International Journal of Advanced Science and Technology
Vol.87 (2016), pp.1-8
http://dx.doi.org/10.14257/ijast.2016.87.01

Feature Selection Using Rough Set For Improving the Performance of the Supervised Learner

D. Asir Antony Gnana Singh, E. Jebamalar Leavline, E. Priyanka and C. Sumathi

University College of Engineering, Bharathidasan Institute of Technology Campus, Anna University, Tiruchirappalli- 620 024
asirantony@gnail.com, jebi.lee@gmail.com, epriyanka76@gmail.com, sumathi205098@gmail.com

Abstract

Prediction plays a significant role in the human life to predict the situation, climate, finance, outcome of the particular event or activities, etc. This predication can be achieved by the classifier which is formally known as supervised learner. The classifier can be built using the dataset and its performance is based on the attributes or features present in the dataset which are highly relevant to the predictive target attributes. The feature selection process removes the redundant and irrelevant features from the dataset to improve the performance of the classifier. This paper proposes a rough set-based feature selection method to remove the redundant and irrelevant features in order to improve the performance the classifier. The proposed method is tested on the various datasets with the various supervised learning algorithms and it is evident that the proposed method producing the better performance than the other methods compared.

Keywords: *Rough set, classification, feature reduction, classifier, data mining*

1. Introduction

Data mining is a process of analyzing data from huge volumes of data for obtaining useful information that can be the desirable pattern to extract the knowledge. In other words data mining helps us to perform the prediction by building the predictive model which is also known as classifier that predicts the unknown data from the known data. The machine learning algorithms are usually divided into two different categories: supervised learning and unsupervised learning algorithm. The supervised learning algorithm is also known as classification algorithm that builds the classifier to perform classification or prediction.

Several classifiers have been developed in the classification literature including k-nearest neighbor algorithm (k-NN), Naïve Bayes (NB), support vector machine (SVM), decision tree, and so on. The unsupervised learning algorithm is the clustering algorithm that builds the clustering model in order to cluster or group the objects into similar categories. The feature selection, variable selection or variable subset selection is a process of obtaining a subset of relevant features from large dataset. Too many features may affect the classification accuracy. Hence, the feature selection is employed in data mining in order to improve the accuracy of classifier.

The feature selection algorithm can be classified into wrapper, filter, and embedded method. In this filter method, the subset selection procedure is independent to the learning algorithm. This method leads to a faster learning process. However, the resulting subset with a specific criterion may not work very well in the learning algorithm. The wrapper method uses a predictive model to select the feature subsets. This method attempts to select the significant features of minimal size according to the criteria based on the output of supervised learning that is adopted for the selection process. The newly selected feature subset is used to train a model. This method trains a new model for all subsets, and hence

ISSN: 2005-4238 IJAST

it is computationally intensive, but it will provide the best performing feature set for particular type of model. Embedded method is a technique which performs feature selection as a part of model construction process. These approaches tend to be between filter and wrapper in terms of computational complexity. The learning algorithm takes advantages of its own variable selection algorithm. The application of the feature selection includes image recognition, bioinformatics, text classification, system monitoring, and clustering.

In this paper, a rough set-based feature selection algorithm is proposed. The rough set is a mathematical approach for incomplete and uncertain data. This theory is a framework to transform data into information. Rough set theory is used to find irrelevant and redundant data and also be useful to find data dependency between huge amounts of data. This rough set is used in various application related in the area of data mining.

The rest of this paper is structured as follows: Section 2 reviews the literature, Section 3 expresses the proposed work and Section 4 provides the details of implementation and experimental setup. Section 5 discusses results and Section 6 draws the conclusion.

2. Related Works

This section explores the research works which are related to the proposed work. Xiaoyue Wang *et al.*, proposed a hybrid approach of rough set and genetic algorithm for web page classification [1]. This approach employs the rough set technique for feature reduction and applies genetic algorithm based on an analogy for biological evolution, and then an initial population is created by randomly generated rules. The results of this paper show that support vector machine with rough set and genetic algorithm seems to be a useful tool for inductive learning. In this paper, they used rough set method for generating classification rules for set of 360 samples of the breast cancer data. The results of this study show that rough set is a promising tool for machine learning and for building the expert system.

Sridevi and Murugan proposed a rough set-based attribute reduction for breast cancer diagnosis [2]. The effectiveness of rough set reductive algorithm is analysed on breast cancer dataset and the result shows that this method yields a better attribute reduction. Shraddha Sarode *et al.*, proposed an approach for dimensionality reduction for web page classification and described a hybrid approach for attribute reduction using rough set and Naïve Bayes method [3]. As compared to the traditional approach, this method requires less processing time because it uses dimensionality reduction technique. Suman Saha *et al.*, proposed a rough set-based approach for ensemble classifier for web page classification [4]. The results of this research show that, by removing redundant features it reduces the CPU load compared to other ensemble classifier techniques.

