Comparative Study on Different Classification Techniques for Spam Dataset

Sanaa Hassan Abou Elhamayed*
ERI, Cairo, Egypt.

* Corresponding author. Email: sanaa-hamayed@hotmail.com
Manuscript submitted May 10, 2018; accepted July 18, 2018.
doi: 10.17706/ijcce.2018.7.4.189-194

Abstract: Nowadays, people and companies use emails for information exchange, email messages, and etc., because they are the fastest and the cheapest way. The main problem that faces email messages is the undesirable emails which known as spams. Spams may cause overflow the internet with considerable copies of the same message or carry malicious content that harms user system and reduce the performance. The purpose of this work is to make a comparative study of several classification techniques on the basis of their performance parameters using spam dataset. The performance of the different classifiers is measured with different ratio of the testing and training dataset. Also, the performance of the classifiers is calculated with and without low variance filter. By applying the low variance filter the accuracy of the KNN classifier is enhanced with about 9% while the accuracy of the other classifier is decreased.

Key words: Email classification, filter, spam, variance.

1. Introduction

Electronic mail message became progressively vital and widespread technique communication due to its time speed [1]. Nowadays, spam has become serious issue for computer security, because it becomes a main source for disseminating threats, including viruses, worms and phishing attacks [2]. Classification techniques are used to overcome the problem of spam. A filter is software designed to restrict the content a reader is authorized to access from the internet via the web, email or other means.

Social networking websites are used by millions of people around the world. People express their views, opinions and share current topics. Millions of data generated every day. Now a day's spammers used this platform to advertise spam content on the social networking websites [3].

In this work different techniques of classification are used. These techniques are lazy, rules, function, tree, and Bayesian. Where KNN, SVM, PART, J48, and Naivbays classification methods are used to detect a spam of email. These algorithms are examined with 10-fold cross validation test mode by using WEKA software with spam email dataset. These techniques are compared in different values of several percentage ratios of training and testing dataset. Then a feature selection by calculating variance of each attribute is used and compared the different techniques before and after using it.

This paper is organized as follows: Section 2 presents some of the previous work. Section 3 describes the dataset which is used in this work. Section 4 discusses some classification techniques. Experimental work and results have been discussed in Section 5. Finally, section 6 is the conclusion of the whole work.

2. Literature Survey

Ref. [1] shows that, they classified email as spam as legitimate using Naïve bayse, K nearest neighbor, and Support Vector Machine. They used WEKA as interface. They found that Naïve Bayes algorithm as simplest most important classifier.

Ref. [2] shows that, he focused on systematically analyzing the strength and weakness of current technologies for spam detection. He discussed the main approaches as well as their weakness and strength for detecting spam.

Ref. [3] shows that, they classified tweets into spam and non spam tweets. They used 120 character tweets for their analysis system. They extracted the tweets from the chosen various active and verified twitter accounts. The dataset was created from words of each tweet and applied to Support Vector Machine model to detect spam or non spam tweeter.

Ref. [4] shows that, they detected spam email by using K nearest neighbor classification method by combining spearman's correlation coefficient as distance measure. Their experimental results presented a significant improvement in accuracy than the traditional methods.

Ref. [5] shows that, he suggested a new effective method that reduced the spam messages by integrating prevention and detection techniques in one scheme. His proposed model allowed the transition of email from one state to another state based on the number of received spam and non spam messages. The name of this model is SREHA. SREHA allowed and enabled each email server to disseminate gained information about the spams and the spammers to share these information with other servers that enabled them to act against spammers.

3. Description of the Dataset

In this work the classification techniques are applied to the spam dataset downloaded from website: https://archive.ics.uci.edu/ml/datasets/Spambase. Creators of this dataset are Mark Hopkins, Erik Reeber, George Forman, Jaap Suermondt Hewlett-Packard Labs, 1501 Page Mill Rd., Palo Alto, CA 94304. The name of this dataset is SPAM E-mail Database and it is used to determine whether a given email is spam or not. It contains 4601 instances with 58 attributes each. The dataset has missing values and the attributes are integer and real numbers. The last attribute is the class label which denotes the e-mail is spam or non-spam. The attributes (55-57) measure the length of sequences of consecutive capital letters. The attributes (49-54) are type char_freq_CHAR= percentage of characters in the e-mail that match CHAR, i.e. 100 * (number of CHAR occurences) / total characters in e-mail. The first 48 attributes are type of word_freq_WORD = percentage of words in the e-mail that match WORD, i.e. 100 * (number of times the WORD appears in the e-mail) / total number of words in e-mail.

