



International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

Improving spam mail filtering using classification algorithms with discretization Filter

Supriya S. Shinde¹, Prof. Rahul Patil²

¹ME Computer Student, ²Professor

^{1,2}Computer Department, Pimpri Chinchwad College of Engineering,
Nigdi, Pune, Maharashtra, INDIA

Abstract: Email spam or junk e-mail is one of the major problems of the today's usage of Internet, which carries financial damage to organizations and annoying individual users. Among the approaches developed till date to stop spam, filtering is an important and popular technique. Common practices for mail filters include organizing incoming email and removal of spam mails and viruses. A rare use is to examine outgoing email at some companies to ensure that employees observe with appropriate rules. Users might also employ a mail filter to prioritize messages, and to sort them into folders based on subject matter or other criteria. For many years these elements have driven pattern recognition and machine learning communities to keep improving email filtering techniques. Mail filters can be connected by the user, either as separate packages, or as part of their email program as email client. In case of emails, users can make manual filters which automatically filter mail according to the criteria chosen by particular user. In this paper, we present a survey of the performance of six machine learning methods in spam filtering techniques. Experiments are carried out on different classification techniques and association techniques using Waikato Environment for Knowledge Analysis (WEKA). Different classifiers are applied on one benchmark dataset in to evaluate which classifier gives better result. The dataset is in Attribute Relation File Format (ARFF). 10 fold cross validation is used to provide well accuracy. Results of classification algorithms are compared on spambase UCI dataset and it is found that no single algorithm performs best for spam mail filtering. For the different dataset it is observed that performance varies with different data sets. Our results prove that the performance of classification improve with filters.

Key Terms: E-mail classification, Spam, Spam filtering, Machine learning, algorithm, WEKA classification Algorithm.

I. Introduction

In recent years, e-mails have become a common and important medium of communication for most Internet users. Data mining [3] is defined as the process of discovering patterns in data. In today's world it can be used in various areas such as fraud detection, scientific discovery, marketing industry, artificial intelligence, Surveillance etc. This paper uses UCI spambase dataset [7]. Preprocessing is done and different classification methods are compared based on their performance. Rules are generated using association rule mining. Classification is a data mining (machine learning) technique which has a set of predefined classes and determine which class a new object belongs to it. There are large number of classifiers available which are used to classify the data such as bays, function, Rule, lazy, meta, decision tree etc. Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. The data must be preprocessed to select a subset of attributes to use in learning. For preprocessing we have applied Discretization filter.

A. Weka:

WEKA is a popular suite of machine learning software .It is developed at the University of Waikato in New Zealand by researchers. WEKA is freely available under GNU General Public License (GPL).This software is written in JAVA contains interactive GUI(Graphical User Interface).It supports many data mining tasks such as preprocessing, Classification, Clustering, Select attribute, Associate, visualization etc. In WEKA three graphical user interfaces are available namely "The Explorer" (exploratory data analysis) "The Experimenter" (experimental environment) "The Knowledge Flow" (new process model inspired interface) [2].WEKA Data format uses flat text files to describe the data. It can work with a wide variety of data files. Data can be imported from a file in various formats: ARFF, CSV, C4.5, binary Data can also be read from a URL or from an SQL database . [3]

B. Dataset:

The UCI Spambase data set has 4601 instances. Among them, 1813 instances are spam and 2788 instances are non-spam. For each instance, there are 58 attributes. The first 48 attributes are all continuous real number

ranging from 0 to 100, with the name word_freq_WORD. The last attribute is a nominal class attribute where 1 denotes that the email is considered spam, 0 denotes that the email is non-spam.

II. Literature Review

This section describes the preprocessing Discretize filtering technique, spam filtering methods and classification methods in detail. We have taken UCI Spambase dataset and examined the classification result using six methods Bayes, Function, Lazy, Meta, Rules, Tree classifiers.

Spam Techniques:

A. Bag-of-Words model based methods:

- **Support vector machine (SVM):** In 2006, the SVM classifier was proposed for spam email filtering. SVM worked based on training email samples and a predefined transformation $\theta : R^s \rightarrow F$, which forms a map from features to create a transformed feature space, storing the email samples of the two classes with a hyperplane in the transformed feature space.[10] The decision rules appear in the following formula:

$$f(\tilde{y}) = \text{sign} \sum_{i=1}^n (a_i x_i K(\tilde{u}, \tilde{o}) + s)$$

where, $K(\tilde{u}, \tilde{o}) = \theta(\tilde{u}) \cdot \theta(\tilde{o})$. Is the kernel function and $a_i, i = 1..n$ and S maximize margin space separate between hyperplane.

