Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF
AGRICULTURAL STATISTICS 75(2) 2021 111–116

Selection of Feature Selection Algorithm for Categorization of Research Abstracts in Agricultural Domain

**Sreekumar Biswas[1], Rajni Jain[2], Sudeep Marwaha[1], Alka Arora[1],
A.R. Rao[3] and Monendra Grover[1]**

[1]ICAR-Indian Agricultural Statistics Research Institute, New Delhi
[2]ICAR-National Institute of Agricultural Economics & Policy Research, New Delhi
[3]Indian Council of Agricultural Research, New Delhi

Received 21 August 2020; Revised 25 July 2021; Accepted 23 September 2021

SUMMARY

Feature selection is one of the most important steps while dealing with classification problems. When it narrows down towards text classification, feature selection becomes indispensable because of the high dimensionality of the feature vector. Feature selection techniques are categorized into a filter, classifier subset evaluation and wrapper. This study presents the empirical results of the comparison of the feature selection for text categorization using research texts in the agricultural domain. The study recommends that the Classifier Subset Evaluator method for feature selection using Naïve Bays as parameter algorithm and MLP as the classifier is the best framework for categorization of agricultural text documents with 90% accuracy.

Keywords: Text categorization, Feature selection, Machine learning, Classifier subset evaluation, Wrapper subset evaluation, Artificial intelligence in agriculture.

1. INTRODUCTION

In the real-world scenario, most of the classification problems require supervised learning where each instance is associated with a class label that determines the category in which a particular instance belongs (Witten *et al.* 1999). Unfortunately, the set of relevant features or attributes for classifying such problems are unknown beforehand (Guyon and Elisseeff 2003; Liu and Motoda 2012). Text Mining refers to the process of finding patterns in data. Text Categorization (TC) is one such task of Text Mining. TC can be defined as the task of assigning labels to text documents by a model, which was trained with some predefined text documents. While dealing with text data for the task of TC, the vector may have a large number of features consisting of many irrelevant and/or redundant features. So, features in a vector can be distinguished in the following ways:

- Relevant features: those features that contribute the most to classification accuracy.
- Irrelevant features: those features that do not contribute to the classification process, rather they may decrease the classification accuracy.
- Redundant features: those features that do not add anything new to the task.

In the case of TC, as stated earlier, the feature vector becomes very large. Hence to perform classification, it is highly recommended to perform feature selection (FS) in the data pre-processing step to increase the classification accuracy to a significant amount. Blum *et al.* (1997) stated various definitions of FS. FS can be defined as the process of selecting a subset of relevant features from the set of all features for model building (Zhou 2007). In the conducted experiment, agricultural research articles are the domain of interest and the

Corresponding author: Rajni Jain
E-mail address: rajnijain67@gmail.com

goal is to classify the articles into different classes to which they belong. Biswas and Jain (2018) have also experimented with agricultural articles consisting of only titles. We have used the WEKA engine and java for classification. We present a comparison of text categorization approaches with different feature selection approaches and identify the best one for this domain.

2. MATERIAL AND METHODS

2.1 Methods of Feature Selection

FS is a technique of generating subsets from the original feature set by a search technique and evaluating them with some evaluation measure (Bijanzadeh *et al.* 2010; Chizi and Oded 2009; Dash and Liu 1997). The selected subset is the set of features that minimizes the error rate. Broadly, there are two categories of FS, *viz*., Filter method and Wrapper method. The filter method uses an attribute evaluator and a ranker search to rank all the features in the dataset. The number of selected features in this approach can be specified beforehand. This approach removes those features that attain lower ranks. Yang *et al.* (1997) showed different ways of the filter method which were applied in this experiment. The wrapper method uses an inductive algorithm for the calculation of the value of a given subset. This method wraps a classifier in a cross-validation loop and searches through the attribute space defined by the feature vector. Then it uses the classifier to find a good attribute set depending on the evaluation criteria. Searching can be forwards, backwards, or bidirectional, starting from any subset. This method has a high cost in terms of time, but the selected features result in improved classification accuracy as compared to the filter method.

