Volume 7, Issue 2, February 2017 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Comparative Study: Classification Algorithms Before and After Using Feature Selection Techniques

Mona Mohamed Nasr
Information Systems Department,
Helwan University, Egypt

Essam Mohamed Shaaban
Information Systems Department,
Beni-Suef University, Egypt

Menna Ibrahim Gabr*
Business Information System Dept.,
Helwan University, Egypt

DOI: 10.23956/ijarcsse/V7I2/01212

Abstract*— Data classification is one of the most important tasks in data mining, which identify to which categories a new observation belongs, on the basis of a training set. Preparing data before doing any data mining is essential step to ensure the quality of mined data. There are different algorithms used to solve classification problems. In this research four algorithms namely support vector machine (SVM), C5.0, K-nearest neighbor (KNN) and Recursive Partitioning and Regression Trees (rpart) are compared before and after applying two feature selection techniques. These techniques are Wrapper and Filter. This comparative study is implemented throughout using R programming language. Direct marketing campaigns dataset of banking institution is used to predict if the client will subscribe a term deposit or not. The dataset is composed of 4521 instances. 3521 instance as training set 78%, 1000 instance as testing set 22%. The results show that C5.0 is superior to other algorithms before implementing FS technique and SVM is superior to others after implementing FS.

Keywords*— Classification, Feature Selection, Wrapper Technique, Filter Technique, Support Vector Machine (SVM), C5.0, K-Nearest Neighbor (KNN), Recursive Partitioning and Regression Trees (Rpart).

I. INTRODUCTION

The problem of data classification has numerous applications in a wide variety of mining applications. This is because the problem attempts to learn the relationship between a set of feature variables and a target variable of interest. Excellent overviews on data classification may be found in Classification algorithms typically contain two phases. The first one is training phase in which a model is constructed from the training instances. The second is testing phase in which the model is used to assign a label to an unlabeled test instance[1].

Classification consists of predicting a certain outcome based on a given input. In order to predict the outcome, the algorithm processes a training set containing a set of attributes and the respective outcome, usually called goal or prediction attribute. The algorithm tries to discover relationships between the attributes that would make it possible to predict the outcome. Next the algorithm is given a data set, called prediction set, which contains the same set of attributes, except for the prediction attribute – not yet known. The algorithm analyses the input and produces predicted instances. The prediction accuracy defines how "good" the algorithm is [2]. The four classifiers used in this paper are shown in (figure 1). But many irrelevant, noisy or ambiguous attributes may be present in data to be mined. So they need to be removed because it affects the performance of algorithms. Attribute selection methods are used to avoid over fitting and improve model performance and to provide faster and more cost-effective models [3]. The main purpose of Feature Selection (FS) approach is to select a minimal and relevant feature subset for a given dataset and maintain its original representation. FS not only reduces the dimensionality of data but also enhance the performance of a classifier. So, the task of FS is to search for best possible feature subset depending on the problem to be solved [4].

This paper is organized as follows. Section 2 refers to the four algorithms to deal with the classification problem. Section 3 describes the used FS techniques. Section 4 demonstrates our experimental methodology then section 5 presents the results. Finally section 6 provides conclusion and future work.

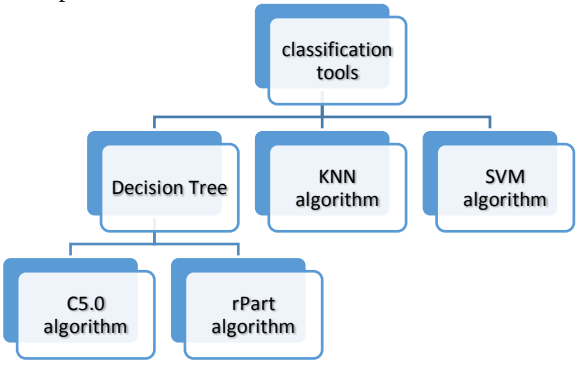


Figure 1: Classification tools.