- **Term frequency-inverse document frequency (TF-IDF):** TF-IDF is used for word weights, as feature for the clustering. The document frequency of the word w is implemented by $DF(w)$ which is defined as the number of email messages in the collected data set where the word w appears in the document at least once as shown in the formula:

$$W_{ab} = TF_{ab} \cdot \log \chi \frac{S}{DF_i}$$

Where, w_{ab} is the weight of a_{th} Word in the b_{th} document (email), TF_{xy} is the occurrences number of the a_{th} word (w) in the b_{th} document (email), DF_a is the number of email messages in which the i_{th} word(w) occurs, and S , as above, is the total number of messages in the training dataset.

- **Naïve bayes classifiers:** It is a simple probabilistic classifier which works based on Bayes theorem with powerful 'naïve' independence conventions. In 2006, Naïve Bayes classifier was proposed for phishing email filtering in Microsoft (Ganger 2006). Naïve Bayes assumption is that the individual attributes are conditionally independent of each other in classification technique.
- **k-Nearest neighbor(k-NN):** It is a classifier and was proposed for spam email filtering by Gansterer. Using this classifier, the decision is made as follows: Based on k nearest training input, samples are chosen using a pre-defined similarity function; after that, the email 'e' is labeled as belonging to the same class as the bulk of k samples mails.

B. Multiple Classification DM algorithms:

Discretization Filter:

Filters are used to preprocess the data to remove the noisy data in data mining. There are two types of filters supervised and unsupervised filter. Filters can be applied to both training and test dataset. There are various methods for supervised and unsupervised filter. In our work we have used Unsupervised Discretization filter this is to quantize each attribute in absence of any knowledge of the classes of the instances in the training set. Discretize filter converts numeric values to nominal [3] and distributes into selected number of bins equally.

[1] Bayes Classifier

Bayes method is also used for classification in data mining. There are six Bayesian method such as AODE, ADOEs, Naive Bayes, Bayesian Net, Naive Bayes Simple, Naive Bayes updatable. In our work we have used Naive Bayes classification method. Naive Bayes classifier makes assumption about independence of the attributes. Bayes Rules are used to predict the class with some feature values [4].

[2] Function Classifier

Neural Network and regression are the concepts used by Function Classifier. Function classifier can be written down as mathematical equation in natural way. Another methods such as decision tree and rules cannot. The various methods of function classifier are Linear Regression, Logistic, FunctionsLogistic,

RBNFNetwork, etc. For our experiment we have used FunctionsLogistic. FunctionsLogistic builds logistic models, fitting them using LogitBoost with simple regression function base learner as base learner and determining how many iterations to perform using cross validation, which supports automatic attributes selection[3].

[3] Rule Classifier

Interesting relationship among all attributes can be found by using Association Rules. More than one conclusion is predicted by the rule classifier. The correctly predicted records are called coverage. Accuracy is total number of records correctly classified from total number of records. Support is coverage divided by total number of records. Different methods in rules classifier are Conjunctive Rule, Decision table, DTNB, PART, Zero, JRip, Rider. For our dataset we have used PART method. PART obtains rules from partial decision trees [3]. It builds the tree using C4.5's heuristics with the same user defined parameters as J4.8.

[4] Lazy Classifier

Lazy learners store the training instances and do no real work until classification time[3]. It supports incremental learning. Different lazy classifier methods are IB1, IBK, K-Star, LWL, LBR. For our data set we have used K-Star algorithm. K-Star is a nearest neighbor method with a generalized distance function based on transformations.

[5] Meta classifier

Meta classifiers operate on output of other learners and hence called as meta learners. Meta classifier includes large number of classifiers. The most important meta classifiers are Bagging, AdaBoostM1, LogitBoost, MultiClass Classifier, CVParameterSelection, Filtered Classifier. In our dataset we have used Filtered Classifier which runs on filtered data.

