Feature Selection based Classification using Naive Bayes, J48 and Support Vector Machine

Dipali Bhosale
Dr. D. Y. Patil School of Engg. & Tech., Savitribai Phule Pune University, India

Roshani Ade
Dr. D. Y. Patil School of Engg. & Tech. Savitribai Phule Pune University, India

ABSTRACT
One way to improve accuracy of a classifier is to use the minimum number of features. Many feature selection techniques are proposed to find out the most important features. In this paper, feature selection methods Co-relation based feature Selection, Wrapper method and Information Gain are used, before applying supervised learning based classification techniques. The results show that Support vector Machine with Information Gain and Wrapper method have the best results as compared to others tested.

General Terms
Feature selection; supervised learning

Keywords
Naïve Bayes; SVM; J48; Correlation Feature Selection; Information Gain; wrapper method

1. INTRODUCTION
In recent research, use of machine learning techniques in data mining has increased. This task of knowledge discovery with the help of a machine learning technique called as supervised learning. In supervised learning class labels are assigned to each and every tuple in training data, this labeled training data is used for deriving a function [1, 2]. This function further can be used for mapping new example. When feature selection is applied before supervised learning it increases the accuracy of classification.

In feature selection, selection of most distinct feature is done. The goal of this technique is to remove redundant and irrelevant features [3, 4]. Due to this dimensionality of the original data set is reduced which results in the efficient performance of the classifier. When features are selected before applying data mining algorithm using some independent approach it is called as Filter method for feature selection. In Wrapper approach, best subset of features is selected targeting data mining algorithm. In this paper, both approaches are used for selecting a subset of the feature. Impact of feature selection for supervised learning can be analyzed by comparing performance of different classification methods.

Naive classifier uses statistical as well as a supervised learning method for classification. It is based on application of Bayes theorem with naive independence assumptions. J48 is used C4.5 decision tree for classification which creates a binary tree. Decision trees are constructs using greedy technique and it uses reduced error pruning. Support Vector Machine (SVM) algorithm is based on statistical learning theory. It shows decision boundary, using a subset of training examples called as "support vectors". The core concept behind SVM is a maximum margin hyperplane, which guarantees maximum separation between two or more classes.

The rest of this paper is organized as follows. Section 2 gives brief idea about different feature selection techniques. Section 3 includes an overview of methodology and data sets used for classification. Section 4 deals with the experimental results. Section 5 includes conclusion with its future scope.

2. FEATURE SELECTION METHOD
Feature selection is one of the dimensionality reduction technique used in data mining. It is often used as data preprocessing method before applying to any classification algorithm. This reduces high dimension data by selecting useful attributes only. Redundant and irrelevant features are omitted while doing feature selection. There are three standard approaches: Embedded, Filter and Wrapper. In the embedded approach algorithm, it decides which approaches are used. A wrapper approach uses target learning algorithms to find relevant feature subset, while in filter approach features are selected before applying a learning algorithm. In this paper, wrapper and filter (Information Gain (IG) and Correlation Feature Selection (CFS)) approaches are used in experiments.

2.1 Correlation Feature Selection
The Correlation Feature Selection (CFS) method selects a subset of features which are highly co-related to class [5]. In each subset attribute are selected by considering the degree of redundancy between them and predictive ability of each individual feature. A function that evaluates best individual feature is:

$$Merit(S_K) = \frac{\mathrm{K} * \mathrm{cr_{fc}}}{\sqrt{K + K * (K-1) * cr_{ff}}}$$

Where, Merit is the heuristic merit of feature subset S containing K feature, $\mathrm{cr_{fc}}$ is the average feature class co-relation and $\mathrm{cr_{ff}}$ is average feature-feature co-relation. The numerator shows the predictive ability of features and the denominator indicates the degree of redundancy between the features.

2.2 Information Gain (IG)
Information Gain [6] calculates the entropy value (i.e. how much information it is giving), for each feature. Entropy is a measure of the uncertainty associated with a random variable. With the help of this value we can determine most useful feature for classification. Higher the entropy value, the feature contains more information. As the data become purer, the entropy value becomes smaller. If the target attribute can take on k different values, then the entropy of the feature relative to this k - wise classification is defined as:

$$\text{Entropy(S)} = \sum_{c=1}^{k} - [\text{p}_\text{c} * (\log_2 * \text{p}_\text{c})]$$

Where, P_C is a proportion of S belongs to class c. As encoding length is measured in bits logarithm has its base 2.

$$IG(S,a) = \text{E(S)} - \text{E(S|a)}$$

where, a is variable value.