II. CLASSIFICATION ALGORITHMS

A. K-Nearest Neighbor (KNN)

K-nearest neighbors is an algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions).A case is classified by a majority vote of its neighbors. If K = 1, then the case is simply assigned to the class of its nearest neighbor. The main advantages of this method are: a) it can estimate both qualitative attributes and quantitative attributes; b) It is not necessary to build a predictive model for each attribute with missing data, even does not build visible models. [5] The limitations of KNN imputation are: a) the choice of the distance function which could be Euclidean, Manhattan, Mahalanobis, Pearson, etc. b) The KNN algorithm searches through all the dataset looking for the most similar instances. This is a very time consuming process and it can be very critical in Data Mining (DM) where large databases are analyzed. c) The choice of k, the number of neighbors [6].

B. C5.0 Algorithm

C5.0 is new Decision Tree (DT) algorithm developed based on C4.5 by Quinlan. It includes all functionalities of C4.5 and apply a bunch of new technologies [7]. The classifier is tested first to classify unseen data and for this purpose resulting DT is used. C4.5 algorithm follows the rules of ID3 algorithm. Similarly C5 algorithm follows the rules of algorithm of C4.5. C5 algorithm has many features such as; i) the large DT can be viewing as a set of rules which is easy to understand; ii) C5 algorithm gives acknowledge on noise and missing data; iii) problem of over fitting and error pruning is solved by the C5 algorithm; and iv) in classification technique the C5 classifier can anticipate which attributes are relevant and which are not relevant in classification[8].

C. Support Vector Machine (SVM)

SVM classification technique analyzes data and recognizes patterns from them. SVM uses a very small sample set and generate pattern from that [8]. SVM represents a powerful technique for general (nonlinear) classification, regression and outlier detection with an intuitive model representation. It includes linear, polynomial, radial basis function, and sigmoidal kernels [9]. The main significance of the SVM is that it is less susceptible for over fitting of the feature input from the input items, this is because it is independent on feature space. SVM is fast accurate while training as well as during testing [10].

D. Recursive Partitioning and Regression Trees (rPart)

The rpart programs build classification or regression models of a very general structure using a two stage procedure; the resulting models can be represented as binary trees[11]. Recursive partitioning creates a DT that correctly classifies members of the population by splitting them into sub-populations. The process is termed recursive because each sub-population may in turn be split an indefinite number of times until the splitting process terminates after a particular stopping criterion is reached.

III. FEATURE SELECTION TECHNIQUES

Attribute selection methods can be broadly divided into filter and wrapper approaches. In wrapper approach the attribute selection method uses the result of the DM algorithm to determine how good a given attribute subset is. The major characteristic of the wrapper approach is that the quality of an attribute subset is directly measured by the performance of the DM algorithm applied to that attribute subset [12]. The advantages of wrapper approaches include the interaction between feature subset search and model selection, and the ability to take into account feature dependencies. Wrapper approach tends to be much slower, as the DM algorithm is applied to each attribute subset considered by the search. [3] Wrapper technique tend to be simpler than filter approach, more accurate and more computationally intensive[13]. Wrapper approach is dependent on the learning algorithm and has mainly three steps. Generation procedure generates or selects a candidate feature subset from the original feature space. Evaluation procedure evaluates the performance of the learning algorithm by using candidate feature subset. So, in this way the learning algorithm guide the search for feature subset. The validation procedure checks the suitability of the candidate feature subset by comparing it with other feature selection and generation method pairs [14]–[16]. In Filter approach the attribute selection method is independent of the DM algorithm to be applied to the selected attributes and assess the relevance of features by looking only at the intrinsic properties of the data [3]. Advantages of filter techniques are computationally simple and fast, and as the filter approach is independent on the mining algorithm so feature selection needs to be performed only once, and then different classifiers can be evaluated. Disadvantages of filter methods are that they ignore the interaction with the classifier which means that each feature is considered separately, thereby ignoring feature dependencies, which may lead to worse classification performance when compared to other types of feature selection techniques[12].

IV. IMPLEMENTATION METHODOLOGY

The dataset which is related to direct marketing campaigns of a banking institution is used from UCI Machine Learning Repository [17]. It contains 45211 instances and seventeen attributes. A brief description about dataset is described in Table 1.