4. Classification Techniques

4.1. Bayes Technique

The Bayesian classifier calculates explicit probabilities for hypothesis and it is robust to noise in input data. It uses the theorem of Bayes theorem that says:

$$P(Z_j \mid d) = P(d|Z_j)\, P(Z_j) / P(d) \quad (1)$$

Consider every attribute and sophistication label as a chance variable and given an instance X with attributes n (X_1, X_2…X_n). The aim of this theorem is to predict category Z [6]. In this work Naive Bayes classification algorithm is used.

4.2. Lazy Technique

The reason of the lazy name is that they store the training instances and do no real work until classification time. KNN classification is based on learning by an evaluation, that is, by comparing a given test instance with training instances that are similar to it [7]. To identify the closest pattern instances the Euclidean distance is calculated as:

$$D\ (p_1, p_2) = ((X_2\text{-}X_1)2 + (Y_2\text{-}Y_1)2)1/2 \quad (2)$$

where, p_1 and p_2 represents the instances in space having coordinates (x_1, y_1) and (x_2, y_2) respectively. KNN in which nearest neighbor is calculated on the basis of value of k that specifies how many nearest neighbors are to be considered to define class of a sample data point.

4.3. Function Technique

SVM is a classifier which is based on vector space where an instance is represented in vector space and each feature (word) represents one dimension. Identical feature denotes same dimension. Two of the parameter particularly term frequency (TF) and TF-inverse document frequency (TF-IDF) add price to those vectors. Wherever TF the quantity of times a word occur in a very document TF-IDF uses the on top of TF multiplied by the IDF (inverse document frequency). DF (document frequency) is that the variety of times that word happens altogether the documents [1], [8].

4.4. Decision Trees Technique

Decision trees are trees that classify instances by sorting them based on attribute values. Each node in a decision tree represents an attribute in an instance to be classified, and each branch represents a value that the node can assume. Instances are classified starting at the root node and sorted based on their attribute values [6]. In this work J48 is applied.

4.5. Rule Technique

Relationship among all attributes can be found by using rules. Rule-based classifier makes use of set of IF-THEN rules for classification. The IF part of the rule is called rule antecedent or precondition. The THEN part of the rule is called rule consequent. In the antecedent part the condition consists of one or more attribute tests and these tests are logically ANDed. The consequent part consist class prediction. If the condition holds the true for a given tuple, then the antecedent is satisfied. Rule is extracted from a decision tree, one rule is created for each path from the root to the leaf node. To from the rule antecedent each splitting criterion is logically ANDed. The leaf node holds the class prediction, forming the rule consequent [6], [7]. In this work PART is applied.

5. Experimental Work and Results

Every word in the dataset has certain probability of occurring in spam or ham e-mail. If the total of words probabilities exceeds a certain limit, the classification method will classify the e-mail to spam or ham email. Five traditional techniques lazy, rules, function, tree, and Bayesian are applied and compared. Several experiments are applied on the spam dataset by changing the training dataset and the testing dataset ratio. The results are obtained for each classifier before and after feature selection.

The performance is evaluated using correctly classified instances, mean absolute error (MAE), time taken to build model, precision, recall, and *f*-measure. Accuracy is total number of instances correctly classified from total number of instances while Mean absolute error represents how close a predicted model to actual model. Also, precision, recall, and *f*-measure are recorded to compute the score of the classifiers.

In Fig. 1, quantitative performance measures are given in terms of accuracy for different percentage ratio for training and testing dataset for the several classifiers.

