bagging improve generalization by decreasing variance. In contrast, methods similar to boosting achieve this by decreasing the bias [4]. [5] demonstrated a technique for building ensembles from simple Bayes classifiers in random feature subsets.
[6] explored tree based ensembles for feature selection. It uses the approximately optimal feature selection method and classifiers constructed with all variables from the TIED dataset.

[7] presented the genetic ensemble feature selection strategy, which uses a genetic search for an ensemble feature selection method. It starts with creating an initial population of classifiers where each classifier is generated by randomly selecting a different subset of features. The final ensemble is composed of the most fitted classifiers.

[8] suggested a nested ensemble technique for real time arrhythmia classification. A classifier model was built for each 33 training sets with enhanced majority voting technique. The nested ensembles can relieve the problem of the unlikelihood of a classifier being generated when

learning the classifier by an old dataset and limited input features. One of the reasons that make the ensemble method popular is that ensemble methods tend to solve dataset problems.

3. METHODOLOGY

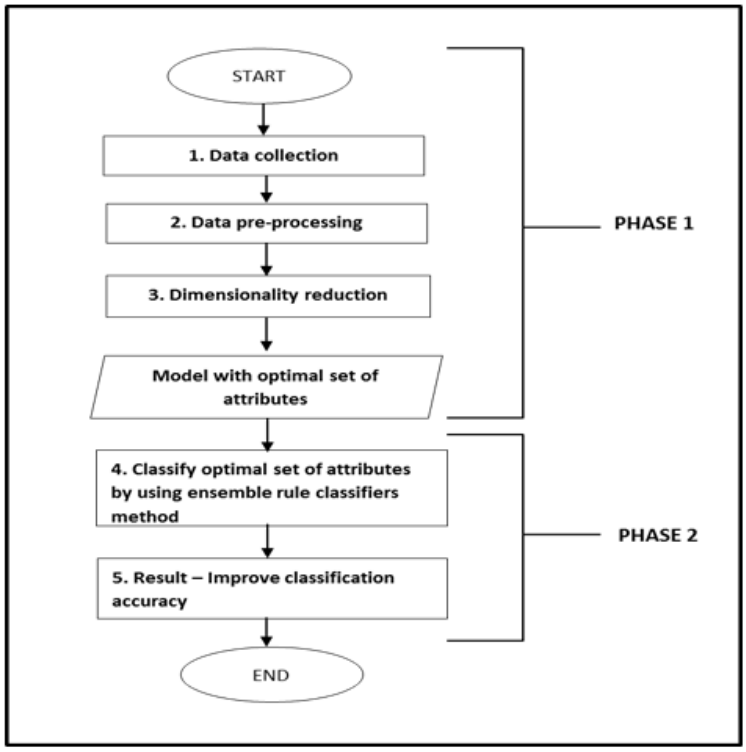

Figure 1. Methodology

The methodology is shown in Fig. 1. It consists of five (5) steps: (1) data collection; (2) data pre-processing; (3) dimensionality reduction; (4) classify an optimal set of attributes by using the ensemble rule classifier method; (5) Result-improved classification accuracy: ensemble rule classifier methods have been compared with datasets that do not use the ensemble rule classifier method. The output of phase 1 (step 1 – 3) is the optimal set of attributes. For phase 2 (step 4 – 5), the output is the improved classification accuracy by adopting an ensemble rule classifier method for the classification task. The details of the steps involved are described below:-

Step 1 (Data Collection): Six (6) different datasets were selected from UCI Machine Learning Repository. Arrhythmia dataset is one of the dataset selected due to its many features that make it challenging to explore [9]. Other five (5) datasets also were taken from different domain in order to confirm the suitability of the ensemble classifiers.

Step 2 (Data Pre-processing): Dataset that has missing values has been pre-processed in order to make sure that the dataset is ready to be experimented. All datasets were discretized since it has numeric data but needs to use classifier that handles only nominal values.

Step 3 (Dimensionality Reduction): 8 search methods and 10 reduction algorithms have been used in order to get the optimal set of attributes. The output of this step is the model consist an optimal set of attributes.

Step 4 (Classify optimal set of attributes by using the ensemble rule classifier method): In this step, the optimal sets of attributes obtained from previous step were classified by adopting the ensemble classifier method.
Step 5 (Model with good accuracy): In this step, the performance (% classification accuracy) of the dataset that used ensemble rule classifier methods has been compared with datasets that do not use the ensemble rule classifier method. The output of this step is the improved classification accuracy with optimal number of attributes.