Table 1: Dataset description

Attribute	Description	Value Range
Age	numeric	The age of the customer
Job	categorical	"admin.","unknown","unemployed","management","housemaid","entrepreneur

		","student", "blue-collar", "self-employed", "retired", "technician", "services".
marital	categorical	"Married", "divorced", "single".
education	categorical	"unknown", "secondary", "primary", "tertiary"
default	binary	Has credit in default? "yes", "no"
balance	numeric	average yearly balance
housing	binary	Has housing loan? "yes", "no"
Loan	binary	Has personal loan? "yes", "no"
contact	categorical	"unknown", "telephone", "cellular"
Day	numeric	last contact day of the month
month	categorical	"jan", "feb", "mar", ..., "nov", "dec"
duration	numeric	last contact duration in seconds
campaign	numeric	number of contacts performed during this campaign and for this client
Pdays	numeric	number of days that passed by after the client was last contacted from a previous campaign, -1 means client was not previously contacted
previous	numeric	number of contacts performed before this campaign and for this client
poutcome	categorical	outcome of the previous marketing campaign "unknown", "other", "failure", "success"
Y	binary	Has the client subscribed a term deposit? "yes", "no"

Randomly set with 10% sample is selected. The percentage of training set and test set is shown in figure 2. In the experiment the KNN algorithm classifies any new object based on a similarity function "distance function" which can be Euclidean, Manhattan, Minkowski or other. It measures how far a new object from its neighbors (the distance between the object and its neighbors) and the number of its neighbors is defined by K. EX: if k=3, so the KNN will search for the closest three neighbors to this object using the distance function and the predicted class of new object is determined by majority class of its neighbors.

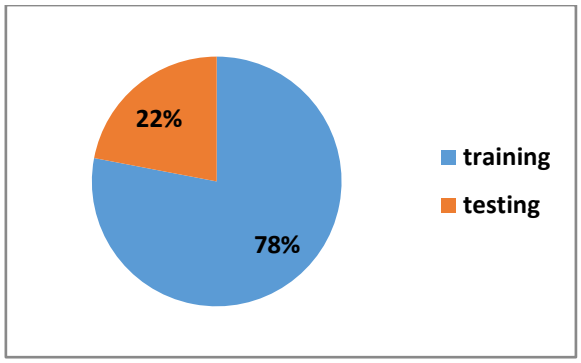


Figure 2: Training and testing set.

The SVM works by mapping data to a high-dimensional feature space so that data points can be categorized, even when the data are not otherwise linearly separable. A separator between the categories is found, and then the data are transformed in such a way that the separator could be drawn as a hyper plane. Following this, characteristics of new data can be used to predict the group to which a new record should belong.

The C5.0 and rPart are kind of decision tree which divides a dataset into smaller subsets. Leaf node represents a decision. Based on feature values of instances, the decision trees classify the instances. Each node represents a feature in an instance in a decision tree which is to be classified, and each branch represents a value. Classification of Instances starts from the root node and sorted based on their feature values[8].

V. EXPERIMENTAL RESULTS

1. Before Feature Selection

Based on the dataset in hand; the results revealed that C5.0 algorithm is the best to solve classification problem and KNN is the poorest algorithm to deal with classification problem. The performance of different methods was compared by calculating the average error rate and accuracy rate of each algorithm using a confusion matrix. The accuracy (AC) is the percentage of the total number of predictions that were correct. It is determined using the following equation:

$$AC = \frac{a+d}{a+b+c+d}$$

C5.0 algorithm was able to correctly predict that 845 clients won't subscribe the bank term deposit and 54 clients will subscribe the term deposit. rPart algorithm correctly predict that 854 clients won't subscribe the term deposit

and 34 clients will subscribe the term deposit. For SVM algorithm the correctly predicted clients that they won't subscribe the term deposit were 881 clients. Also for Knn algorithm 870 clients are correctly predicted that they won't subscribe the term deposit and 5 clients will subscribe it. More information for the correctly and incorrectly predicted records is shown in figure 3.