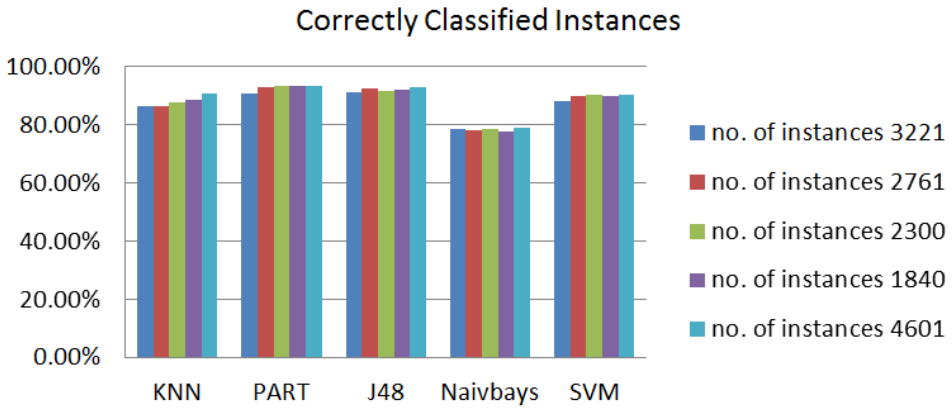

Fig. 1. Comparison of algorithms with different ratio of training and testing dataset.

Table 1. Mean Absolute Error with Different Ratio of Training and Testing Dataset

Train-test %	no. of instances	KNN	PART	J48	Naivbays	SVM
30-70%	3221	0.1377	0.1067	0.1106	0.2134	0.1183
40-60%	2761	0.1347	0.092	0.1008	0.2192	0.1021
50-50%	2300	0.1225	0.079	0.1033	0.2146	0.0978
60-40%	1840	0.1144	0.0791	0.0988	0.2208	0.0989
100-0%	4601	0.0924	0.0769	0.0892	0.2066	0.0958

Table 1 shows that PART classifier gives less mean absolute error for all different ratio of training and testing dataset than the other algorithms. Furthermore, the use of increased number of training dataset, the performance of all classifiers is enhanced.

Table 2. Time Taken to Build Model with Different Ratio of Training and Testing Dataset

Train-test %	no. of instances	KNN	PART	J48	Naivbays	SVM
30-70%	3221	0.01	9.2	3.33	0.47	1.95
40-60%	2761	0	9.27	3.14	0.3	2.01
50-50%	2300	0	9.35	3.08	0.45	2.23
60-40%	1840	0	9.22	2.59	0.42	2.08
100-0%	4601	0.01	15.67	7.19	2.41	1.64

From Table 2 it is clear that the KNN classifier is the best concerning the time taken to build the model.

Table 3. Comparison of Algorithms without Feature Selection

	KNN	PART	J48	Naivbays	SVM
Accuracy	90.78%	0.93.59%	0.92.98%	79.29%	90.42%
MAE	0.0924	0.0769	0.0892	0.2066	0.0958
Precision	0.921	0.947	0.94	0.956	0.896
Recall	0.927	0.947	0.944	0.69	0.952
F-Measure	0.924	0.947	0.942	0.801	0.923

The features are selected by calculated the variance of each attribute. Low variance filter where

considering that attribute with little changes carries little information. Thus all attributes with variance lower than a given threshold are removed. Table 3 shows the performance of the classifiers before applying low variance filter. The attributes are decreased to 12 attributes and Table 4 shows the results of the classifiers after applying the low variance filter.

Table 4. Comparison of Algorithms with Feature Selection

	KNN	PART	J48	Naivbays	SVM
Accuracy	99.76%	90.68%	92.72%	73.29%	80.439
MAE	0.003	0.1401	0.1159	0.244	0.1956
Precision	0.996	0.943	0.944	0.918	0.819
Recall	1	0.901	0.936	0.614	0.87
F-Measure	0.998	0.921	0.94	0.736	0.843

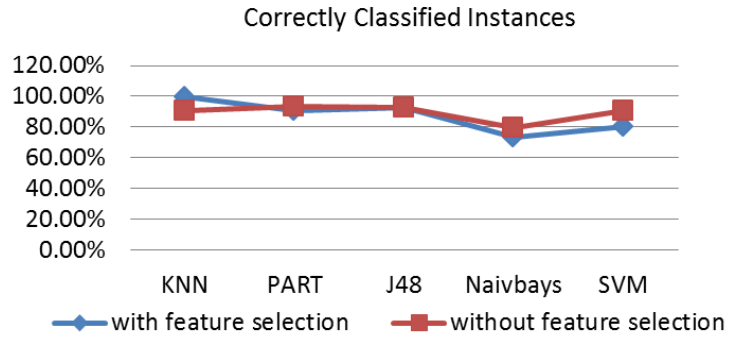

Fig. 2. Comparison of algorithms with and without feature selection.

