Empirical study of feature selection methods over classification algorithms

N. Bhalaji*, K.B. Sundhara Kumar and Chithra Selvaraj

Department of Information Technology, SSN College of Engineering, Chennai, 603110, India
Email: bhalajin@ssn.edu.in
Email: sundharakumarkb@ssn.edu.in
Email: chithras@ssn.edu.in
*Corresponding author

Abstract: Feature selection methods are deployed in machine-learning algorithms for reducing the redundancy in the dataset and to increase the clarity in the system models without loss of much information. The objective of this paper is to investigate the performance of feature selection methods when they are exposed to different datasets and different classification algorithms. In this paper, we have investigated standard parameters such as accuracy, precision and recall over two feature selection algorithms namely Chi-Square feature selection and Boruta feature selection algorithms. Observations of the experiments conducted using R studio resulted around 5–6% increased performance in above said parameters when they were exposed to Boruta feature selection algorithm. The experiment was done on two different datasets with different set of features and we have used the following five standard classification algorithms – Naive Bayes, decision tree, support vector machines (SVM), random forest and gradient boosting.

Keywords: classification; feature selection; Boruta; Chi-square; ensemble classifiers.


Biographical notes: N. Bhalaji is an Associate Professor of Information Technology, has over 13 years of teaching experience. He received his BE and ME both in the discipline of Computer Science and Engineering and PhD specialising in Trust Based Routing approach for MANET’s from Anna University, Chennai. His current research interests include security in distributed systems and wireless mobile network. He has published over 25 research papers in refereed journals in the areas of ad hoc networks, soft computing and internet of things. He is an active member in Professional Bodies such as IEEE, ACM, IACSIT, IAENG, and ACEEE. He is an active reviewer for reputed publishers like Elsevier, IEEE and Springer.

K.B. Sundhara Kumar is a Junior Research Fellow at SSN College of Engineering. He is currently pursuing his PhD at Anna University, Chennai. He received his BTech and MTech both in the discipline of Information Technology. He has published over five research papers in the areas of data
Empirical study of feature selection methods over classification algorithms

mining, big data, cloud computing and internet of things. His research interests include data analytics, data mining and machine learning.

Chithra Selvaraj is an Associate Professor in SSN College of Engineering, Chennai has over 18 years of teaching experience. She has completed her Bachelor of Engineering degree in Computer Science and Engineering from Bharathiar University with University Sixth rank. She received her Masters’ degree in Computer Science and Engineering from Sathyabama University. She has received her PhD from Anna University, Chennai and indulged in research work in the Networks domain. She is a recognised Supervisor in Anna University. She has received ‘Best Paper’ award in the conference ICCNT 2009, Singapore. She has published over 30 research articles in refereed International and National journals. She is IEEE Senior member.

This paper is a revised and expanded version of a paper entitled ‘Empirical study of feature selection methods over classification algorithms’ presented at International Conference on Innovative Systems, Bangalore, India, 16–17 December, 2016.

1 Introduction

Feature selection is the mechanism of selecting a subset of features from the given set of features in order to reduce the redundant and irrelevant features without much loss of the facts and details according to Dash and Liu (2003), Blum and Langley (1997) and Guerra-Salcedo and Whitley (1999). Classification algorithms find itself in a large scope of pattern recognition problems. For such problems, the dataset may contain a wide variety of features, few of which may be redundant or irrelevant to that problem. Ensemble classifiers have shown that the combination of various classifiers along with the relevant feature selection methods provides greater accuracy in predictive analytics. Another striking advantage of using these feature selection methods is that the time used for training the model is drastically reduced because of the limited number of features being used.

There are three main categories of feature selection techniques as – wrappers, filters and embedded methods. Wrapper methods use black box techniques to score the subsets of variables according to predictive power. Filter methods perform the variable selection as a preprocessing irrespective of the chosen predictor. Embedded methods, however, select the variables as a training process and they depend on the learning machines.