	Actual Class	Actual Prediction	
KNN		Positive	Negative
	Positive	870/1000 (87%)	11/1000 (1.1%)
	Negative	114/1000 (11.4%)	5/1000 (.5%)

3-a KNN.

	Actual Class	Actual Prediction	
C5.0		Positive	Negative
	Positive	845/1000 (84.5%)	36/1000 (3.6%)
	Negative	65/1000 (6.5%)	54/1000 (5.4%)

3-b C5.0.

	Actual Class	Actual Prediction	
rPart		Positive	Negative
	Positive	854/1000 (85.4%)	27/1000 (2.7%)
	Negative	85/1000 (8.5%)	34/1000 (3.4%)

3-c rPart.

	Actual Class	Actual Prediction	
SVM		Positive	Negative
	Positive	881/1000 (88.1%)	119/1000 (11.9%)
	Negative	0/1000 (0%)	0/1000 (0%)

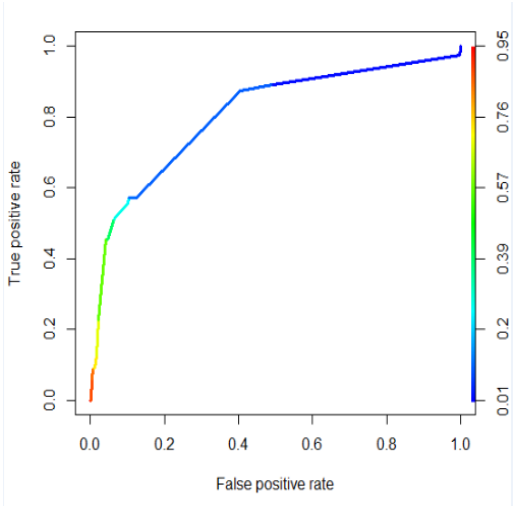
3-d SVM.

Figure 3: confusion matrix for the classifiers.

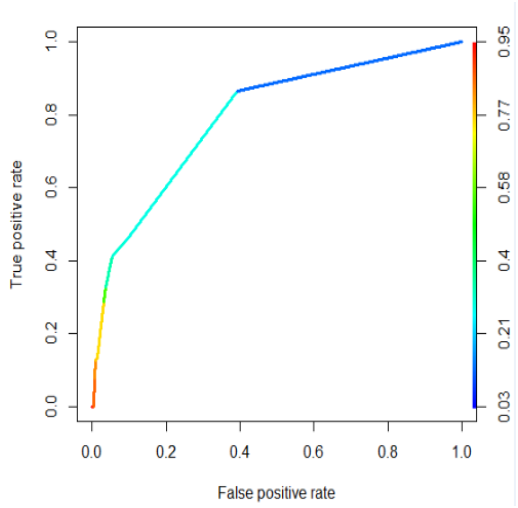
Roc curve is implemented for the four classifiers as in figure 4. The closer the curve to the left-hand border and then the top border of the ROC space, the more accurate the test. The Area Under ROC Curve (AUC) quantifies the overall ability of the test to differentiate between those who will subscribe the term deposit and those who won't. In the worst case scenario the area under ROC curve will be at 0.5 and in the best case scenario (one that has zero false positives and zero false negatives) the area will at 1.00. More details about the accuracy and error rate for each classifier are shown in table 2.

Table 2: Details of experimental results

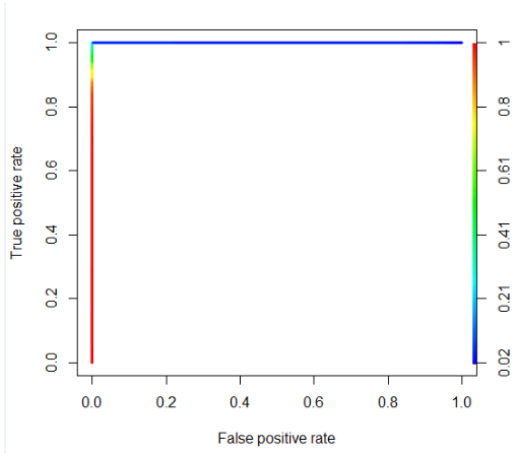
Classification Tool	**Accuracy Rate**	**Error Rate**
C5.0 Algorithm	89.9%	10.1%
rPart Algorithm	88.8%	11.2%
Support vector machine(SVM)	88.1 %	11.9%
K-nearest neighbor(KNN)	87.5%	12.5%