The above equation calculates Information Gain that training example S obtains from observation that a random variable V takes some value a.

2.3 Wrapper Method

The wrapper is technique for selecting best subset of features using a specific classification algorithm [7]. The difference between embedded and the wrapper approach is that the wrapper has internal cross validation while embedded is not having. It uses a targeted data mining algorithm as a black box to select a best feature subset. This method takes into account feature dependencies while searching and building a model.

The evaluation function used here is five-fold cross validation. The search is conducted using possible parameters. The goal of search method is finding the state which is having maximum evaluation. In this experiment best first search method is used.

3. METHODOLOGY USED

Fig.1 illustrates the overall flow of the experiment. First, classification results are noted without doing any kind of feature selection techniques (Co-relation based Feature Selection, Wrapper, and Information Gain) on data sets. Then, using three feature selection techniques, separate feature subsets are chosen for each technique. The selected features are passed to the classifiers and results are noted.

In this paper, Naïve Bayes (NB), J48 and Support Vector Machine (SVM) classifiers are used for the classification of different data sets (Iris and Glass). The data is preprocessed by using different feature selection techniques, namely, IG, Wrapper and CFS.

3.1 Naive Bayes

The Naïve Bayes classifier is based on Bayesian probability model. If a class is provided, Naïve Bayes classifier assumes that the value of one feature is independent of any other feature [8, 9]. It is based on the mathematical principle of Conditional probability. If n attributes are given, independent assumptions made by the Naïve Bayes classifier is 2n!. A Conditional probability model for above classifier is given as:

P (Ci | x)

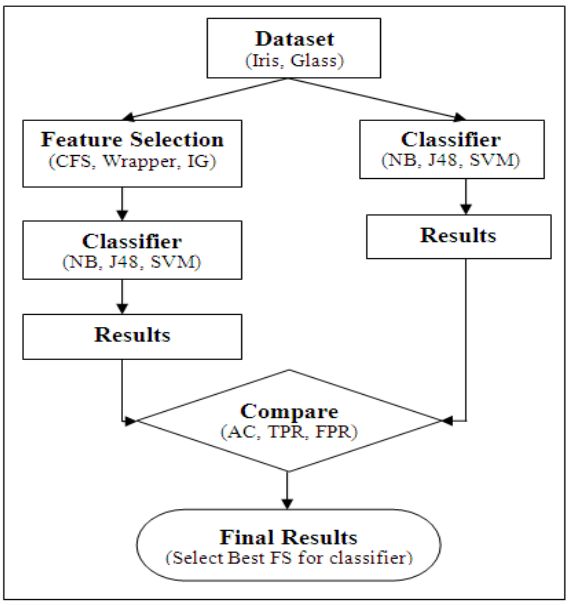

Fig 1: System Architecture

where, C_i is the ith class and x is input vector.

In this case, class variable C is conditional on several features variable x = x1,……., xn

Using Bayes theorem equation (1) can be written as:

$$\text{P}(\text{C}_\text{i} \,|\text{x}\,) = \frac{\text{P}(\text{C}_\text{i}) * \text{P}(\text{x} \mid \text{C}_\text{i})}{\text{P}(\text{x})}$$

Constructing a classifier from probability model:

For classification purposes Naïve Bayes classifier combines above model with decision rule. The commonly used rule is the maximum a posteriori or MAP decision rule. This rule selects the hypothesis which is most probable.

3.2 J48

In WEKA data mining tool J48 is implementation of C4.5 algorithm [10, 11]. C4.5 builds decision trees with the help of information entropy. At every node of the tree, attribute is selected which is most effectively splitting itself into multiple subset. Splitting is done based on Information Gain (IG) value. For decision making, the attribute with highest normalized IG is used. This algorithm has the limitation of handling numeric data only.