[6] Decision Trees

Decision Trees specify the sequence of decision that need to be made along with the resulting recommendation. A Divide and Conquer approach to the problem of learning from a set of independent instances leads naturally to a style of representation called a decision tree[3]. There are different methods for decision tree such as ADTree, BFTree, J48, J48graft, DecisionStump etc. For our data set we have used J48. Based on the highest value of the Information Gain and Entropy, it creates a tree of attributes which depicts the arrangement of attribute in tree structure. Improved version of C4.5 is J48.[5]

C . Classifiers Models-Based Features:

These approaches work based on building full models that are able to create new features with many adaptive algorithms and classifiers to produce the final results.

D. Hybrid Approach:

These approaches work based on combining different algorithms of classifiers working together to enhance results. SVM and Naïve Bayes algorithms if combined together generates better results in terms of more accuracy and less false positive rate for spam email classification.

E. Clustering of spam mails:

Clustering is the process of defining data grouped together according to similarity. It is usually an unsupervised machine learning algorithms. This group of machine-learning techniques depends on filtering phishing emails based on clustering emails via online or offline mode. One of the most used clustering techniques is k-means clustering. K-means clustering is an offline and unsupervised algorithm which begins by determining cluster k as the assumed center of this cluster. Any random email object or features vector can be selected as initial center then undertake the three steps provided below: Determine the Centre coordinate; Determine the distance of each email object(vector) to the Centre group of email objects based on a minimum distance. :

- 1- Determine the distance of each email object (vector) to the Centre
- 2- Group of email object based on minimum distance.

F. Image based Spam Filtering:

It has become a new type of e-mail spam classification. Spammers insert the message into the image and then attach it to the mail. Some traditional methods created on analysis of text-based information do not work in this case. In the paper [11] It is proposed three layer such as Mail Header Classifier, the Image Header Classifier and the Visual Feature Classifier image spam filtering respectively. In the First layer it is applied Bayesian classifier and SVM classifier in the remaining layers. In paper [9] it is offered statistical feature extraction for classification of image-based spam using artificial neural networks. They consider statistical image feature histogram and mean value of block of image for image classification.

No	Techniques Used	Advantages	Disadvantages
1	Bag-of-Words model based methods	-Build secure connection between user's mail transfer Agent and mail user Agent.	-Time consuming - Large number of features -consuming memory
2	Multiple Classification DM algorithms	-Provide clear idea about the effective level of each classifier on spam email detection	These are Non standard classifiers defined
3	Hybrid system	-High level of accuracy by take the advantages of many classifiers	-Time consuming because this technique has many layers to make the final result
4	Classifiers Model-Based Features	- High level of accuracy can be achieved	-huge number of features -many algorithm for classification which means time consuming -higher cost -need large mail server and high memory requirement
5	Clustering Technique	-classification process is fast	-Less accuracy because it depend on unsupervised learning , need feed continuously
6	Image based Spam Filtering	-Efficiency will be more than text based approaches	-It is very Costly -Time Consuming

Table1: Comparison the advantages and disadvantages between machine learning techniques

III. Proposed Work

The figure shows the flow of experiment performed. In this work the performance of various classification techniques are analyzed with and without preprocessing technique.

Steps to apply classification techniques on data set and get result in Weka:

Step 1: Take the spambase dataset as input.

Step 2: Apply the classifier algorithms on the whole data set.

Step 3: Note the accuracy given by it and time required for execution.

Step 4: Repeat step 2 and 3 for different classification algorithms.

Step5: Apply the Discretization filter and Repeat Step2 and Step 3.