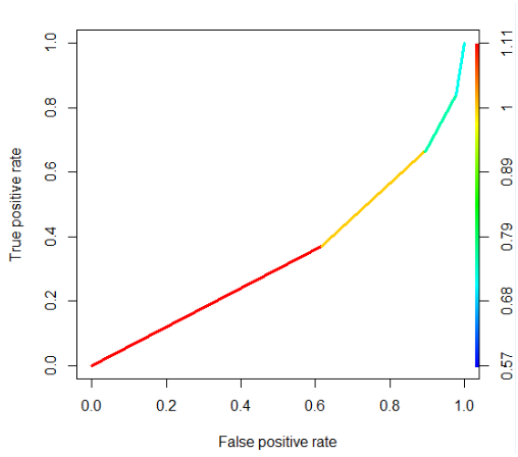
4-a: ROC Curve for C5.0.



4-b: ROC Curve for rPart.



4-c: ROC Curve for SVM.



4-d: ROC Curve for KNN

Figure 4: ROC Curve for the Algorithms.

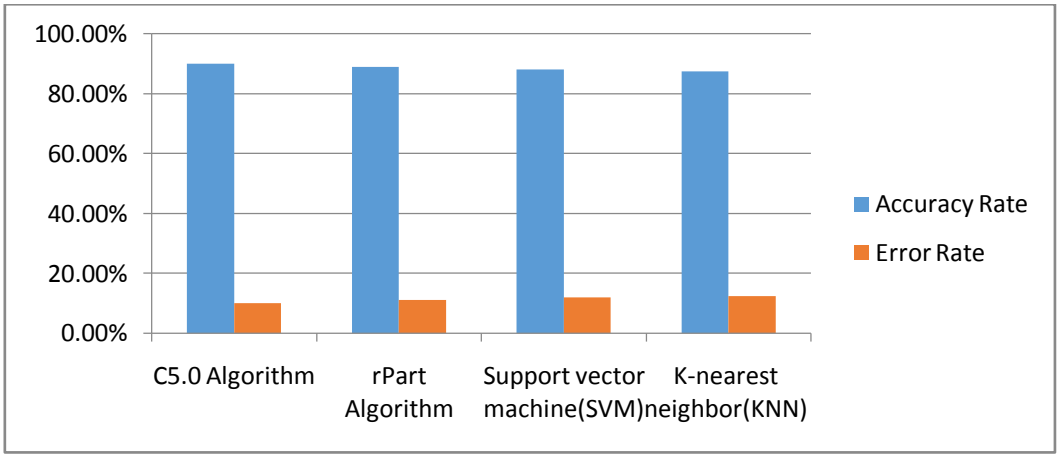


Figure 5: Comparison between algorithms according to accuracy rate

2. After Feature Selection

During this stage a Chi squared test and information gain techniques are used to filter the features, then the relevant important features are used in doing classification by our four classifiers. Also wrapper technique is used to extract the most important features and then the results are used to do classification by the four classifiers. A comparison is made between wrapper and filter techniques, with the selected features that result from both techniques. Accuracy rate for each classifier using relevant features from both techniques are shown in table 3. Based on this result we used wrapper technique to do feature selection as it gives higher results.

Table 3: Comparison between Wrapper and Filter Approach

Classification Tool	**Accuracy Rate**		
	Filter Approach		**Wrapper Approach**
	Chi-squared	*Info gain*	
SVM	94.7%	94.5%	96.7 %
C5.0 Algorithm	89.8%	89.2%	90.4%
rPart Algorithm	88.6%	89.4%	89.3 %
K-nearest neighbor(KNN)	89.1%	88.8%	88.9%

- **Wrapper Technique**

The experimental results revealed that SVM algorithm is superior to others to solve classification problem when cost = 10. Even when cost = 1 SVM gives impressive results the accuracy rate was (96.7 %) which is higher than the rest. Accuracy and error rate for each classifier is shown in table 4. A comparison is made between four classifiers before and after applying FS in table 5.