Standard six datasets namely Arrhythmia, Bio-degradation, Ionosphere, Ozone, Robot Navigation and Spam-base from the UCI [10] were used in the experiments. These datasets include discrete and continuous attributes and represent various fields of data. The reason for choosing this dataset is to confirm the ensemble classifier is suited to all fields of data. The information on the datasets is shown in Table I.

Table 1. Dataset Characteristics.

Dataset	# of Attributes	# of Instances	# of Classes
Arrhythmia	279	452	16
Bio-degradation	41	1055	2
Ionosphere	34	351	2
Ozone	72	2536	2
Robot Navigation	24	5456	4
Spam-base	57	4601	2

All six (6) datasets were tested using 8 search methods and 10 reduction algorithms.

4. Results And Discussion

The outputs for phase 1 and phase 2 are presented in this section. The performance results are presented in the percentage of classification accuracy with the optimal set of attributes.

4.1. Phase 1 (Step 1 – 3)

Table 2. List of an optimal set of attributes selected.

Dataset	Search Method	Reduction Algorithm	# of Attr	#of Sel Attr
Arrhythmia	Best First Search	WrapperSubsetEval	279	19
Bio-degradation	Best First Search	WrapperSubsetEval	41	10
Ionosphere	Greedy Stepwise	WrapperSubsetEval	34	8
Ozone	Race Search	ClassifierSubsetEval	72	5
Robot Navigation	SubsetSizeForward	CFSSubsetEval	24	6
Spam-base	Genetic Search	WrapperSubsetEval	57	18

Table 2 shows the results of an optimal set of attributes selected by using various search method and reduction algorithm. In phase one (1), eight (8) search methods, namely Best First Search, Genetic Search, Exhaustive Search, Greedy Stepwise Search, Linear Forward Selection Search, Scatter Search, Subset Size Forward Selection Search and Ranker Search were applied. In addition, ten (10) reduction algorithms that are CfsSubsetEval, ClassifierSubsetEval, ConsistencySubsetEval, FilteredSubsetEval, ChisquaredAttributeEval, FilteredAttributeEval, GainRatioAttributeEval, InfoGainAttributeEval, PrincipalComponent and WrapperSubsetEval were adopted. It can be seen that Arrhythmia and Ozone dataset produced a massive attribute reduction, which is more than 90% reduction. Best first search (BSF) was used with WrapperSubsetEval for Arrhythmia dataset since BFS is a robust searching [11] and better for dataset studied [12]. The rest of the dataset achieved more than 60% attribute reduction. Wrapper

method (WrapperSubsetEval) performed better for 4 out of 6 datasets selected with combination of various search method. These experiments confirmed that significance attribute reduction can be accomplished by combining the right search method and reduction algorithm.

4.2. PHASE 2 (STEP 4 – 5)

In phase 2, each selected set of attributes for the six (6) various dataset namely Arrhythmia, Bio-degradation, Ionosphere, Ozone, Robot Navigation and Spam-base were classified using ensemble rule classifier methods of boosting, bagging and voting. In this phase, rule classifiers like Repeated Incremental Pruning to Produce Error Reduction (RIPPER), PART, Prism, Nearest Neighbor With Generalization (NNge) and OneR were evaluated with ensemble method. 70% of the dataset being used as training and the remaining 30% was used for testing data. The results are shown in Table 3 through Table 6.

Table 3. Classification Result of using RIPPER and RIPPER with Ensemble Rule Classifier Method.

Dataset	Without Ensemble Rule Classifier	With Ensemble Rule Classifier	
	RIPPER	Boosting + RIPPER	Bagging + RIPPER
	Acc (%)	Acc (%)	Acc (%)
Arrhythmia	73.67	73.41	73.80
Bio-degradation	83.50	83.94	83.72
Ionosphere	92.87	93.63	93.54
Ozone	93.62	93.67	93.62
Robot Navigation	96.28	97.73	97.16
Spam-base	92.65	93.24	92.92

Table III shows the classification result of using RIPPER and RIPPER with the ensemble method. RIPPER [13] with boosting and bagging method improves the classification accuracy of 4 datasets namely Bio-degradation, Ionosphere, Robot Navigation and Spam-base. These results are in line with the strength of the RIPPER that it tries to increase the accuracy of rules by replacing or revising individual rules [14]. It uses a reduced error pruning, which isolates some training data in order to decide when to stop adding more conditions to a rule. It also used a heuristic based on the minimum description length principle as stopping criterion.