Rudnicki (2016) studied the impact of feature selection for failure analysis of parameter-induced simulation crashes in climate models. The study was made on several parameters which were weakly constrained that can potentially lead to simulation crashing. Feature selections can be useful for predicting which attributes can act as a potential threat that leads to simulation crashing.

The objective of this work is to provide a detailed analysis of two feature selection methods – Chi-square and Boruta methods applied over five classification methods namely, Naive Bayes, decision tree (J48), support vector machines (SVM), random forest and gradient boosting method. US census data and Iris data are the two different datasets taken from the UCI machine-learning repository for calculating accuracy, precision and recall values to provide an in-depth analysis of these feature selection methods.
2 Related works

Feature selection methods are considered as one of the most important phases in machine-learning techniques. There are a lot of literatures available for feature selection methods such as those mentioned in Kantarcioglu and Clifton (2004), Setiono and Liu (1997), etc., as they are very conventional and long established. It finds its applications in various areas not limited to signal processing, statistics, classification techniques suggested by Zheng et al. (2004) and clustering techniques, machine learning as discussed by Canuto et al. (2006). Of late in machine learning, feature selection with ensembles described by Santana et al. (2007), Caragea et al. (2003) is gaining high attention. Feature selection methods reduce the training time significantly. Without losing much information, it removes the unnecessary and redundant data.

Vaidya and Clifton (2002) propounded that selecting subsets of variables can be classified into three categories namely wrappers, filters and embedded methods. The wrapper methods deploy a black box method for feature extraction. The method proposed by Guyon and Elisseeff (2003) defines the search space for all possible variable subsets, assessing the performance of the learning machine to guide the search and halt and finally which predictor to be used. The problem of multi-variate inputs and over fitting of data persisted which was solved by using any predictor of the user’s choice and using a variable ranking method using correlation coefficient or with a nested subset selection that uses forward feature selection or backward feature selection.

Pehlivan et al. (2014) compared Bayesian classification, classification and regression trees (CART), decision tree and sequential minimal optimisation classification algorithms over gain ratio, RelieF, Cfs subset evaluator and consistency subset evaluator. The implementation had over 3500 Android apk files and the analysis of permission-based android malware detection. The evaluation methods comprised of measures such as overall accuracy, true-positive rate, false-positive rate and precision. Their findings revealed a higher performance for random forest and decision tree method for most of their feature selection methods.

Application of feature selection algorithms on supervised learning techniques is called as the sequential feature selection algorithms as advocated by Aha and Bankert (1996). The analysis was done on forward sequential selection and backward sequential selection. The work reports the positive empirical results on the application of these techniques. For datasets with minimal features, backward method mostly had outperformed the forward sequential selection technique. The variants of backward sequential selection (BSS) and forward sequential selection (FSS) were studied for finding a cloud pattern classification that had a sparse dataset with many attributes. The authors had used K-nearest neighbour method as the classifier and suggested that BSS does not always outperform FSS technique.

K-nearest neighbour is a simple non-parametric relatively mature pattern selection method. Zhang et al. (2017) proposed a two-stage plant species recognition combining local K-nearest neighbours and weighted sparse representation. The plant species recognition is a very tough phenomenon. The authors had used K-nearest neighbour for finding out the species that are closely similar to the training samples.

Extracting complex spatial and temporal patterns from a noisy multi-dimensional time series of electro encephalo graphy (EEG) data were found to be extremely complicated as studied by Garrett et al. (2003). They had used feature selection based on genetic algorithms to obtain preliminary results for the classification of EEG using finger
Empirical study of feature selection methods over classification algorithms

movements. This work also uses the brain computer interfaces (BCI) since the standard input devices like keyboard and mouse could not be used efficiently by disabled persons. The authors had implemented SVM for the classification of EEG and compared the results with the neural networks classifiers. The preliminary results were limited to a single dataset and two classification methods.

To find the subset of features used by the classification algorithms, the feature selection methods are applied in ensembles. In many of the previous works, the researchers used single classifiers for the pattern recognition problem. Santana et al. (2007) had used six different feature selections over six different methods to carry out their experiment. They had used ensembles of classifiers such as those provided in the literature of Ho (1998), Rodriguez et al. (2006) and Tumer and Oza (2003).

