International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2016 IJSRCSEIT | Volume 1 | Issue 1 | ISSN : 2456-3307

Improving Classifier Performance Using Feature Selection with Ensemble Learning

Bhavesh Patankar*[1], Dr. Vijay Chavda[2]

*[1] Research Scholar, Department of Computer Science, Hemchandracharya North Gujarat University, Patan, Gujarat, India.

[2]NPCCSM, Kadi SarvaVishwaVidyalaya, Gandhinagar, Gujarat, India.

ABSTRACT

One of the critical task in data mining is classification. It is very much important in classification to achieve maximum accuracy. In the field of data mining, numerous classifiers are present for the classification task. Each classification techniques have their pros and cons. Some of the techniques work well with certain data sets while other techniques work well with other data sets. There have been many techniques evolved for improving classification accuracy. One of such technique is pre-processing which helps in improving quality of the data. Another method is to combine the classifiers, which will in turn improve the classification accuracy. In this paper, empirical study is been done on various techniques for improving classification accuracy. One of the technique is feature selection, which will select best features from the available features in the data set. Other approach is ensemble learning which combines many classifiers to improve the classification accuracy.

Keywords: Classification; Pre-processing; Feature Selection; Ensemble Learning;

I. INTRODUCTION

In data mining, it is evident that classification accuracy is the critical factor for classification techniques. Many classification techniques are been evolved in data mining, but not every technique is suitable for all data sets. They are various techniques available in order to improve the classification accuracy. Sometimes, data, which used to do classification, is not as of required quality. Therefore, it is good to improve the quality of the data, which will result in improving the classification accuracy. In data mining, pre-processing is one of the task, which deals with the data set. It has been seen that a wide variety of techniques are available for data pre-processing like noise reduction, data cleaning which includes filling missing values, feature selection, dimensionality reduction, etc [1]. Ensemble techniques have appeared as an influential technique for improving the strength as well as the accuracy of both solutions (i.e. supervised and unsupervised). In addition, as massive amounts of data constantly produced from different sights, it is vital to combine different concepts for smart decision-making. In the past few years, there have been various studies on the problem of combining models into a single model, and the success of ensemble techniques seen in multiple disciplines, including anomaly detection, intrusion detection, recommendation systems and web applications [2].

Many papers are been reviewed to figure out various parameters to be taken into consideration in order to improve the classification accuracy. It is good to have pre-processing step before the classification done in order to achieve the increasing accuracy of the classification. The available source data set is been converted into more qualitative data set. In some cases, it may occur that data set can contains high dimensions; many of the dimensions may be irrelevant for our classification approach. Hence, it becomes necessary to perform Feature selection to utilize the best features for achieving the greater accuracy in classification. Many techniques recommended reducing noise and outliers for the improvement of classification accuracy.

II. FEATURE SELECTION

Achieving greater accuracy is very much important in any data mining process. An aim of feature selection is selecting a subset of relevant features for generating strong learning models. Camelia Vidrighin et al.[3] have considered the wrapper approach, as a combination of three steps: model generation, model evaluation and model validation. They have focused on uniting feature selection with filling the missing values in order to improve the performance of the learning schemes. Analysis on various approaches for feature selection have been done and based on the result best models have been identified which have consistently improved the accuracy of classification.

Feature selection can be termed as combination of search technique to find out the best features out of the available features in the given data set. The simplest algorithm, which minimizes the error rate, is been considered. As seen earlier wrapper methods use predictive model to get the relevant feature subsets. Wrapper methods are considered computationally very much intensive, but generally provide best feature sets from the given data set for the given classification model. Filter methods use proxy measure to select the optimum feature set. Filter techniques are generally computationally less intensive than wrapper techniques. Hence they produced feature set which are not tuned to specific models and so classification accuracy from filters are generally lesser than what we can achieve from wrapper methods.

III. ENSEMBLE LEARNING

Ensemble learning techniques are learning algorithms that generate a set of classifiers and then classify new data points by considering a (weighted) vote of their estimates. The novel ensemble technique is Bayesian averaging, however more recent techniques include error-correcting output coding, boosting, and bagging. Dietterich et. al. [4] have reviewed these methods and explained why ensembles often perform better than any single classifier. They have reviewed some previous studies comparing ensemble methods and some new experiments is been shown to expose the causes that Adaboost does not overfit rapidly.