Table 4. Classification Result of using PART and PART with Ensemble Rule Classifier Method

Dataset	Without Ensemble Rule Classifier	With Ensemble Rule Classifier	
	PART	Boosting + PART	Bagging + PART
	Acc (%)	Acc (%)	Acc (%)
Arrhythmia	74.13	74.98	76.93
Bio-degradation	83.94	83.86	84.69
Ionosphere	90.78	92.62	91.95
Ozone	93.76	93.86	93.81
Robot Navigation	96.88	99.06	97.74
Spam-base	93.53	93.81	93.98

Table 4 shows the classification result of using PART and PART with ensemble method. PART rule classifier with bagging method increased the classification accuracy of all the datasets. The PART algorithm [15] is a simple algorithm that does not perform global optimization to produce accurate rules. It adopts the separate-and-conquer strategy by building a rule, removes the instances; it covers, and continues creating rules recursively for the remaining instances until there are no more instances left. In addition, many studies have shown that aggregating the prediction of multiple classifiers can improve the performance achieved by a single classifier [16]. In this case, Bagging is known as a “bootstrap” ensemble method that creates individuals for its ensemble by training each classifier on a random redistribution of the training set. In contrast,

Boosting method with PART rule classifier performed better accuracy for Robot Navigation dataset with more than 3% accuracy. In this case, these results are consistent with data obtained in [17] which proved that PART algorithm is the effective algorithm to be used for classification rule hiding.

Table 5. Classification Result of using PRISM and PRISM with Ensemble Rule Classifier Method

Dataset	Without Ensemble Rule Classifier	With Ensemble Rule Classifier	
	Prism	Boosting + Prism	Bagging + Prism
	Acc (%)	Acc (%)	Acc (%)
Arrhythmia	62.36	61.71	66.07
Bio-degradation	53.42	58.71	54.61
Ionosphere	89.77	91.87	91.03
Ozone	93.79	93.75	93.88
Robot Navigation	95.21	95.35	97.38
Spam-base	80.46	81.08	81.22

Table 5 shows the Classification result of using Prism and Prism with the ensemble rule classifier method. Prism is an algorithm used different strategy to induce rules which are modules that can avoid many of the problems associated with decision trees [18]. Prism rule classifier with bagging method performed well to enhance all the dataset. In addition, boosting method with Prism produced better accuracy result for Ionosphere and Spam-base Dataset.

Table 6. Classification Result of using OneR and OneR with Ensemble Rule Classifier Method

Dataset	Without Ensemble Rule Classifier	With Ensemble Rule Classifier	
	OneR	Boosting + OneR	Bagging + OneR
	Acc (%)	Acc (%)	Acc (%)
Arrhythmia	59.76	59.69	59.37
Bio-degradation	77.03	81.69	77.03
Ionosphere	87.26	91.53	87.17
Ozone	93.88	93.84	93.82
Robot Navigation	76.01	85.09	75.99
Spam-base	79.19	90.80	79.79

Table 6 shows the classification result of using OneR and OneR with the ensemble rule classifier method. Boosting Method with OneR rule classifier performed a lot better accuracy for Bio-degradation, Ionosphere, Robot Navigation and Spam-base dataset. Huge accuracy improvement using OneR rule classifier with Boosting method for Spam-base dataset which is more than 10% accuracy increased. In this case, OneR demonstrated the efficacy as an attribute subset selection algorithm in similar cases in [20].

In summary, results have shown significant improvement in term of classification accuracy when using the ensemble rule classifier method.

5. CONCLUSIONS

In this paper, eight (8) search methods with ten (10) reduction algorithms were tested with 6 datasets. Experimental results benchmark dataset demonstrates that the ensemble method, namely bagging and boosting with rule classifiers which are (RIPPER), PART, Prism, (NNge) and OneR significantly perform better than other approaches of not using the ensemble method. Beside these, it is found that right combination between search methods and reduction algorithms shown good performance feature selection model on extracting an optimal number of attributes. For future research, methods of finding the suitable match between search method, reduction algorithm and ensemble classifiers can be developed to get a better view of the datasets.