M. Zhang and J. T. Yao presented a rough set approach to feature selection [5]. In this paper, they used a rough set feature selection method known as parameterized average support heuristic. Huilian Fan *et al.*, [6] proposed a rough set- based feature selection based with wasp swarm optimization. Nick Cercone *et al.*, [7] presented a feature selection with rough set for web page classification. This paper indicates that the rough set feature selection approach can improve the classification accuracy. Xiaoguang Qi *et al.*, [8] developed a web page classification with feature selection algorithm and they used a web mining techniques. Jan Bazan *et al.*, presented a view on rough set concept approximation. This paper indicates that approximation is a main fundamental concept for rough set [9].

Yan Wang *et al.*, proposed a rough set-based feature selection for medical dataset. It says that rough set feature selection known as feature forest algorithm can improve classification accuracy [10]. From these literatures, it is observed that the rough set based feature selection produces better performance in feature selection

and various applications in the data mining field. Therefore the proposed method adopts the rough method for selecting the significant features from the dataset.

3. Proposed Method

This section explains the proposed method and the algorithm developed. The overall architecture of the proposed method is illustrated in the Figure 1. Initially, the dataset which is to be pre-processed is loaded. Then the developed rough-set based feature selection algorithm is performed on the loaded dataset and the significant features are obtained. Then the dataset with the obtained features are split into training and testing datasets. Then the classification model is built by the training dataset using the classification algorithms NB, k-NN, SVM, and decision tree. Then the constructed classifier is tested individually with the testing dataset for calculating the classification accuracy of the classifier.

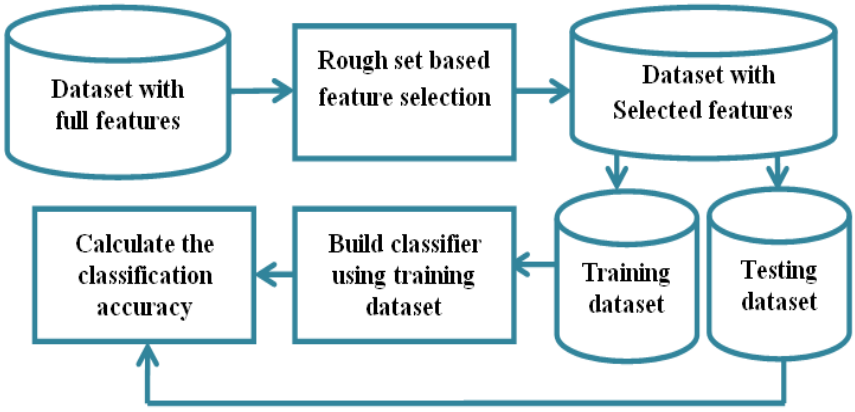

Figure 1. Overall Architecture of the Proposed System

3.1. Description of the Algorithm:

The proposed algorithm is implemented with the following seven steps. The sample dataset is illustrated in the Table 1. In step 1, the unique rows from the dataset are separated as shown in the Table 2. Step 2 finds the differing features in each instance by comparing it with all other instance as shown in Figure 2.

Step 3 finds the overall concatenated comparability matrix. Step 4 finds the lower and upper bounds of the dataset. Step 5 finds the sum of occurrence of each feature in the comparability matrix. Step 6 finds the significant features. Step 7 obtains the new dataset with significant features and finds classification accuracy.

Step 1: Find the unique rows in the data set.
Step 2: Find the differing features in each instance by comparing it with all other instance.
Step 3: Find the overall concatenated comparability matrix (Table 3).
Step 4: Find the lower and upper bounds of the data set.

Lower Bound=Union(Y ε Ind(Data Set):Y is subset of X)
where X is the given data set, Ind is the indiscernibility relation
Ind (Data Set) = {{1},{2,5},{3},{4},{6}}
X Yes= {1,2,3,6}
X No= {4,5)
Lower Bound= {1,3,6}
Upper Bound= Union(Y ε Ind(Data Set):Y intersection X !=empty)

Upper Bound= {1,2,3,5,6}
Difference set between lower and upper bounds:{2,5}
Therefore, it is a rough set since difference set is non-empty.

Step 5: Find the sum of occurrence of each feature in the comparability matrix.

Here, Feature 1's occurrence=6
Feature 2's occurrence=4
Feature 3's occurrence=9

Step 6: Finding the significant features:

Since the comparability matrix is designed based on the features that vary in value, the maximum varying feature gets a greater value. So, features 3 and 1 are significant in the sample dataset illustrated in Table 1.

Step 7: Obtain classification accuracy for new dataset with significant features.