With applying FS technique the SVM algorithm correctly predicts all the records with zero false positives and zero false negatives giving 100 % as accuracy rate which means that there is an improvement in the performance of the classifier. Also there is an improvement in the performance of C5.0, rPart and Knn compared with their performance

before applying FS. C5.0 algorithm correctly predict 904 records out of 1000 records. rPart correctly predict 893 records out of 1000 records. And Knn was able to correctly predict 889 records out of 1000 record. Confusion matrix for the four classifiers with FS using wrapper technique is shown in figure 6.

	Actual Class	Actual Prediction	
SVM		Positive	Negative
	Positive	881/1000 (88.1%)	0/1000 (0%)
	Negative	0/1000 (0%)	119/1000 (11.9%)

6-a: confusion matrix for SVM.

	Actual Class	Actual Prediction	
C5.0		Positive	Negative
	Positive	853/1000 (85.3%)	28/1000 (2.8%)
	Negative	68/1000 (6.8%)	51/1000 (5.1%)

6-b: confusion matrix for C5.

	Actual Class	Actual Prediction	
rPart		Positive	Negative
	Positive	857/1000 (85.7%)	24/1000 (2.4%)
	Negative	83/1000 (8.3%)	36/1000 (3.6%)

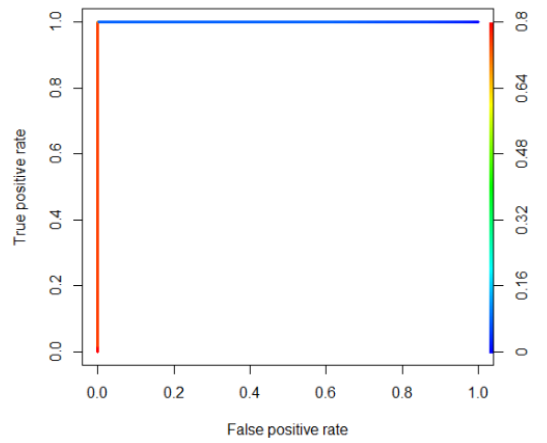
6-c: confusion matrix for rPart.

	Actual Class	Actual Prediction	
KNN		Positive	Negative
	Positive	879/1000 (87.9%)	2/1000 (.2%)
	Negative	109/1000 (10.9%)	10/1000 (1%)

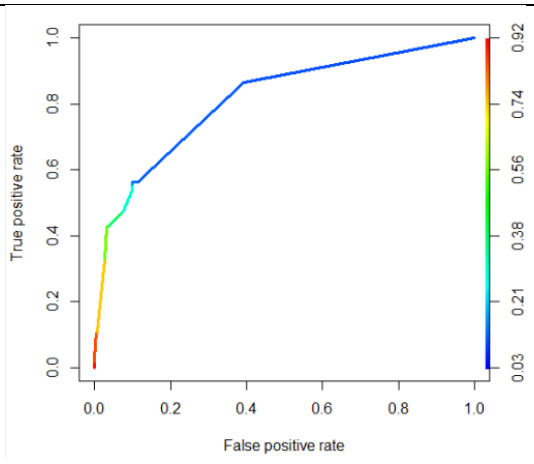
6-d: confusion matrix for KNN.

Figure 6: confusion matrix for the classifiers with FS.