In this experiment, we also attempted the filter-based approaches (Yang *et al.* 1997; Tsanas *et al.* 2010)) but the results were not so optimistic because our feature vector is more like a sparse matrix, where the information gain (Cao *et al.,* 2012) and correlation coefficient are considered negligible. For this reason, results of filter-based FS are ignored (Table1). We have used the wrapper method and the classifier subset evaluator for feature selection. In both cases, Naïve Bayes (NB) and Sequential Minimal Optimization (SMO) as the evaluator and best-first search strategy. Table 2 describes all the techniques we have used in this research.

Table 1. Performance of Filter-based approach

Feature Selection method	Method	Size of reduced feature set (number)	Accuracy	Time Taken in Seconds
Filter Method	Correlation Attribute Eval	4732	58.76	271
	Info Gain Attribute Eval	5563	53.23	105
	Cfs Subset Eval	No Change	NA	335

Table 2. Description of feature section techniques

Approach	Description
CSNB	classifiers using classifier subset evaluation using Naïve Bayes as parameter algorithm
CSSM	classifiers using classifier subset evaluation usingSMO as parameter algorithm
WSNB	classifiers using wrapper subset evaluation using Naïve Bayes as parameter algorithm
WSSM	classifiers using wrapper subset evaluation using SMO as parameter algorithm

2.2 Data and Experiment

The data has been collected from Prof. M.S. Swaminathan Library, Indian Agricultural Research Institute. The data is summarized in Table 3.

Table 3. Data description

Name of the file	Ag.arff
Number of records	182
Source	Prof. M. S. Swaminathan Library, Indian Agricultural Research Institute
Number of classes	4
No of attributes	5781
Type of attributes	Numeric
Missing values	Nil

In this experiment, we collected the data from IARI Library and stored it as a text file with a .txt extension. This data included the titles and abstracts of the research papers and has been stored in the local system. Preprocessing of the data was the next step which included stemming, stop-word removal and tokenization. Then the data was converted into vector format and further processing was done. The format of the data for the experiment is Attribute Relation File Format (ARFF) as explained by Frank *et al.*, 2016. The research papers were selected from four domains *viz*., Agronomy, Economics, Genetics and Soil Science.

The next step was feature selection. WEKA 3.8.0 was used to experiment along with java. WEKA stands for Waikato Environment for Knowledge Analysis, developed from the University of Waikato, New Zealand. It is an open-source machine learning tool licensed under the GNU General Public License. This software contains a collection of visualization tools and algorithms for the task of data analysis and predictive analysis (Frank *et al.*, 2016). For feature selection, we have used wrapper based FS technique and classifier subset evaluator. In both the cases, Naïve Bayes and Sequential Minimal Optimization algorithms were used as an evaluator and for searching the attributes from the data, best-first strategy was used (Xu *et al.*, 1988). Best First Algorithm (BFA) is a heuristic search strategy taken from the field of Artificial Intelligence and is applied to feature selection. Xu *et al.* (1988) proved that this algorithm could guarantee the globally best subset without exhaustive enumeration for any criterion that satisfies monotonicity. They showed that the number of subsets evaluated by BFA was much less than that needed by the branch and bound algorithm, an optimal feature selection algorithm proposed by Marendra and Funkunaga (1977). BFA is an informed search strategy that uses problem-specific knowledge beyond the definition of the problem itself, so it can find solutions more efficiently than an uninformed strategy. Best-first search is an instance of the general TREE-SEARCH or GRAPH-SEARCH algorithm in which a node is selected for expansion based on an evaluation function, f(n). A key component of these algorithms is a heuristic function denoted h(n) where h(n) refers to the estimated cost of the cheapest path from node n to a goal node, if n is a goal node, then h(n) = 0. We refrain from going into details as they can be read from any textbook on Artificial intelligence (Nilsson 2014).

At first, we made a comparison test for the selection of the feature selection method. For this, we have conducted the feature selection procedure using the methods mentioned above. Then we ran classification algorithms on the same training and test data over a 10-fold CV for the evaluation of the FS methods. A pairwise t-test was used to obtain the significance of the experiment and further, to select one of the two FS techniques to carry the experiment so that a model could be prepared for the categorization of agricultural research articles.