Step 6: Results as accuracy among different spam Classifiers are analyzed

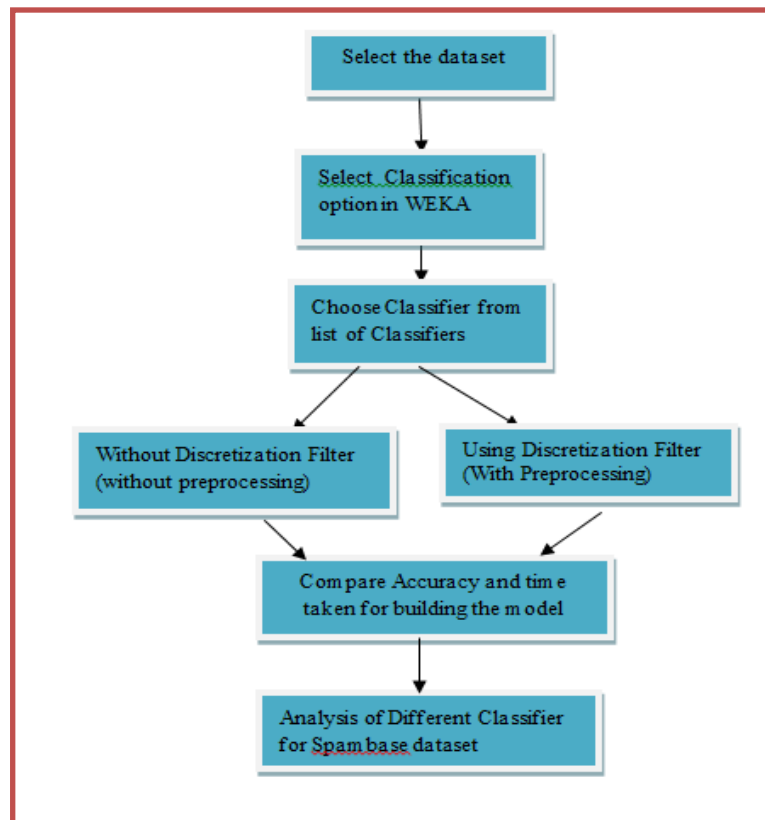


Figure 1: Workflow of Experimented Classification process.

IV. Experimental Result and Discussion

To investigate the performance on the selected classification methods or algorithms namely Bayes-Naive Bayes, Functions-Logistic, J-48, PART, K-Star, FilteredClassifier we use the same experiment procedure as suggested by WEKA i.e preprocessing->classification->Evaluation of different spam classifier Algorithms. In WEKA, all data is considered as instances and features in the data are known as attributes [4]. For experimentation the UCI spambase dataset is used which has a total of 4601 data instances. The results of the simulation are shown in Tables 1 and 2 below. Table 1 mainly summarizes the result based on accuracy and time taken for each simulation when discretization filter is not used. Meanwhile, Table 2 shows the result based on accuracy and time taken for each simulation when Discretization Filter is used for data preprocessing. Figures 1 and 2 are the graphical representations of the simulation result.

Algorithms	Correctly classified Instances(%)	Incorrectly classified instances(%)	Time taken to build model
NavieBayse	79.28	20.71	0.13
Logistic	92.41	7.5	0.65
KStar	93.84	6.15	0
FilteredClassifier	92.80	7.19	0.2
PART	93.58	6.41	1.3
J48	92.57	7.23	0.19

Table 2: Result Before Preprocessing of classification methods

Algorithms	Correctly classified Instances(%)	Incorrectly classified instances(%)	Time taken to build model
NavieBayse	90.19	9.8	0.04
Logistic	94.45	5.54	2.13
KStar	93.84	6.15	0
FilteredClassifier	92.76	7.23	0.12
PART	93.91	6.08	0.36
J48	92.97	7.02	0.42

Table 3: Result After Preprocessing using Discretization Filter of classification method

From above tables 1, 2 we have proven that the correctly classified instances are more if we preprocess the data using Discretization filter. Table 1 shows the performance of different classification algorithms. Logistic classifier has the best performance where 4346 instances are correctly classified from total 4601 instances. From table 2 we can see if the dataset is preprocessed using Discretization filter the performance of the same classification algorithm is increased. With the Logistic classifier the number of correctly classified instances are 4346 out of 4601 instances. Thus from the overall analysis we can say the performance of the classification algorithms can be improved by preprocessing the data set before applying the classification algorithms. With the help of figures we are showing the working performance of various algorithms used in WEKA in fig.1 below. Every algorithm has its own importance and we use them on the behavior of the data.