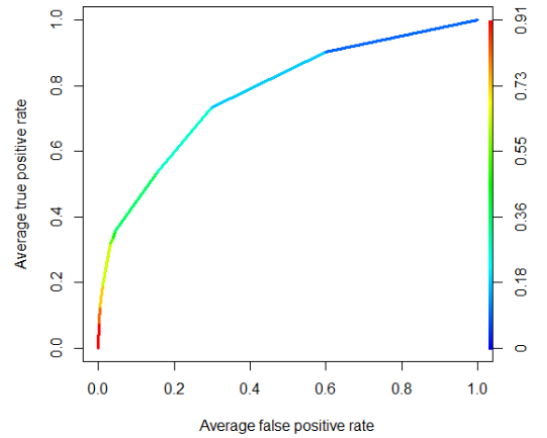
As known ROC curve is able to visualize the performance of the classifiers, the AUC is used to get the one with best performance. So as shown in figure 7 the AUC for SVM is at 1.0 which means higher performance compared to others. The AUC for KNN is under .5 which gives poor performance. For C5.0 and rPart the AUC is higher than .5 which give good performance.



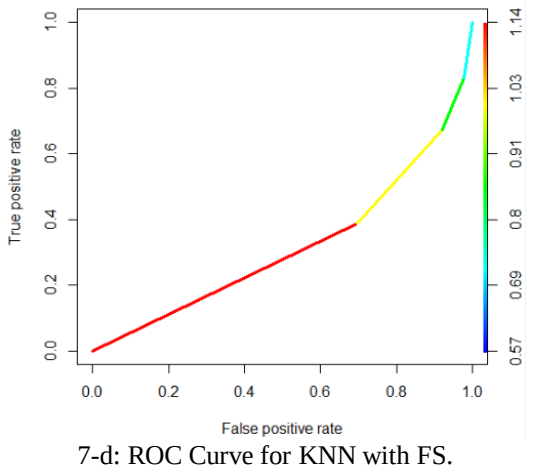
7-a: ROC Curve for SVM with FS.



7-b: ROC Curve for C5.0 with FS.



7-c: ROC Curve for rPart with FS.



7-d: ROC Curve for KNN with FS.

Figure 7: ROC Curve for the Algorithms with FS.

Table 4: Experimental results with FS.

Classification Tool	Accuracy Rate	Error Rate
Support vector machine(SVM)	100%	0%
C5.0 Algorithm	90.4%	9.6%
rPart Algorithm	89.3 %	10.7%
K-nearest neighbor(KNN)	88.9%	11.1%