3.3 Support Vector Machine

It is a classification technique based on statistical learning theory (SLT) [12]. It uses support vectors [13] to represent decision boundaries. It finds number of support vectors that represent the training data. The only portion of data is used to train the model. The SVM [14, 15, 16, 17, 18] is originally designed for binary classification. The multiclass classification problem can be solved by the serial combination of binary classifier. Linear classification requires mainly largest classification interval, i.e. maximum margin hyperplane [19] as shown in Fig.2.

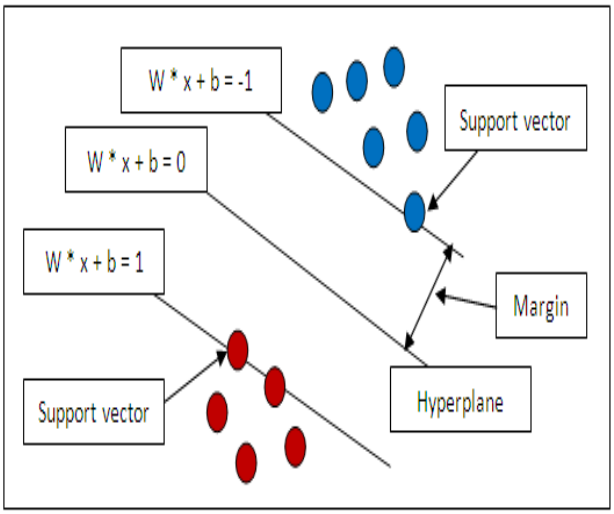

Fig 2: Basic idea for SVM

Maximum margin hyperplane can be obtained by minimizing the following function:

$$min = \frac{1}{2} \; || \, w^2 ||$$

Subject to: $y_I \, (w * x_I + b) \geq 1, \quad I = 1,2,3,\ldots\ldots\ldots\ldots..,n$

$$L = \tfrac{1}{2} \; || \, w^2 || \; - \sum_{i=1}^{n} \lambda_i \, [y_i (w \; * x_i \; + \; b) - 1]$$

where, λ_i is Lagrange multiplier

Problem having non-linear decision boundary can be solved by transforming data from original co-ordinate space into a new space ø(x) to solve it as a linearly separable case.

This transforms space has the following form of linear decision boundary:

$$w. \, ø(x) + b = 0$$

Mapping is done by introducing kernel function K. It is the dot product of two vectors x_i and x_j as shown below:

$$K(x_I, x_j) = ø(x_i) * ø(x_j)$$

3.4 Data Set Used

Data sets from UCI Machine Learning Repository are used for training purpose. In this paper, two standard data sets are used for classifications which are Iris and Glass [20, 21].

3.4.1 Iris

The Iris data set contains three classes each having 50 tuples with four attributes. Three classes are: Iris Setosa, Iris Versicolour, IrisVirginica. In this data set, Iris Setosa is linearly separable from other two, while remaining two are non-linearly separable from each other. This dataset is available online at UCI machine learning repository.

3.4.2 Glass

The Glass data set is intended towards the study of criminological investigation. It includes 11 attributes and is a multiclass type of data set. It has seven different classes according to type of glass.

4. EXPERIMENTAL RESULTS

As per the methodology discussed earlier, experiments are performed on two data sets with and without feature selection. Analysis of results is done using following evaluation metrics.

4.1 Evaluation Measure

There various parameters to measure the performance, among them only Accuracy, True Positive Rate (TPR), False Positive Rate (FPR) are considered in this paper. Accuracy is the proportion of the total number of predictions that were correct. TPR gives the proportion of correctly classified instances out of total classified. FPR shows the proportion of negative cases that were incorrectly classified as positive.

$$AC = \frac{TN+TP}{TP+TN+FP+FN} \tag{11}$$

$$TPR = \frac{TP}{TP+FN} \tag{12}$$

$$FPR = \frac{FP}{FP+TN} \tag{13}$$

To compute these metrics, first confusion matrix for the data set is then computed using these values into above equations to find Accuracy, TPR, and FPR.

4.2 Without Feature Selection

Iris and Glass data sets are used for training three different classifiers which are Naive Bayes (NB), J48 and Support Vector Machine (SVM). The obtained results are as shown in Table 1. Results are compared based on the above mentioned evaluation measures.