It is known that a neural network ensemble unites a finite number of neural networks or other types of interpreters, which are trained concurrently for a common classification assignment. After the experimentation, on comparing with a single neural network, the ensemble is able to efficiently improve the classification accuracy of the classifier. Zhao et. al. [5] have surveyed many ensemble techniques on different data sets to see the effect of it. And in the survey they have found that ensemble of neural network always perform better than the single neuron. Lira et. al. [6] have developed an ANN-based automatic classifier for power system disturbance waveforms. In the training process, actual voltage waveforms applied and then Signals processed in two steps that is decomposition and Principal Component analysis (PCA) which results in reducing the input space of the classifier to a much lower dimension. Classification task was carried out using a combination of six Multilayer perceptrons with different. The result of experiment with real data indicate that the random committee is clearly an effective way in order to improve disturbance classification accuracy when it compared with the average and the separate models. Natesan et. al. [7] have worked on secure communication between two parties. They have proposed an Adaboost algorithm for network intrusion detection system with single weak classifier. The classifiers as Naive Bayes, Bayes Net and Decision tree are been used as weak classifiers. Experiments carried out with the help of benchmark data set to reveal that boosting algorithm can significantly improve weak classifiers classification accuracy. Finally, the results were very much effective. Base classifiers Naive Bayes and Decision Tree have shown comparatively better performance as a weak classifier with Adaboost.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

Experiment are been carried out using Weka. Weka (Waikato Environment for Knowledge Analysis) is a widespread machine-learning tool developed in JAVA language. It is evident that it is one of the free open source softwares available under the GNU General Public License. Considering the experiment, it executed on base classifier and then accuracy is measured. Consequently, the experiment carried out on the classifier with feature selection followed by boosting and then the accuracy is measured. Data sets used in the experiment is been collected from UCI machine repository. At the end, results are been compared and conclusion is drawn.

Following datasets from the UCI Machine Learning Repository are been collected to initiate the experiment.

Sr.No	**Dataset Information**		
	Dataset	***Instances***	***Attributes***
1	Iris	150	5
2	Diabetes	768	9
3	Ionosphere	351	35

Table 1. Data set information

The experiment is been performed using Multilayer perceptron, J48 and Naïve Bayes classifier. While carrying out the experiment the data sets are been chosen and not a single filter is applied on them. Firstly experiment is performed using single base classifier on the data set without feature selection applied then experiment is carried out using single base classifier with adaboost and data set with feature selection applied on it. The experiment is been carried out using weka 3.8.0.

Accuracy of the base single classifier and base classifier with adaboost and feature selection is measured which is displayed in given below table.

Classifier	Datasets		
	Iris	***Diabetes***	***Ionosphere***
Multilayer Perceptron	97.3	75.39	91.16
Multilayer Perceptron with AdaBoost and feature selection	95.33	75.52	94.30
J48	96.00	73.82	91.45
J48 with AdaBoost and feature selection	94.67	73.58	94.30
Naïve Bayes	96.00	76.30	82.62
Naïve Bayes with AdaBoost and feature selection	96.00	77.47	92.30

Table 2. Accuracy measures of Multilayer perceptron, J48 and Naïve Bayes on Iris, Diabetes and Ionosphere data set with feature selection and adaboost and without feature selection and without adaboost.

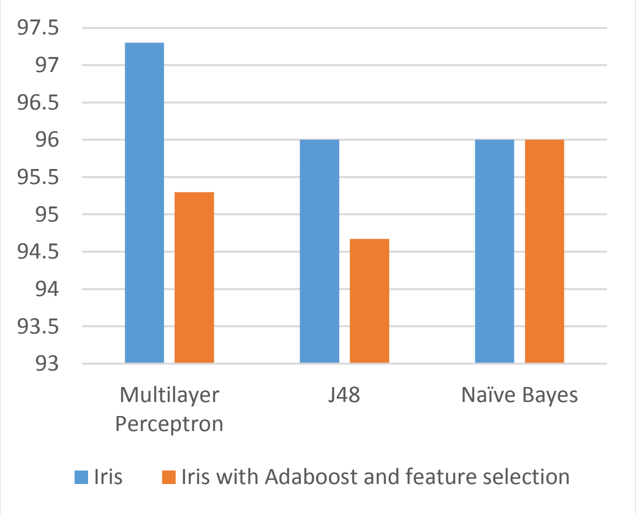

Figure 1. Comparison of Multilayer perceptron, J48 and Naïve Bayes with feature selection and adaboost and without feature selection and adaboost on Iris data set.