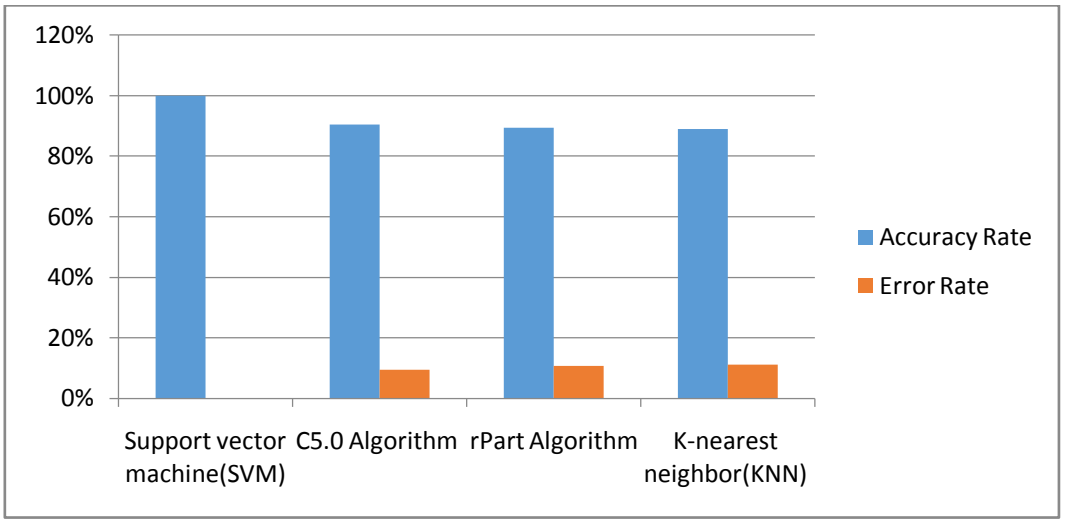


Figure 8: Comparison between algorithms with fs according to accuracy rate

Removing the irrelevant features from a dataset before doing any data mining has a great influence on the performance of the classifiers. Notably the accuracy rate of the four classifiers is increased and the error rate is decreased. In SVM the accuracy rate moved from 88.1% to 100% resulting in zero error rate. With C5.0 algorithm the accuracy rate moved from 89.9 % to 90.4% and the error rate reduced by .5%. With rpart the accuracy rate increased from 88.8% to 89.3 resulting in reducing the error rate by .5%. And in Knn the accuracy rate moved from 87.5% to 88.9% with error rate reduced by 1.4%. Summarization of the accuracy rate for the classifiers before and after using FS is shown in table 5.

Table 5: Accuracy rate before and After Wrapper Technique

Classification Tool	Accuracy Rate	
	Before FS	After FS
Support vector machine(SVM)	88.1 %	100%
C5.0 Algorithm	89.9%	90.4%
rPart Algorithm	88.8%	89.3 %
K-nearest neighbor(KNN)	87.5%	88.9%

VI. CONCLUSION

DM includes many tasks, classification task is one of them which can be solved by many algorithms. The data on which DM tasks depend may contain several inconsistencies, missing records or irrelevant features, which make the knowledge extraction very difficult. So, it is essential to apply pre-processing techniques such as FS in order to enhance its quality. In this paper we compared between the performance of SVM, C5.0, KNN and rpart before and after Feature Selection. From the obtained results we conclude that after implementing the FS it improves the data quality and the performance of the four classifiers, and SVM algorithm gives impressive and high results over other algorithms.

REFERENCES

[1] CC Aggarwal, Classification Algorithms and Applications Data, 2014.
[2] F. Voznika and L. Viana, “Data mining classification,” pp. 1–6, 1998.
[3] S. Beniwal and J. Arora, “Classification and Feature Selection Techniques in Data Mining,” vol. 1, no. 6, pp. 1–6, 2012.
[4] D. Tomar and S. Agarwal, “A Survey on Pre-processing and Post-processing Techniques in Data Mining,” vol. 7, no. 4, pp. 99–128, 2014.
[5] I. Technologies, “Missing Value Imputation in Multi Attribute Data Set,” vol. 5, no. 4, pp. 5315–5321, 2014.
[6] E. Acu, “The treatment of missing values and its effect in the classifier accuracy,” no. 1995, pp. 1–9.

[7] P. Su-lin and G. Ji-zhang, "C5 . 0 Classification Algorithm and Application on Individual Credit Evaluation of Banks," Syst. Eng. - Theory Pract., vol. 29, no. 12, pp. 94–104, 2009.

[8] R. Pandya, "C5 . 0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning," vol. 117, no. 16, pp. 18–21, 2015.

[9] D. Meyer, "Support Vector Machines," vol. 1, pp. 1–8, 2015.

[10] I. J. Of, "Research in Computer Applications and Robotics a Survey on Trust Based," vol. 4, no. 4, pp. 55–58, 2016.

[11] T. M. Therneau and E. J. Atkinson, "An Introduction to Recursive Partitioning Using the RPART Routines," pp. 1–62, 2015.

[12] M. L. Raymer, E. D. Goodman, L. A. Kuhn, and A. K. Jain, "Dimensionality Reduction Using Genetic Algorithms," vol. 4, pp. 164–171, 2000.

[13] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," Bioinformatics, vol. 23, no. 19, pp. 2507–2517, 2007.

[14] H. Hsu, C. Hsieh, and M. Lu, "Expert Systems with Applications Hybrid feature selection by combining filters and wrappers," Expert Syst. Appl., vol. 38, no. 7, pp. 8144–8150, 2011.

[15] R. Jensen, "Combining rough and fuzzy sets for feature selection Doctor of Philosophy School of Informatics University of Edinburgh," 2005.

[16] D. Tomar and S. Agarwal, "Twin Support Vector Machine Approach for Diagnosing Breast Cancer , Hepatitis , and Diabetes," vol. 2015, 2015.

[17] http://archive.ics.uci.edu/ml/datasets/Bank+Marketing