Table 1. Performance of classifiers for Iris data set

Classifier	Accuracy	TPR	FPR
NB	96	0.96	0.02
J48	96	0.96	0.02
SVM	**96.66**	0.967	0.01

Table 2. Performance of classifiers for Glass data set

Classifier	Accuracy	TPR	FPR
NB	48.59	0.48	0.18
J48	66.82	0.66	0.13
SVM	**68.69**	0.68	0.14

From Table 1, 2 is clear that, in case of the Iris dataset SVM is giving better results (96.66%) than other two classifiers. In case of Glass data set also SVM is showing more Accuracy than other two. To improve the results feature selection techniques are applied to both data sets.

4.3 With Feature Selection

Here, results after applying feature selection techniques are compared using three different classifiers.Both the data sets are undergone through CFS, Wrapper and IG feature selection techniques.

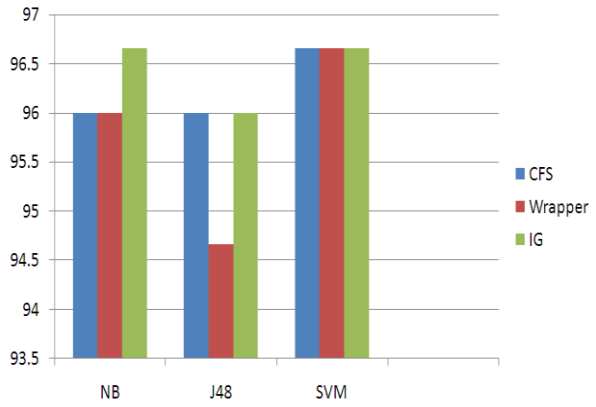

Fig. 3. Accuracy of classifier using FS techniques for Iris data set

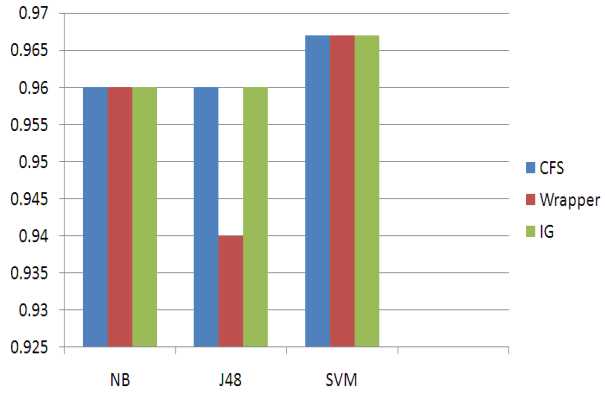

Fig. 4. TPR of classifier using FS techniques for Iris data set

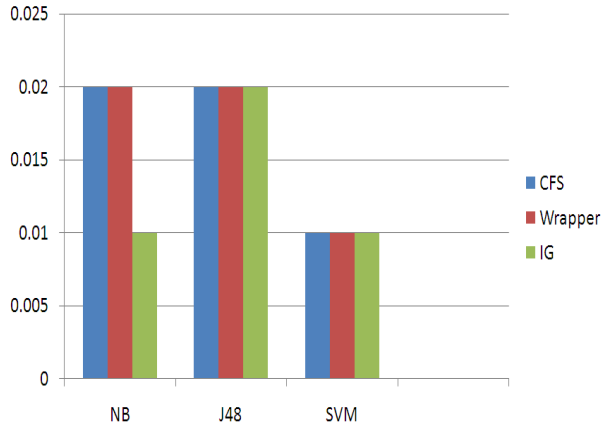

Fig. 5. FPR of classifier using FS techniques for Iris data set

Figure 3-4 gives the results of classifier after applying feature selection. From the table it is clearly observed that FS improve the classifier's result. SVM is giving best AC, TPR, and FPR than other two.

Wrapper FS includes a classifier oriented selection of feature; it is time consuming than other FS techniques. Figure 6-8 provides results on glass data set after applying FS.

