© 2016 IJSRST | Volume 2 | Issue 5 | Print ISSN: 2395-6011 | Online ISSN: 2395-602X
Themed Section: Science and Technology

Effect of Feature Selection Using Best First Search on the Performance of Classification

Bhavesh Patankar*1, Dr. Vijay Chavda2

*1Research Scholar, Department of Computer Science, Hemchandracharya North Gujarat University, Patan, Gujarat, India.
2NPCCSM, Kadi SarvaVishwaVidyalaya, Gandhinagar, Gujarat, India

ABSTRACT

Performance in terms of high accuracy is very much required to any data mining system. A very much influencing factor on the success or failure of the data mining process is the quality of data used in the process. Feature selection is one of the preprocessing techniques in data mining which can prepare a quality data set before feeding it into the data mining process. This paper focuses on the Best First Search technique of feature selection. Feature selection is done using Best First Search. After selecting the features in given data set, it is taken as an input to the data mining process. It is observed that when classification is performed with feature selection, performance of classifier Performance of the classification. The success of a data mining problem heavily depends upon the quality of the data which is the most influencing factor. Moreover, feature selection represents one of the tools which can refine a dataset before presenting it to a learning scheme. In this paper, analyses of a wrapper approach for feature selection, with the purpose of boosting the classification accuracy is done. Experimental evaluations have been performed of feature selection on several data sets. The results showed that feature selection improves the overall performance in classification. Accuracy and speeds up the training process.

Keywords: Data Mining; Classification; Preprocessing; Feature Selection;

I. INTRODUCTION

The process of extracting knowledge from huge amount of data accessible from numerous data sources which are collected in data warehouse is termed as Data mining. In Data mining, accuracy of classification depends on lot of factors. One of the major factor is selection of relevant features and eliminating the irrelevant features from the data sources. Classification accuracy can be high if high quality data is fed as an input. In this paper reviews for feature selection methods have been made and explained why feature selection can often perform better to improve the classification accuracy. Identifying the relevant feature is a central problem in data mining. Many attempts have been made in defining relevance in machine learning. In 1994 Kohavi, John and Pfleger have noted in their analysis of defining relevant features. The earlier methods to feature selection were filtering methods. The simplest technique is to test each possible subset of features finding the one which help in minimizing the error rate. There are three main categories of the feature selection algorithm: filters, wrappers and embedded methods.

Filters, generally produce a feature set which is not tuned to a specific type of classification model. Filters are less computationally intensive. A feature set produced from filter is generally more general. Filter methods are also being used as pre-processing step for wrapper methods.

Wrapper techniques use predictive model to generate optimum features. Each new feature set is used to train the model, which is tested on hold out set. Then counting the number of mistakes made on that hold out set provide the score of the sub set. Because wrapper methods train a new model for each and every subset, which are computationally intensive and generally produces good results.

IJSRST16224 | Received : 28 August 2016 | Accepted: 03 September 2016 | September-October-2016 [(2)5: 01-03]

Embedded techniques are a group of techniques which perform feature selection as a part of the model building process.

II. METHODS AND MATERIAL

Literature Review

Feature selection is a practical approach towards selecting optimum features in order to improve the classification accuracy in machine learning.

Bratu et. al. have analyzed wrapper approach for q/0feature selection Their experiment results shows feature selection improves the accuracy and speed up the training process. They have proposed two robust combination. One which constantly achieves highest accuracy and other which significantly boost initial accuracy [1]. A. Jain et. al. studied the problem of choosing an optimal feature set for land use classification based on SAR satellite images using four different texture models. It showed that pooling features resulting from diverse texture models which is followed by a feature selection results in a significant improvement in the classification accuracy. They also illustrated the endangerments of using feature selection in very small size of samples[2]. M. Dash et. al. have surveyed many feature selection methods since 1970 [3]. Pat Langley had defined problem of selecting relevant feature in machine learning with the view of searching relevant feature set [4]. George H. John et. al. have derived that the features selected should be determined by induction algorithm and not only on the features and the target concept.

III. RESULT AND DISCUSSION

Weka tool is used in order to carry out the experiment. Weka (Waikato Environment for Knowledge Analysis) is a very famous machine learning tool developed in JAVA language. It is open source software and accessible under the GNU General Public License. So, it is available with no cost. Here, the experiment is performed on base classifier with feature selection then accuracy is measured then without feature selection and then accuracy is measured. The data sets used in the experiment is collected from UCI machine repository. Finally, results are compared which results in the conclusion.

UCI Machine Learning Repository is the source of data set which is used perform the experiment.

Sr.No	**Dataset Information**		
	Dataset	***Instances***	***Attributes***
1	Glass	214	10
2	Vehicle	846	19
3	Iris	150	5

The experiment is carried out on Multilayer Perceptron, J48 and Naïve Bayes classifier. On the chosen datasets no filter is applied while carrying out the experiment.