The experimental phase started with dividing the data manually in 10-folds, then all the classifiers were evaluated by 10-fold cross-validation, iterated over 10 times; thus, making 10x10-fold cross-validation (CV). The 10x10-fold CV gave a better evaluation for comparing the performances of each classifier.

The classification is performed using J48, Naïve Bayes, Random Forest, K-NearestNeighbor (K=3), Support Vector Machine (SVM) and Multi-Layer Perceptron (MLP) based algorithms. Both the methods of feature selection were performed on the data and all these algorithms were applied to perform classification. The objective was to check the accuracy, whether it has improved or not.

ANOVA test on the results obtained from 10X10-fold CV confirmed that there was no significant difference between the FS techniques tested. Therefore, the method by which maximum accuracy was obtained was selected as the technique of FS for agricultural documents.

3. RESULTS AND DISCUSSION

The first objective was selection of a parameter algorithm. We experimented with 5 parameter algorithms for both the feature selection methods namely classifier subset evaluation and Wrapper subset evaluation. We used the same classifier (MLP) to compare the performance of the parameter algorithm. A small dataset of 182 research papers was used for this

Table 4. Performances of Different Algorithms for Feature Selection on a small size text dataset

Feature Selection method	Parameter Algorithm	Classifier	Size of reduced feature set (number)	Accuracy	Time Taken in Seconds
Classifier Subset Evaluation	KNN	MLP	44	68.12	332
	J48	MLP	20	62.63	147
	RF	MLP	1	0.00	6437
	NB	MLP	43	68.72	508
	SVM	MLP	37	69.83	537
Wrapper Subset Evaluation	KNN	MLP	15	63.18	198
	J48	MLP	10	69.27	203
	RF	MLP	1	0.00	5230
	NB	MLP	33	75.95	1410
	SVM	MLP	30	78.63	1215

purpose. Table 4 shows the performance of different parameter algorithms used in this study concerning the size of the reduced feature set, accuracy and time taken for feature selection.

As seen in Table 4, though the time taken by NB, and SVM are a bit high compared to other algorithms in both feature selection methods, they showed more accuracy in classification. Hence, these two parameter algorithms (NB and SVM) were used in the subsequent steps with a text dataset of abstracts. Table 4 values are based on experiments with the real-time dataset. They vary with the dataset. Results are indicative of comparative performance. The performance will improve if the dataset is large enough and consisting of thousands of documents

The second objective is a selection of feature selection methods (Fig. 1). While comparing the approaches of wrapper subset evaluation and classifier subset evaluation for feature selection using paired t-test, we obtained a p-value of 0.002. Thus, we can reject the null hypothesis at a 95% confidence interval and can say that there is a significant difference in the result obtained by the two methods applied for FS. We selected the method of FS that provided us with the best accuracy which is the classifier subset evaluation. Fig. 1 presents the comparison of the CS and WS approaches for feature selection. In the majority of the cases, the CS approach is outperforming the WS approach. Hence, we adopted CS for further categorization.

The results of the 10x10 CV are shown in the subsequent tables. Table 5 and Table 6 show the results for the classifier subset feature selection with NB and SMO as parameter algorithms, respectively. These tables reflect the individual accuracy obtained from each of the 10 runs of the cross-validation results. For example, in iteration 10, we observed accuracy of 92.3% with NBRF and 87.9 with NBJ48 (Table 5). Please note that each iteration accuracy itself is based on 10 times the execution of the algorithm. In Fig. 1, we have compared the average accuracy after considering