The accuracy of the ensembles was found to be strongly affected by the combination of classifiers used. They had also used non-random feature selection algorithms and genetic algorithms. Different configurations were used for the different combinations of classifiers. Ensembles with three and nine classifiers were used. They concluded that selection-based methods were affected most by using feature selection algorithms than the fusion-based techniques.

The classification where the features are distributed among the agents which cannot be combined due to privacy issues is considered in the experimentation by Modi and Kim (2005). A distributed asynchronous decision tree algorithm was proposed which differed in the fact that it can be applied to vertically partitioned data that had multiple categorical features. Since the proposed algorithm is asynchronous in nature, it is a more apt algorithm for the agents to communicate in distributed domain. Information gain is considered as a parameter for the discrete diffraction transform (DDT) algorithm proposed in this work. The performance in the distributed meetings was analysed and the predictions were accurate as a group rather than individual and the privacy of agents are also maintained using the information gain parameter.

Kursa (2014) studied the influence of gene selection in micro array data analysis as it can give fruitful information. Since the nature of the micro-array data is very noisy, it is not very simple to perform the gene selection. The results showcased how robust were the random forest methods on the noisy micro-array data. Boruta algorithm was used as feature selection method in order to remove the unwanted features present in the data in order to perform gene selection. The author concluded that the use of Boruta feature selection algorithm helped to identify the false positives in the noisy data.

3 Experimental setup

The experimental setup for this work comprises of two feature selection algorithms namely, Chi-square feature selection method and Boruta wrapper feature selection algorithm. These algorithms are applied to two different datasets, US census dataset and Iris dataset, taken from the UCI machine-learning repository. After the feature selection step is done, the five classification algorithms under consideration are applied and the performance metrics like accuracy, precision and recall are calculated. The implementation is carried our using R, which is a free software for statistical computing. R Studio is an integrated development environment (IDE) built using R which is used as the programming environment for this analysis.
A basic feature selection algorithm to pick up $k$-best features is given in Figure 1 as suggested by Manning et al. (2008).

**Figure 1** Algorithm to pick best $k$-features

```
SELECTFEATURES(D,C,K)
V ← EXTRACTVOCABULARY(D)
L ← []
for each $t$ ∈ $V$
do $A(t,C) ← COMPUTEFEATUREUTILITY(D,t,C)$
```

In the above algorithm, for a given class $C$, first the utility measure $A(t, C)$ for every term of the vocabulary is calculated. Then the $k$ terms that have the highest value of the utility measure is selected. All the other terms are discarded and are not considered for classification. The flow diagram shown in Figure 2 describes the steps involved in our proposed system. The abstract view of the processes in our work is described in the flow diagram.

The Iris dataset consists of four numeric attributes and one predictive attribute having three classes. It has 150 instances (50 each in the three classes). It is one of the best-used datasets in pattern recognition.

The US census dataset contains the census data conducted by the US Census Bureau. This has 41 attributes. However, the last attribute should not be used in the classifiers. The training record consists of 199,523 instances and the test file contains 99,762 instances. The description about the feature selection algorithms is given below.

### 3.1 Chi-square feature selection

Chi-square feature selection explained comprehensively by Mesleh (2007) and Zhu et al. (2007) in statistics is generally used to test the independence of two events. In feature selection, Chi-square method takes one event as the occurrence of an attribute and another event as the occurrence of the class.