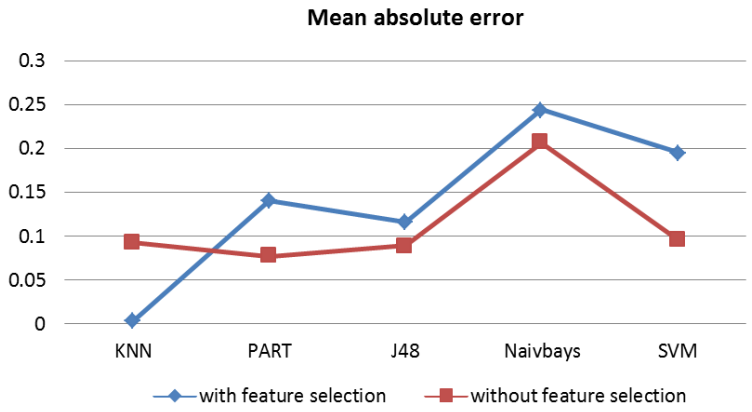

Fig. 3. Comparison of algorithms with and without feature selection.

From the careful observation of the results in Fig. 2 and Fig. 3, it can be seen that the use of low variance filter leads to increased classification accuracy of KNN classifier about to 9% as compared without using it. While for the other classifiers the accuracy is decreased.

6. Conclusion

A brief comparative study on the performance of different classifiers algorithms are tested using spam dataset with and without low variance filter. The classification algorithms' performance is calculated with different percentage ratio values of the training and testing dataset. The results with a varying number of training dataset indicate that the performance of the various classifiers increases when the number of training dataset increases. It is clear from this work that PART and J48 outperform the other classifiers for

the given dataset without using filter. Using low variance filter for feature selection enhances the performance of KNN classifier while not suitable for the others. The future work will focus on the improvement of classifier's performance by using different ways of feature selection.

References

[1] Elifenesh, Y., & Manisha, T. (2016). Email classification using classification method. *International Journal of Engineering Trends and Technology (IJETT), 32(3)*, 142-145.

[2] Muhammad, I., Malik, M. A., Mushtaq, A., & Faisal, K. (2016). Study on the effectiveness of spam detection technologies. *International Journal of Information Technology and Computer Science*, 11-21.

[3] Abha, T., & Smita, J. (2016, November). Spam filtering methods and machine learning algorithm - A survey. *International Journal of Computer Applications, 154(6)*, 8-12.

[4] Ajay, S., & Anil, S. (2016, February). A Novel method for detecting spam email using KNN classification with spearman correlation as distance measure. *International Journal of Computer Applications, 136(6)*, 28-35.

[5] Adwan, F. Y. (2016). Spam reduction by using e-mail history and authentication (SREHA). *International Journal Computer Network and Information Security, 7*, 17-22.

[6] Satyanarayana, N., Ramalingaswamy, C., & Ramadevi, Y. (2014). Survey of classification techniques in data mining. *IJISET - International Journal of Innovative Science, Engineering & Technology, 1(9)*, 268-278.

[7] Ritu, S., Shiv, K., & Rohit, M. (2015). Comparative analysis of classification techniques in data mining using different datasets. *International Journal of Computer Science and Mobile Computing, 4(12)*, 125-134.

[8] Kishansingh, R., & Bhavesh, A. O. (2017). A comparative study of classification techniques in data mining. *International Journal of Creative Research Thoughts, 5(3)*, 154-163.

Sanaa Hassan Abou Elhamayed is a PhD of engineering holder from Cairo, Egypt. She works as a part of Informatics Research Department in ERI. Her research interests are natural language processing, information system, and machine learning. She is a researcher in Electronic Research Institute.

List of her latest publication:

1. Enhancement of agriculture classification by using different classification systems. International Journal of Computer Applications (IJCA), 2016.

2. Classifying datasets using some different classification methods. International Journal of Engineering and Technical Research (IJETR), 2016.