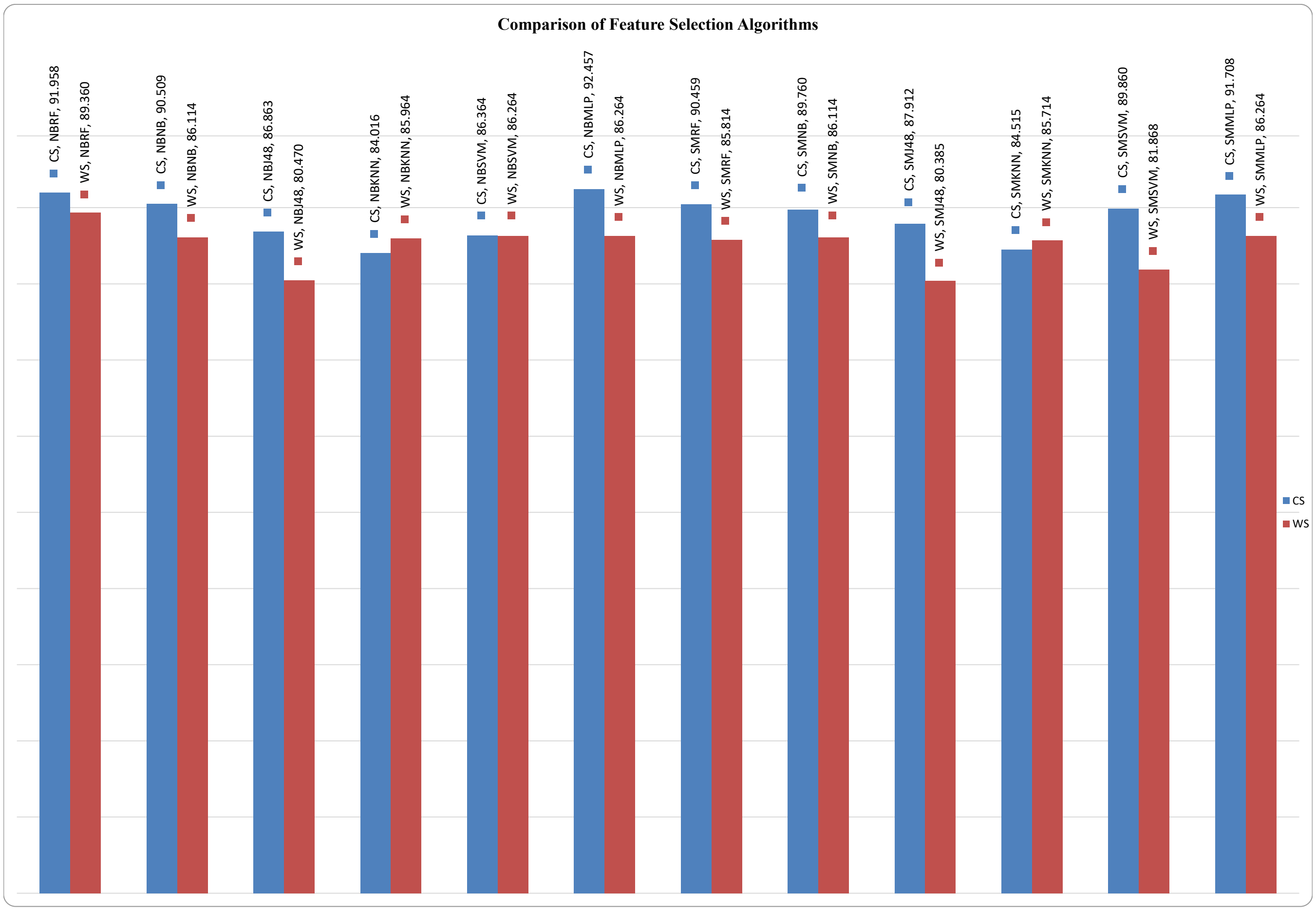

Fig. 1. Comparison of Feature Selection Algorithms

Table 5. Classifier Subset feature selection with NB algorithm (CSNB) for each of the 10x10-fold CV on the same training and test dataset