Firstly, data sets are taken without feature selection applied on them and classification is done using above mentioned classifiers. Secondly, data sets are taken with feature selection applied on them and classification is performed using above mentioned classifiers after feature selection. Here, weka 3.6.12 is used to perform the experiment. Accuracy of the base classifier on different data sets with feature selection is measured then without feature selection is measured which is displayed in below table.

Classifier	**Datasets**		
	Glass	***Ionosphere***	***Iris***
Multilayer Perceptron	67.75	91.16	97.33
Multilayer Perceptron with feature selection	65.83	93.45	95.33
J48	66.82	91.45	96
J48 with feature selection	68.69	90.6	96
Naïve Bayes	48.59	82.62	96
Naïve Bayes with feature selection	47.66	90.6	96

We can see the result of the classifiers with feature selection and without feature selection. The columnar chart clearly shows the effect of feature selection on various data sets.

From the graph, it is visible that when feature selection is applied on the data set, Multilayer perceptron performed better in terms of accuracy and speed. However, in some cases Multilayer perceptron performed better on the data set for which relevant features are not selected.

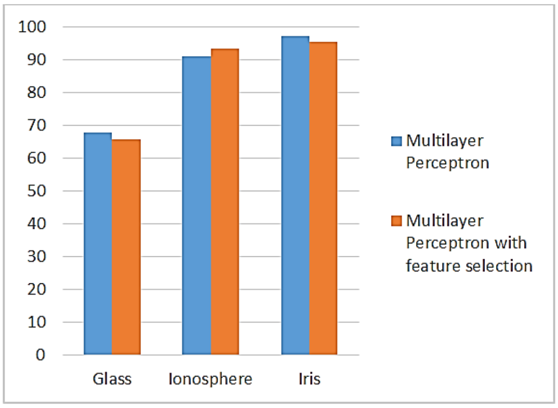

Figure 2. Multilayer Perceptron with and without feature selection comparison

It is clearly seen from the graph that when feature selection is applied on the data set, J48 performed better in terms of accuracy and speed. However, in Ionosphere data set J48 performed better on the data set for which relevant features are not selected.

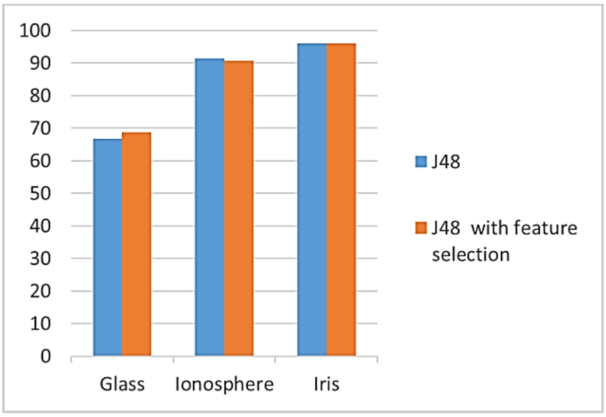

Figure 3. J48 with and without feature selection comparison

It is clearly seen from the graph that when feature selection is applied on the data set, Naïve Bayes performed better in terms of accuracy and speed. However, in Glass data set Naïve Bayes performed better on the data set for which relevant features are not selected.

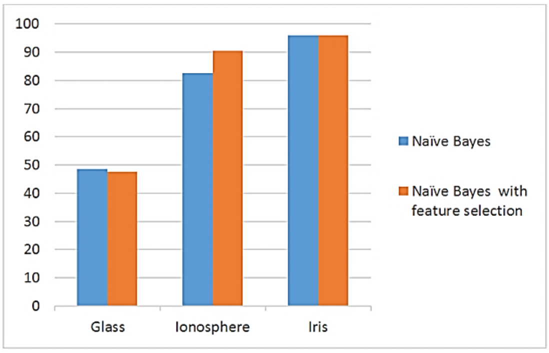

Figure 4. Naïve Bayes with and without feature selection comparison

IV. CONCLUSION

The paper shows the effect of feature selection on classification accuracy by using different classifiers on different data sets. The experiment was carried out using weka 3.6.12. For attribute evaluation CfsSubsetEval algorithm and BestfirstSearch Search method is utilized. The results depicted the effect of feature selection on various base classifiers. Moreover, it was observed that in majority of cases for all the selected datasets, there is increase in classification accuracy when feature selection is applied on the data sets instead of using data sets without applying feature selection. In a nutshell, feature selection helps in improving the accuracy of classification. Future work can include the effects of changing the feature selection techniques like genetic algorithm.

V. REFERENCES

[1] 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.

[2] Jain, Anil, and Douglas Zongker. "Feature selection: Evaluation, application, and small sample performance." IEEE transactions on pattern analysis and machine intelligence 19.2 (1997): 153-158.

[3] Dash, Manoranjan, and Huan Liu. "Feature selection for classification." Intelligent data analysis 1.3 (1997): 131-156.

[4] Langley, Pat. "Selection of relevant features in machine learning." Proceedings of the AAAI Fall symposium on relevance. Vol. 184. 1994.

[5] John, George H., Ron Kohavi, and Karl Pfleger. "Irrelevant features and the subset selection problem." Machine learning: proceedings of the eleventh international conference. 1994.