The chi-square values can be calculated using the formula given below:

$$
\chi^2 = \frac{(O_i - E_i)^2}{E_i}
$$

(1)

In the above formula, $O_i$ is the occurrence of attribute and $E_i$ is the occurrence of the class. A rank is calculated based on these two factors. When Chi-squared algorithm is run on a dataset, it outputs the importance of the attribute for that dataset. On the basis of the attribute importance, the number of attributes essentially needed for predicting the class variable is chosen. Chi-square tests are of two types. The Chi-square goodness of fit test determines whether the sample data matches the class. Chi-square test of independence tests whether variables are related. Chi-square is present in the FSelector library of $R$ as a function.
3.2 Boruta feature selection algorithm

Boruta is an all-relevant feature selection wrapper algorithm. The important features are identified by comparing the importance of original attributes with the importance that is achieved at random, done by estimation of using the permuted copies. This is taken from the literature of Kursa and Rudnicki (2010). When the Boruta is run on a dataset, an important method called as attStats is shown in an attribute-centred way. Boruta uses shuffling principle to iteratively compare the importance of attributes. It is available under Boruta library of $R$.

The procedure adopted the implementation of Boruta algorithm is as follows:

- An extensive system is built where each descriptive variables are replicated. These values are randomly permuted across objects.
- Several random forest runs are performed on these variables maintaining randomness among the variables.
- For each run the importance values are calculated for all the attributes.

By default, 100 runs are performed on all the attributes. An extensive statistical test is performed on all the attributes. The hypothesis for this procedure is that the importance of the variable is equal to the maximal importance of the random attributes (MIRA).
For each attribute, the count on how many times the importance of the attributes are higher than MIRA is taken. The expected number of hits $N$ is $0.5 N$ with standard deviation of $\sqrt{0.25 N}$ as in the case of binomial distribution. A variable is said to be important or accepted if the number of hits is significantly higher than the expected value and is said to be rejected if the number of hits is significantly lesser than the expected value as projected by Kursa and Rudnicki (2010).

Five different classifiers will be used for the analysis: Naive Bayes, decision tree (J48), SVM, random forest and gradient boosting method. For each dataset, feature selection methods are first applied and with the resultant features, each classification techniques are applied and the performance metrics such as the accuracy, precision and recall values are calculated.

### 4 Results and discussions

In this study, two different feature selection methods and five different classification algorithms were chosen. These are applied to two different datasets, the US census dataset containing 41 attributes and the Iris dataset containing five attributes. In these datasets, one attribute is the predictor attribute and hence one attribute is left out while using these datasets for analysis. Thus 40 attributes from the US census dataset and four from Iris dataset were chosen for applying the feature selection algorithms.

First, the Chi-square feature selection algorithm was applied on the US census dataset with 40 attributes. The implementation yielded 28 attributes as important attributes. Next, the Boruta algorithm is run on the same dataset with 40 attributes which rejected 29 attributes leaving out 21 attributes for consideration. Then the five classification algorithms, Naive Bayes, decision tree (J48), SVM, random forest and gradient boosting method were run individually on both the datasets with reduced features. The accuracy, precision and recall were calculated and the results are obtained as shown in Table 1.

Out of 199,523 instances for training and 99,762 instances for testing, we had used 75,000 for training and 25,000 instances for testing. Table 1 shows that random forest (93.8) and gradient boosting (94.6) methods provide more accuracy when combined with Boruta algorithm. Figure 3 depicts the comparison of the accuracy, precision and recall values of all the five algorithms for the two feature selection algorithms we have taken under consideration.

<table>
<thead>
<tr>
<th>Classification algorithms</th>
<th>Chi-square</th>
<th>Boruta</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>Precision</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>86.2</td>
<td>85.3</td>
</tr>
<tr>
<td>Decision tree (J48)</td>
<td>84.3</td>
<td>79.8</td>
</tr>
<tr>
<td>Support vector machines (SVM)</td>
<td>86.4</td>
<td>81.8</td>
</tr>
<tr>
<td>Random forest</td>
<td>88.1</td>
<td>88.7</td>
</tr>
<tr>
<td>Gradient boosting method</td>
<td>88.3</td>
<td>89.6</td>
</tr>
</tbody>
</table>

Table 1 Performance metrics for US census data
It is evident from Figure 3 that the accuracy, precision and recall for random forest and gradient boosting is very high for Boruta feature selection algorithm while it is comparatively lower when combined with Chi-square feature selection. Also, it is interesting to note that Boruta yielded only 21 attributes for this census dataset and it performed much better than Chi-square that gave 28 attributes for the same census dataset, for all the classification algorithms.