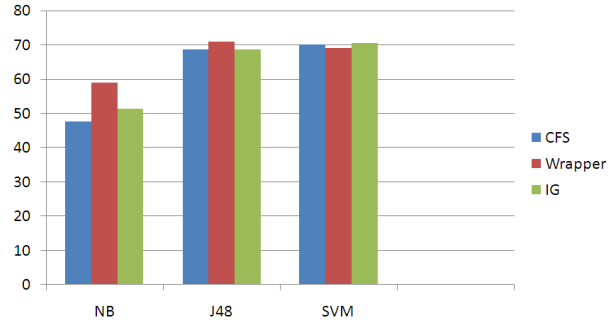

Fig. 6. Accuracy of classifier using FS techniques for Glass data set

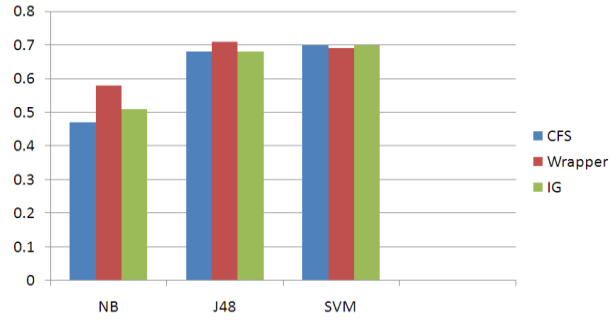

Fig. 7. TPR of classifier using FS techniques for Glass data set

From the figure it is observed that SVM is giving higher TPR with improved accuracy. This means more number of attributes are correctly classified with minimum misclassification.

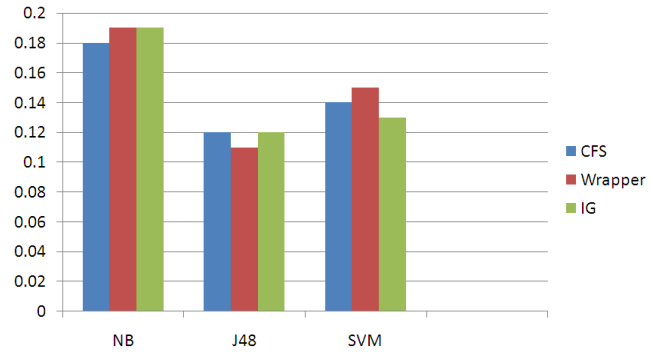

Table 8. FPR of classifier using FS techniques for Glass data set

Table 3 is describing feature selected by different FS techniques. CFS and IG are improving classifier performance by selecting minimum no. of attributes. Due to this time complexity gets reduced and TPR is increased.

Table 3. Selected features for Iris data set

Classifier	CFS	Wrapper	IG
NB	3,4	3,4	3
J48	3,4	4	3,4
SVM	3,4	1,2,3,4	3,4

For glass data set feature selected are as shown in Table 4. Observation from table concludes that wrapper method is selecting less no. of features as well as improving accuracy and TPR of all three classifiers.

Table 4. Selected features for Glass data set

Classifier	CFS	Wrapper	IG
NB	1,2,3,4,6,7,8,9	1,4	3,4,6,7
J48	1,2,3,4,6,7,8,9	1,3,4,6,8	2,3,4,6,7,8
SVM	1,2,3,4,6,7,8,9	2,4,5,7	2,3,4,6,7

5. CONCLUSION AND FUTURE WORK

In this paper, effect of feature selection on supervised learning based classifiers is compared. Accuracy, TPR and FPR are used as an evaluation metric for comparison. From the experimental results it has been observed that Information Gain and Wrapper method improves accuracy and True Positive Rate and minimizes False Positive Rate.

The future work will include combining different classifier using ensemble method and applying feature selection technique before classification.

6. REFERENCES

[1] Shengyan Zhou, Iagnemma K., "Self-supervised learning method for unstructured road detection using Fuzzy Support Vector Machines",International Conference on Intelligent Robots and Systems, pp. 1183–1189, IEEE, 2010.

[2] Techo, J. Nattee, C. Theeramunkong, T., "A Corpus-Based Approach For Keyword Identification Using Supervised Learning Techniques", 5th International Conference On Electrical Engineering/Electronics, Computer, Telecommunications And Information Technology, Ecti-Con, Vol.1, pp. 33 – 36, 2008.