The new data set is the one that contains the selected significant features. This data set is passed to the classifiers such as support vector machine (SVM), Naive Bayes (NB), k-nearest neighbour and decision Tree. The original unprocessed data set is also evaluated using the same classifiers for comparison.

Table 1. Sample Dataset

Object	Decisional Attribute	Attribute 1	Attribute 2	Attribute 3
1	1	0	1	2
2	1	1	0	2
3	1	1	1	3
4	0	0	1	1
5	0	1	0	2
6	1	0	1	3

3.2. Implementation

The proposed method is implemented using MATLAB12 version with the specification of Windows operating system, 4 GB RAM, and 120 GB hard disk. The proposed method is tested on various datasets which are discussed in the Section 4.1 and the accuracy of the supervised learning algorithms namely k-nearest neighbour algorithm (k-NN), Naïve Bayes (NB), support vector machine (SVM), decision tree are obtained with the proposed algorithm and the results are discussed in Section 5.

3.3. Dataset Used

For the conduction of the experiment the datasets namely Lung Cancer dataset and Reuter corn dataset are collected from the UCI repository [11] and WEKA software dataset repository [12], respectively. The Lung Cancer dataset contains the 56 attributes and 32 instances. The features are extracted for the Reuters corn dataset using WEKA filter approach and the total number of features and the instances are 2233 and 1554, respectively, from the 2233 features randomly the 62 features with 45 instances are selected for the conduction of the experiment.

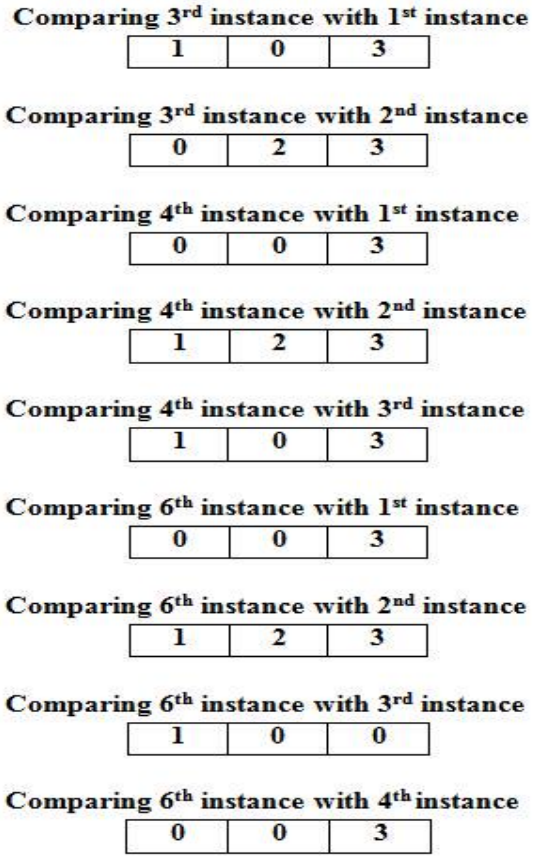

Figure 2. Comparing the Instances with other Instance

Table 2. Unique Rows Obtained from the Dataset shown in Table 1

Object	Decisional Attribute	Attribute 1	Attribute 2	Attribute 3
1	1	0	1	2
2	1	1	0	2
3	1	1	1	3
4	0	0	1	1
6	1	0	1	3

Table 3. Overall Concatenated Comparability Matrix

Attribute 1	Attribute 2	Attribute 3
1	0	3
0	2	3
0	0	3
1	2	3
1	0	3
0	0	3
1	2	3
1	0	0
0	0	3

4. Results and Discussion

The results are obtained using the proposed method tested on the dataset Reuter con with the various classification algorithms are illustrated in the Figure 3.

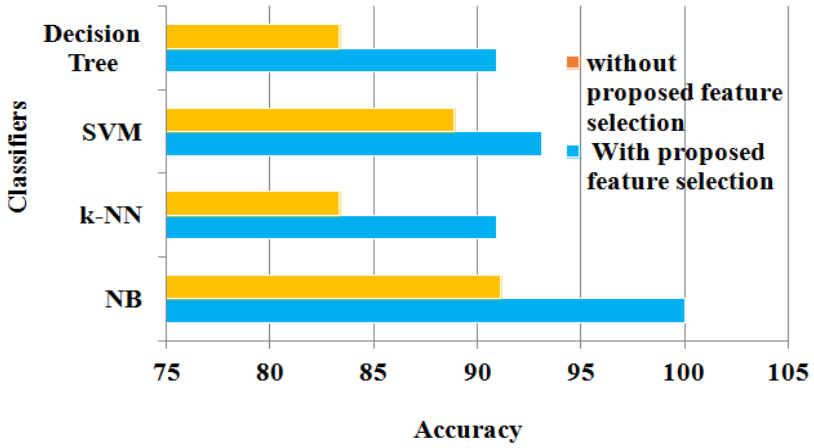

Figure 3. The Classification Accuracy of Various Classifiers with and without Proposed Method on Dataset Reuters Corn