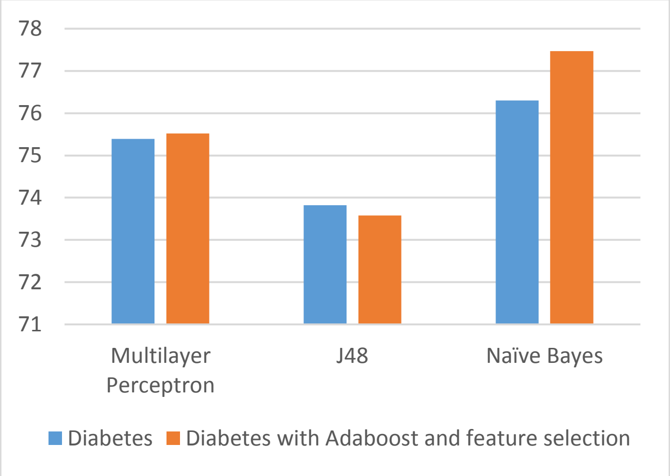

Figure 2. Comparison of Multilayer perceptron, J48 and Naïve Bayes with feature selection and adaboost and without feature selection and adaboost on Diabetes data set.

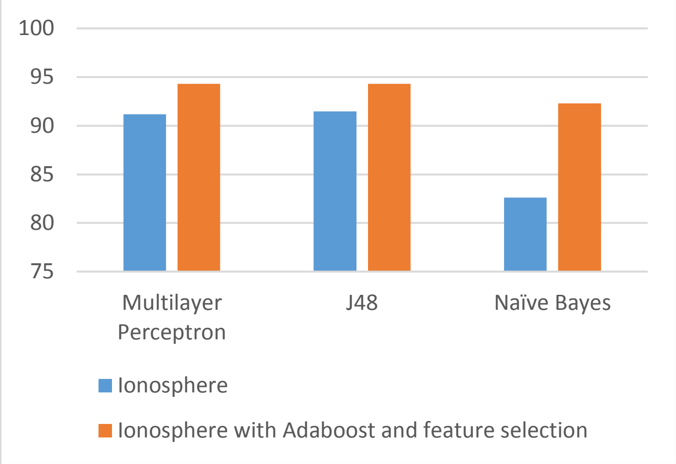

Figure 3. Comparison of Multilayer perceptron, J48 and Naïve Bayes with feature selection and adaboost and without feature selection and adaboost on Ionosphere data set.

V. CONCLUSION

In this paper, it is evident that classification accuracy improved with the help of feature selection and ensemble technique like Adaboost, which is been used in this experiment. Here, Best First method with CFS Subset Evaluation is been used to select the optimum feature in order to improve the classification accuracy. After that ensemble technique is used which combines the multiple classifier in order to improve the classification accuracy. Here Adaboost ensemble technique is been used for the improvement of the classification accuracy. From the results of the experiment, it is clear that in most of the cases feature selection with ensemble technique definitely improves the classification accuracy of the classifier. Future work include using different feature selection approach than what is been used in this paper. In addition, instead of AdaBoost any other ensemble technique can be utilize to see the result.

VI. REFERENCES

[1] Moeinzadeh, H, Nasersharif, B, Rezaee, A., Pazhoumand-dar, H., “Improving Classification Accuracy Using Evolutionary Fuzzy Transformation”, 11th Annual Conference on Genetic and Evolutionary Computation Conference (GECCO 2009), Montreal, Canada, 2009 (1)

[2] Han, Jiawei, Micheline Kamber, and Jian Pei. Data mining, southeast asia edition: Concepts and techniques. Morgan kaufmann, 2006.

[3] Bratu, Camelia Vidrighin, Tudor Muresan, and Rodica Potolea. "Improving classification accuracy through feature selection." Intelligent Computer Communication and Processing, 2008. ICCP 2008. 4th International Conference on. IEEE, 2008.

[4] Dietterich, Thomas G. "Ensemble methods in machine learning." International workshop on multiple classifier systems. Springer Berlin Heidelberg, 2000.

[5] Zhao, Ying, Jun Gao, and Xuezhi Yang. "A survey of neural network ensembles." *2005 International Conference on Neural Networks and Brain*. Vol. 1. IEEE, 2005.

[6] Lira, Milde MS, et al. "Combining multiple artificial neural networks using random committee to decide upon electrical disturbance classification." 2007 International Joint Conference on Neural Networks. IEEE, 2007. Nilsson,R., Statistical Feature Selection, with Applications in Life Science, PhD Thesis, Linkoping University, 2007.

[7] Natesan, P., P. Balasubramanie, and G. Gowrison. "Improving the attack detection rate in network intrusion detection using adaboost algorithm." Journal of Computer Science 8.7 (2012): 1041.