Iteration	Accuracy (%) of Algorithms					
	RF	NB	J48	KNN	SVM	MLP
1	88.462	84.615	83.517	81.3187	86.813	90.110
2	93.956	90.110	89.560	83.517	86.264	94.506
3	90.695	91.209	85.714	83.517	86.264	91.209
4	92.857	90.110	87.363	82.418	86.813	92.308
5	94.506	91.209	87.912	86.813	90.110	93.956
6	93.596	90.659	87.912	87.363	88.462	93.956
7	92.308	100	87.363	86.813	87.363	92.857
8	92.308	85.165	87.912	80.220	82.967	91.758
9	89.560	87.912	84.615	83.517	82.418	90.659
10	92.308	93.407	87.912	85.165	86.264	94.506
Average	91.958	90.509	86.863	84.016	86.364	92.457

Table 6. Classifier Subset feature selection with SMO algorithm (CSSM) for each of the 10x10-fold CV on same training and test dataset

Iteration	Accuracy (%) of Algorithms					
	RF	NB	J48	KNN	SVM	MLP
1	90.110	89.560	87.912	82.418	91.209	92.857
2	90.110	93.4066	87.912	84.615	90.110	92.308
3	88.462	87.912	86.264	82.967	86.264	89.560
4	92.308	93.407	89.56	84.066	90.659	92.857
5	92.857	90.659	91.209	87.912	92.857	92.308
6	93.407	90.110	90.659	85.714	92.308	95.604
7	91.209	91.758	89.56	86.813	92.308	91.758
8	89.56	86.813	86.813	83.517	88.462	89.011
9	87.363	87.363	85.165	82.418	87.363	91.209
10	91.209	88.462	89.011	86.264	90.659	91.758
Average	90.459	89.760	87.912	84.515	89.860	91.708

all 100 iterations (10X10CV). Similarly, in Table 6, we have shown the results from the SMO algorithm.

The 10x10-fold CV provides better reliability to state the accuracy of the classifiers and we can see that the error rate is pretty minimal to accept all the learners. However, we have to select the classifier that gives us the highest accuracy. Fig. 2 enlightens the fact to select the classifier which is considered to be the best categorizer of agricultural texts. Y-axis shows the accuracy of the various algorithms presented on the X-axis.