ACKNOWLEDGEMENTS

The authors wish to thank Universiti Malaysia Terengganu (UMT), Universiti Utara Malaysia (UUM) and Kementerian Pendidikan Malaysia (KPM). This work was supported by UMT, UUM and KPM, Malaysia.

REFERENCES

[1] R. Jensen and Q. Shen, "Finding rough set reducts with ant colony optimization," Proc. 2003 UK Work., vol. 1, no. 2, pp. 15–22, 2003.
[2] N. Suguna and K. Thanushkodi, "A Novel Rough Set Reduct Algorithm for Medical Domain Based on Bee Colony," vol. 2, no. 6, pp. 49–54, 2010.
[3] B. Yue, W. Yao, A. Abraham, and H. Liu, "A New Rough Set Reduct Algorithm Based on Particle Swarm Optimization," pp. 397–406, 2007.
[4] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee, "Boosting the margin: A new explanation for the effectiveness of voting methods," Ann. Stat., vol. 26, no. 5, pp. 1651–1686, 1998.
[5] A. Tsymbal, S. Puuronen, and D. W. Patterson, "Ensemble feature selection with the simple Bayesian classification," Inf. Fusion, vol. 4, no. 2, pp. 87–100, 2003.
[6] E. Tuv, "Feature Selection with Ensembles , Artificial Variables , and Redundancy Elimination," J. Mach. Learn. Res., vol. 10, pp. 1341–1366, 2009.
[7] D. W. Opitz, "Feature selection for ensembles," Proc. Natl. Conf. Artif. Intell., pp. 379–384, 1999.
[8] M. E. A. Bashir, M. Akasha, D. G. Lee, G. Yi, K. H. Ryu, E. J. Cha, J.-W. Bae, M.-C. Cho, and C. W. Yoo, "Nested Ensemble Technique for Excellence Real Time Cardiac Health Monitoring.," in International Conference on Bioinformatics & Computational Biology, BIOCOMP 2010, July 12-15, 2010, Las Vegas Nevada, USA, 2 Volumes, 2010, pp. 519–525.
[9] E. Namsrai, T. Munkhdalai, M. Li, J.-H. Shin, O.-E. Namsrai, and K. H. Ryu, "A Feature Selection-based Ensemble Method for Arrhythmia Classification," J Inf Process Syst, vol. 9, no. 1, pp. 31–40, 2013.
[10] D. Aha, P. Murphy, C. Merz, E. Keogh, C. Blake, S. Hettich, and D. Newman, "UCI machine learning repository," University of Massachusetts Amherst, 1987. .
[11] M. GINSBERG, Essentials of Artificial Intelligence. Elsevier, 1993.
[12] R. Kohavi and G. H. John, "Wrappers for feature subset selection," Artif. Intell., vol. 97, no. 1–2, pp. 273–324, 1997.
[13] W. W. Cohen, "Fast effective rule induction," in Proceedings of the Twelfth International Conference on Machine Learning, 1995, pp. 115–123.
[14] F. Loen, M. H. Zaharia, and D. Galea, "Performance Analysis of Categorization Algorithms," in Proceeding of th 8th International Symposium on Automatic Control and Computer Science, lasi, 2004.
[15] E. Frank and I. H. Witten, "Generating accurate rule sets without global optimization," in Work, 1998, pp. 144–151.
[16] C. D. Sutton, "Classification and Regression Trees, Bagging, and Boosting," Handbook of Statistics, vol. 24. pp. 303–329, 2004.
[17] S. Vijayarani and M. Divya, "An Efficient Algorithm for Classification Rule Hiding," Int. J. Comput. Appl., vol. 33, no. 3, pp. 975–8887, 2011.
[18] J. Cendrowska, "PRISM: An algorithm for inducing modular rules," lnL J. Man-Machine Stud., vol. 27, pp. 349–370, 1987.
[19] A. Tiwari and A. Prakash, "Improving classification of J48 algorithm using bagging,boosting and blending ensemble methods on SONAR dataset using WEKA," Int. J. Eng. Tech. Res., vol. 2, no. 9, pp. 207–209, 2014.
[20] C. Nevill-Manning, G. Holmes, and I. H. Witten, "The Development of Holte's 1R Classifier."

AUTHORS

Mohammad Aizat Bin Basir is currently a lecturer in Universiti Malaysia Terengganu (UMT), Malaysia.
Faudziah Binti Ahmad is currently Assoc. Prof. in computer science in Universiti Utara Malaysia (UUM), Malaysia.