On considering the other case, for the Iris dataset, both the feature selection methods suggested that all the four attributes are important for this dataset.

We first applied Chi-square method followed by the Boruta technique. One drawback in using Boruta is that the time taken to complete the algorithm is much higher than Chi-square. The time is very high for using a dataset that has only four attributes and with only 150 instances. Table 2 shows the performance metrics for Iris dataset applying the Chi-Square and Boruta and the five classification algorithms.

Out of 150 instances, we used 105 instances for training and 45 instances for testing. Table 2 shows the values of accuracy, precision and recall for all the five classification algorithms. Random forest and gradient boosting once again showed better results than the other classification algorithms.

### Table 2  Performance metrics for Iris dataset

<table>
<thead>
<tr>
<th>Classification algorithm</th>
<th>Chi-square</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>Precision</td>
<td>Recall</td>
<td>Accuracy</td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>83.6</td>
<td>81.5</td>
<td>84.1</td>
<td>83.6</td>
<td>81.5</td>
<td>84.1</td>
</tr>
<tr>
<td>Decision tree (J48)</td>
<td>86.4</td>
<td>85.4</td>
<td>86.6</td>
<td>86.4</td>
<td>85.4</td>
<td>86.6</td>
</tr>
<tr>
<td>Support vector machines (SVM)</td>
<td>87.0</td>
<td>86.1</td>
<td>86.8</td>
<td>87.0</td>
<td>86.1</td>
<td>86.8</td>
</tr>
<tr>
<td>Random forest</td>
<td>89.4</td>
<td>86.4</td>
<td>84.9</td>
<td>89.4</td>
<td>86.4</td>
<td>84.9</td>
</tr>
<tr>
<td>Gradient boosting method</td>
<td>88.6</td>
<td>80.9</td>
<td>81.6</td>
<td>88.6</td>
<td>80.9</td>
<td>81.6</td>
</tr>
</tbody>
</table>
But an interesting point to be noted from Table 2 is that the results of classification algorithms for both feature selection methods are the same. The reason is that both the feature selection methods listed all the attributes to be important. The only advantage is that Chi-Square saves a lot of time. So there are possibilities where feature selection techniques need not be applied on some datasets. Figure 4 shows the comparison of accuracy, precision and recall for Iris dataset.

Figure 4  Comparison of accuracy, precision and recall for Iris dataset (see online version for colours)

From this analysis, we can conclude that feature selection algorithms do play a major role in classification algorithms. For a dataset using a wide range of attributes, the Boruta algorithm selects a minimal number of attributes without compromising on the performance metrics. Similarly for a limited number of attributes in a dataset, Chi-square can be used as its execution time is very less. On the basis of the scope of the dataset, the feature selection algorithms may be chosen wisely. Boruta also has more chance of taking multiple-feature associations simultaneously without much computational overhead which may lead to less complexity. The results show that for dataset with more values, Boruta shows around 5–6% increased performance in the metrics which can be seen from the above tables.

5 Conclusion

A detailed analysis of the two different feature selection algorithms is carried out. For this analysis, two different datasets and five classification algorithms were used. With the results, it is clear that for a dataset that contains more number of attributes, the Boruta wrapper feature selection algorithm performs better whereas for a dataset that comprises of fewer attributes the Chi-square feature selection yields a better performance. Also, the running time of Boruta algorithm is higher compared with Chi-square. Even for
a dataset containing only four attributes, Boruta runs 100 times to check whether an attribute can be considered. This gives an edge to Chi-Square algorithm when the dataset has very limited number of attributes. As a general perspective, it is a very interesting result since it shows that Boruta feature selection method applied on a dataset with a large number of attributes and using the random forest and gradient boosting classification methods provided greater accuracy, precision and recall values. The results of this paper shall provide the foundation for our future research work on predicting whether a crime is a violent crime or a non-violent crime by applying these algorithms on the crime data.

References