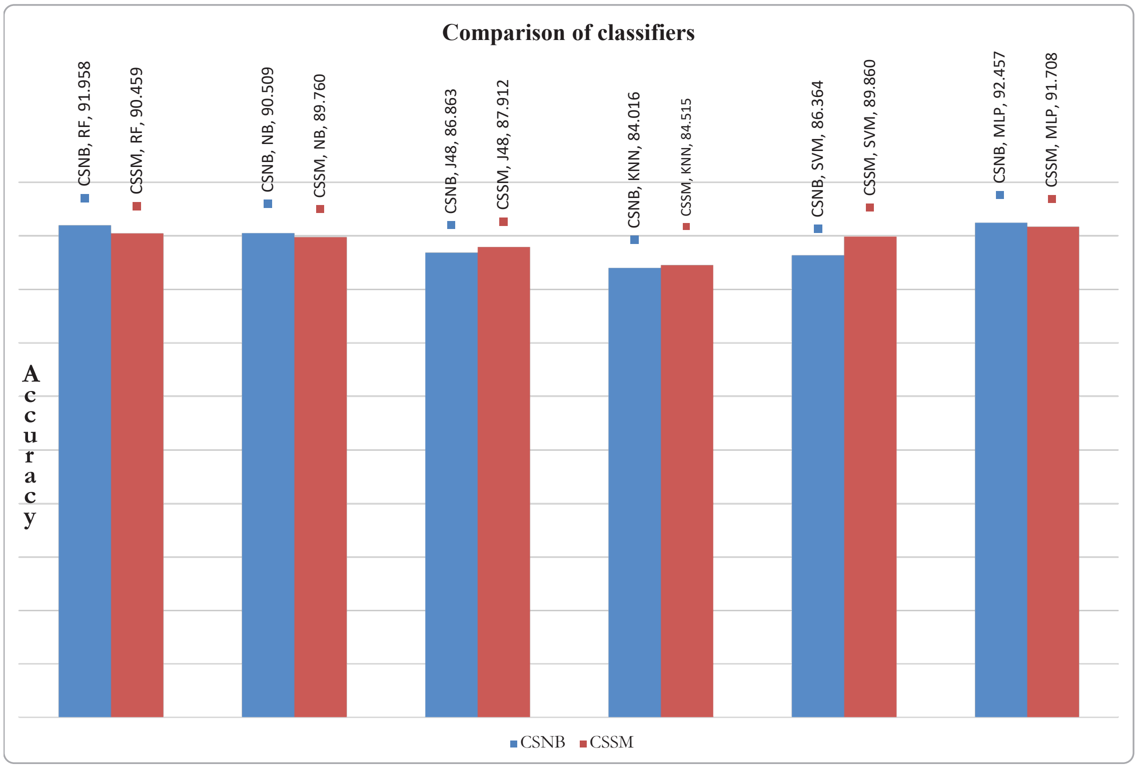

Fig. 2. Comparison of Classifiers

We see that the use of MLP combined with the classifier subset evaluation scheme of FS and NB as the parameter algorithm provided more than 90% accuracy to classify the agricultural texts.

4. CONCLUSION

The experimental results show that both techniques are very much efficient for feature selection for agricultural text. But the classifier subset feature selection provides us better performance over wrapper based feature selection. The accuracy of the classifiers is more than 90% and hence are good enough to be used as the classification model for categorizing agricultural texts. However, the Classifier Subset Evaluator method for feature selection using Naïve Bays as parameter algorithm and MLP as the classifier (CSNBMLP) is the best classifier in this context. In future, this technique may be extended to some ensemble learning techniques with FS techniques to further improve the efficiency of the classification model while dealing with much larger datasets.

ACKNOWLEDGEMENTS

Authors are thankful to the anonymous referee for the useful comments which helped in improving the paper contents.

REFERENCES

Bijanzadeh, E., Yahya E., and Esmaeil, E. (2010). Determining the most important features contributing to wheat grain yield using supervised feature selection model. *Australian Journal of Crop Science,* **4(6)**, 402.

Biswas, S. and Jain, R. (2018). Text Document Categorization using Machine Learning Algorithm in Agricultural Domain. *J. Ind. Soc. Agril. Statist.,* **72(1)**, 61-69.

Blum, A.L. and Pat, L. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence* 97(1): 245-271.

Cao. D., Ma, N., Liu, Y. and Guo, J. (2012). A feature selection algorithm for continuous attributes based on the information entropy. *Journal of Computational Information Systems,* **8(4)**, 1467-1475

Chizi, B., and Oded, M. (2009). Dimension reduction and feature selection. *Data Mining and Knowledge Discovery Handbook*. Springer, Boston, MA, pp 83-100.

Dash, M. and Liu, H. (1997). Feature selection for classification. *Intelligent Data Analysis,* **1(1-4)**, 131-156.

Frank, E., Hall, M.A. and Witten, I.H. (2016). The WEKA workbench. *Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research,* **3**, 1157-1182.

Liu, H. and Motoda, H. (2012). *Feature Selection for Knowledge Discovery and Data Mining* (Vol. 454). Springer Science and Business Media.

Narendra, P.M. and Funkunaga, A.K. (1977). A branch and bound algorithm for feature subset selection. *IEEE Transactions,* **C-26(9)**, 917-922.

Nilsson, N. J. (2014). *Principles of Artificial Intelligence*. Morgan Kaufmann.

Tsanas, A., Little, M.A. and McSharry, P.E. (2010). A simple filter benchmark for feature selection. *Journal of Machine Learning Research,* **1,** 1-24.

Witten, I., Frank, E., Trigg, L., Hall, M., Holmes, G. and Cunningham, S. (1999). *Weka: Practical machine learning tools and techniques with java implementations*.

Xu. L., Yan, P. and Chang, T. (1988). Best first strategy for feature selection. *In 9th International Conference on Pattern Recognition,* pp. 706-707. IEEE Computer Society.

Zhou, H., Wu, J., Wang, Y. and Tian, M. (2007). Wrapper approach for feature subset selection using GA. *In 2007 International Symposium on Intelligent Signal Processing and Communication Systems,* pp. 188-191. IEEE.