[3] Cecille Freeman, Dana Kuli´Cand OtmanBasir, "Feature-Selected Tree-Based Classification", IEEE Transactions on Cybernetics, Vol. 43, No. 6, pp. 1990-2004, December 2013.

[4] Wald R., Khoshgoftaar T.M., Napolitano A., "Stability Of Filter- And Wrapper-Based Feature Subset Selection", IEEE 25th International Conference On Tools With Artificial Intelligence, pp. 374 – 380, 2013.

[5] Dr.Saurabh Mukherjee, Neelam Sharma, "Intrusion Detection Using Naive Bayes Classifier With Feature Reduction", Elsevier, pp. 119 – 128, 2012.

[6] Amirasayed A. Aziz, Ahmad Taherazar, Mostafa A. Salamaand Sanaa El-Ola Hanafy, "Genetic Algorithm With Different Feature Selectiontechniques For Anomaly Detectors Generation", IEEE Federated Conference On computer Science And Information Systems, pp. 769–774, 2013.

[7] Altidor W.,Khoshgoftaar T.M., Van Hulse J., "An Empirical Study On Wrapper-Based Feature Ranking", 21st International Conference On tools With Artificial Intelligence, pp.75-82, IEEE 2009.

[8] Yuguang Huang,Lei Li, "Naive Bayes classification algorithm based on small sample set", IEEE International Conference on Cloud Computing and Intelligence Systems, pp. 34 – 39, 2011.

[9] GuoQiang,"An Effective Algorithm for Improving the Performance of Naive Bayes for Text Classification", IEEE Second International Conference on Computer Research and Development, pp. 699–701, 2010.

[10] Menkovski, V.,Efremidis, S., "Oblique Decision Trees using embedded Support Vector Machines in classifier ensembles", 7th IEEE International Conference on Cybernetic Intelligent Systems (CIS), pp. 1-6, 2008.

[11] Mohamed W.N.H.W., Salleh, M.N.M., Omar, A.H.,"A comparative study of Reduced Error Pruning method in decision tree algorithms", IEEE International Conference on Control System, Computing and Engineering, pp. 392 – 397, 2012.

[12] HaoWu, Xianmin Zhang, HongweiXie, Yongcong Kuang, "Classification of Solder Joint Using Feature Selection Based on Bayes and Support Vector Machine", IEEE Transactions On Components, Packaging and Manufacturing Technology, Vol. 3, No. 3,pp.516-522, March 2013.

[13] A.M.Chandrasekhar and K.Raghuveer, "Intrusion Detection Technique by using K-means, Fuzzy Neural Network and SVM classifiers", IEEE International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, 2013.

[14] Vasilis A. Sotiris, Peter W. Tse, and Michael G. Pecht, "Anomaly Detection Through a Bayesian Support Vector Machine", IEEE Transactions On Reliability, Vol. 59, No. 2,pp.277-286, June 2010.

[15] PreechaSomwangand Woraphon Lilakiatsakun, "Computer Network Security Based On Support Vector Machine Approach", IEEE 11th International Conference on Control, Automation and Systems, pp.155-160, 2011.

[16] Mrs. Snehal A. Mulay, Prof. P. R. Devale, "Decision Tree based Support Vector Machine for Intrusion Detection", IEEE International Conference on Networking and Information Technology, pp. 59-63, 2010.

[17] Kurt Kramer, Dmitry B.Goldgof, Lawrence O. Hall and Andrew Remsen, "Increased Classification Accuracy and Speedup Through Pair-wise Feature Selection for Support Vector Machines", IEEE, 2011.

[18] Ajay Urmaliya, Dr. JyotiSinghai, "Sequential Minimal Optimization for Support Vector Machine with Feature Selection in Breast Cancer Diagnosis", IEEE Second International Conference on Image Information Processing (ICIIP), pp.481-486, 2013.

[19] Changjing Shang and Dave Barnes, "Support Vector Machine-Based Classification of Rock Texture Images Aided by Efficient Feature Selection", WCCI 2012 IEEE World Congress on Computational Intelligence, June2012.

[20] https://archive.ics.uci.edu/ml/data sets/Iris

[21] http://archive.ics.uci.edu/ml/dataset/Glass+Identification