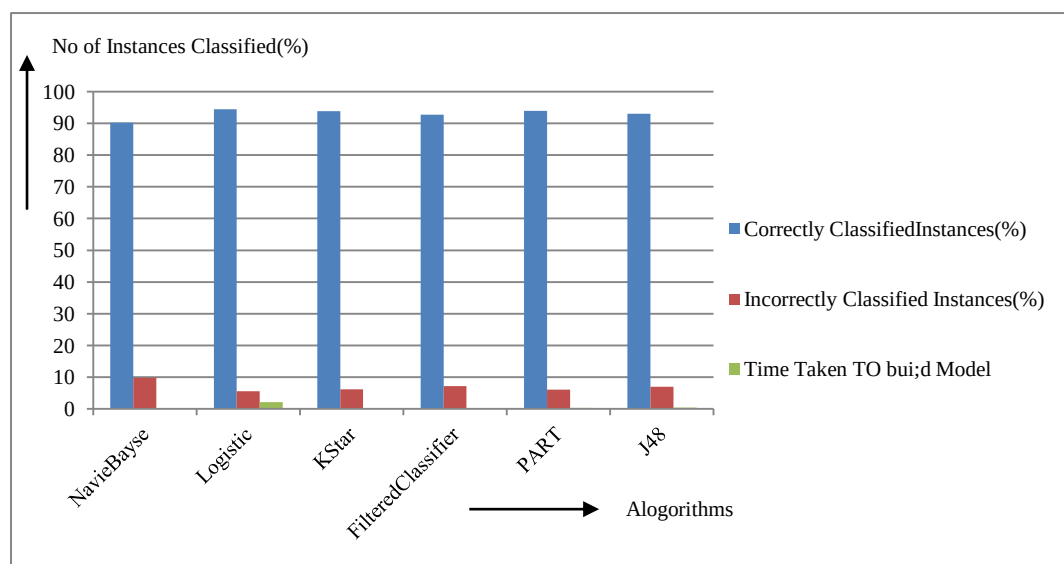


Figure 1: Classification methods comparative performance

V. Conclusion

Spam mails are becoming a very serious problem for today's Internet community, threatening both the integrity of the networks and the productivity of the users. Through this work, we have met our objective which was to analyze the six selected classification algorithms based on Weka and various spam filtering techniques. The result shows the best classifier algorithm is Logistic classifier for UCI Spambase dataset and performance of each of this six algorithm can be improved if the dataset is preprocessed using Discretization Filter. Among the spam filtering techniques described hybrid approach generates the best spam mail filtering results in terms of more accuracy and less false positive rate.

VI. References

- [1] Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, "The WEKA Data Mining Software" An Update.
- [2] Ian H. Witten and Elbe Frank, (2005) "Data Mining Practical Machine Learning Tools and Techniques," Second Edition, Morgan Kaufmann, San Francisco.
- [3] Mohd Fauzi bin Othman, Thomas Moh Shan Yau, "Comparison of Different Classification Techniques Using WEKA for Breast Cancer". Control and Instrumentation Department, Faculty of Electrical Engineering, Universiti Teknologi Malaysia, Skudai, Malaysia.
- [4] Vaithiyanathan1, K. Rajeswari, Kapil Tajane, Rahul Pitale. "Comparison of Different Classification Techniques Using Different Datasets".
- [5] Saadat Nazirova, "Survey on Spam Filtering Techniques" Communications and Network, 2011, 3, 153-160. Published Online August 2011 (<http://www.SciRP.org/journal/cn>).
- [6] UCI Machine Learning Repository: <https://archive.ics.uci.edu/ml/datasets/Spambase>
- [7] R. Kishore Kumar, G. Poonkuzhali, P. Sudhakar, Member, "Comparative Study on Email Spam techniques", IMECS 2012
- [8] Saadat Nazirova "Survey on Filtering Techniques" Science Research-Communication and Network (2011)
- [9] M. Soranamageswari and C. Meena, "Statistical Feature Extraction for Classification of Image Spam Using Artificial Neural Networks," *Second International Conference on Machine Learning and Computing*, Bangalore, 9-11 February, 2010, pp. 101-105.
- [10] W.A. Awad and S.M. Elseuofi, "Machine Learning Methods For Spam E-mail Classification", *International Journal of Computer Science & Information Technology (IJCSIT)*, Vol 3, No 1, Feb. 2011.
- [11] T.-J. Liu, W.-L. Tsao and C.-L. Lee, "A High Performance Image-Spam Filtering System," *Ninth International Symposium on Distributed Computing and Applications to Business, Engineering and Science*, 10-12 August 2010, Hong Kong.