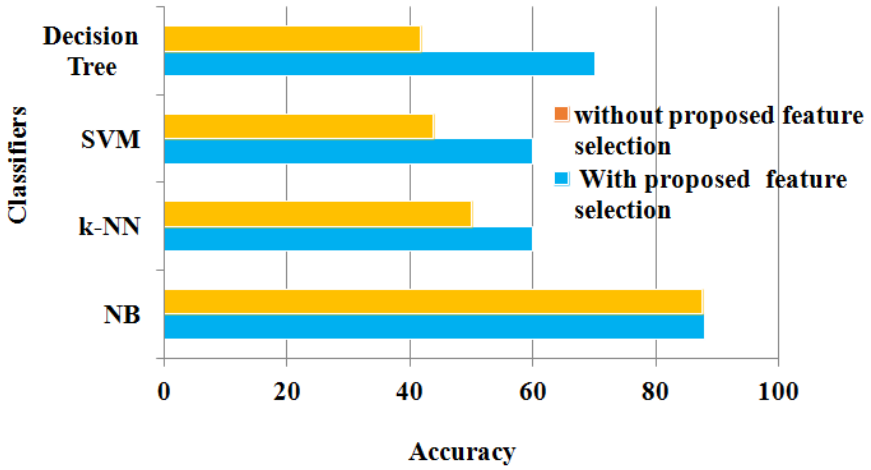

Figure 4. The Classification Accuracy of Various Classifiers with and without Proposed Method on Dataset Lung Cancer

The classification accuracy of various classifiers with and without proposed method on the dataset Lung cancer is illustrated in the Figure 4. From the Figure 3, it is observed that the proposed method with NB gives better classification accuracy compared to all other classifiers on the Reuters Corn dataset. Without feature selection algorithm, accuracy with the various classifiers is very less on the Reuters Corn dataset. From Figure 4, NB is producing the same accuracy for with and without features selection and the proposed method is producing better accuracy with the Decision tree compared to SVM and k-NN.

5. Conclusion and Future Work

This paper proposed rough set-based feature selection. This proposed method is tested on the Reuters Corns and Lung cancer datasets with the various classifiers Decision tree, NB, SVM, and k-NN. This proposed method is producing better

accuracy for various classifiers. This proposed method produces better accuracy with NB compared to other classifiers. This proposed feature selection algorithm can be implemented with different selection strategies for feature selection process with various classification algorithms.

References

[1] X. Wang, Z. Hua and R. Bai, "A Hybrid Text Classification model based on Rough Sets and Genetic Algorithms", Proceedings of IEEE Ninth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, **(2012)**, pp. 971-977.
[2] T. Sridevi and A. Murugan, Indian Journal of Innovations and Developments, vol. 1, no. 5, **(2012)**.
[3] S. Sarode and J. Gadge, International Journal of Computer Applications, vol. 99, no. 14, **(2014)**.
[4] S. Saha, C. A. Murthy and S. K. Pal, Fundamenta Informaticae, vol. 76, **(2007)**.
[5] M. Zhang and J. T. Yao, "A rough sets based approach to feature selection", Proceedings of IEEE Annual Meeting of the Fuzzy Information Processing, **(2004)**, pp. 27-30.
[6] H. Fan and Y. Zhong, Journal of Computational Information Systems, vol. 8, no. 3, **(2012)**.
[7] A. An, Y. Huang, X. Huang and N. Cercone, "Transactions on Rough Sets II", Springer Berlin Heidelberg, pp. 1-13.
[8] X. Qi and B. D. Davison, "ACM Computing Surveys", vol. 41, no. 2, **(2009)**.
[9] J. Bazan, N. Hung Son, A. Skowron and M. Szczuka, "Fuzzy Sets, Data Mining, and Granular Computing", Springer Berlin Heidelberg, **(2003)**, pp. 181-188.
[10] Y. Wang and L. Ma, "Feature selection for medical dataset using rough set theory", Proceedings of 3rd WSEAS international conference on Computer engineering and applications, **(2009)**, pp. 68-72.
[11] M. Lichman, "UCI Machine Learning Repository", Irvine, CA: University of California, School of Information and Computer Science. Available from: http://archive.ics.uci.edu/ml. [Accessed 15 June 2015], **(2013)**.
[12] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten, "The WEKA Data Mining Software: An Update", SIGKDD Explorations, vol. 11, no. 1, **(2009)**.

