ABSTRACT

Spam refers to immaterial or unwanted messages sent over the Internet to a large number of users to spread inappropriate advertisement messages and malware. The paper aims to propose the most feasible machine learning techniques to be used for spam detection in computer networks. The most common form of spam messages out there are email spams. The purpose of the paper is to provide a comprehensive comparison amongst the different techniques and thereby come up with the most appropriate technique for the same. The paper seeks to describe each technique in detail by providing the way in which it can perform spam detection. The focus is on obtaining a single technique that would be helpful and suitable in all ways to perform spam detection with maximum efficiency. The techniques compared are ZeroR, Bagging, Random Forest, Rotation Forest and J48.

Keywords

Computer Networks, Machine Learning, Data Mining, Spam detection, random forests, rotation forest, bagging.

I. INTRODUCTION

Recently, the rate of spam intrusion has been rising at an alarming rate. There are several data mining, machine learning and artificial neural network algorithms that are used in the personal as well as commercial world to reduce the amount of spam intrusion and detect spam messages[1],[2],[3],[4]. Out of the above mentioned ways, data mining techniques are the most commonly employed techniques to provide spam detection. The spam intrusions are taking place all over the world. It harasses people personally as well as various organizations commercially. Several financial and IT based companies have had problems in the past, and this seems to continue even today. In spite of the extensive development of various data mining and machine learning techniques for spam detection, the problem seems to persist.

With the help of this paper, the comparison of the most efficient techniques will be carried out so that a complete view of the protective measures towards spam intrusion can be looked at. The techniques that are compared are as follows: ZeroR, Rotation forest, Random forest, Bagging and J48.

The techniques are compared on the basis of advantages, disadvantages and classes they support.

The remainder of the paper is organized as follows:

Section II describes the techniques in detail followed by Section III which will provide in depth comparison of the techniques mentioned above. Section IV proposes the most feasible technique that can be used for spam detection. Finally we conclude the paper in the last section.

II. TECHNIQUES

A. ZeroR

The technique is the most basic classification method. It entirely relies on the target class and ignores all predictors[5]. It uses the majority category (class) in order to obtain the classification results. The predictability power in ZeroR is null but due to its usefulness for determining a baseline performance, it can be used as a benchmark for other classification methods.

The following is an example based on ZeroR technique:

![Figure 1: Example of ZeroR on a dummy dataset](image)

For the example shown above the data set has weather conditions as its attributes and play golf as the target class. The ZeroR algorithm will ignore all
the attributes except the target attribute. It does not follow the prediction capabilities as exhibited by other techniques, but simply looks at the final target class. It counts the occurrence of target class and correctly predicts the majority class.

In the scenario for spam detection, the target class will be considered as spam or non-spam. ZeroR will correctly identify the class with maximum occurrences. Hence, the complexity for the algorithm is very low, since it is extremely simple to implement but at the same time the efficiency is very low.

B. Bagging

The Aggregating technique is amongst the most prevalent multiple classifier systems[6]. Bootstrap aggregating, also called bagging, is a machine learning ensemble meta-algorithm. The technique Bagging, increases the efficiency of the prediction of classifiers that are mostly unstable. Initially the process starts with designing bootstrap samples from available training datasets and then they are amassed to produce the bagged predictor.

It is used to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. It is usually applied to decision tree methods.

Bagging generates new training sets from existing one known as bootstrap samples. Assume a standard training set, say D of size n. Bagging creates m new training sets $D_i$, each of size $n'$, by sampling from D.

For spam detection, m models are fitted using the above m bootstrap samples and combined by voting. The classes supported by Bagging technique are Date class, Numeric class, Nominal class, Missing class values and Binary class. The attributes supported by this technique are Binary attributes, Missing values, Unary attributes Numeric attributes, Date attributes, Empty nominal attributes and Nominal attributes. Atleast one instance has to be created in order for this technique to be implemented. The algorithm is as follows:

1. Repeat for $b=1, \ldots, B$
   a) Take a bootstrap replicate $X_b$ of the training set $XTRAIN$.
   b) Construct a base classifier $C_b(x)$ (with decision boundary ($C_b(x)=0$) on $X_b$.
2. Combine base classifiers $C_b(x)$, $b=1,2, \ldots, B$; by the simple majority rule to a final decision rule

\[ \beta = \arg\max_{y \in \{-1,1\}} \sum_{b=1}^{B} \delta_{\text{sign} (C_b(x)),y,\delta_{ij}} \]

where $\delta_{ij}$ is the Kronecker symbol, and $y$ is the possible decision (class label) of the classifier.

C. Random Forest

This technique involves the use of an algorithm that is a classifier and it requires the formation or the growth of several classification trees[7]. The technique is a continuation of the bagging approach that is typically used. Normally, there is always the use of specific attributes and major attributes needed when there is a node split involved in a decision tree, but this approach selects the splitting of nodes in the tree on the basis of a random number of attributes and hence the name Random Forest. The technique will therefore be helpful in reducing the number of features that can be taken into consideration. This will thereby help in scaling, inter-connection amongst the features and will reduce the amount of noise or disturbance involved in the data. The Random Forest algorithm is one of the ensemble methods used for the creation of multiple number of CART-like trees. The trees are trained individually on a bootstrapped sample of the original training data. A majority vote of the trees is then needed to determine the output. The RF algorithm needs to determine a split for each node involved and for that it will randomly search for the subset of the input variables. The initial value is set to the square root of the number of inputs.

The most important advantage of this technique over the other techniques like bagging is that the computational complexity has reduced to a great extent because of limiting the input variables, and this has made the algorithm relatively simpler to be applied. Secondly, since the trees are not cut, the algorithm can easily manage dimensional data. Thus this technique, if used effectively, can be helpful in the reduction of spam intrusion and thereby help in spam detection.

D. Rotation Forest

This technique is again a modification of the bagging technique. The crucial point to be considered while using this technique is that in order to provide and implement diversity, the technique has been brought to action. Similar to the bagging technique, here too the ensemble is consisting of decision trees to be built on bootstrap samples[8]. The difference is in the formation or the
construction of the decision tree. The random forest algorithm considered a random number of features to be selected whereas in this case while splitting a node, a particular feature has to be considered to be selected as the best feature among a subset of L randomly chosen features, where L is a parameter used in the algorithm. The smallest of changes are used in the techniques mentioned, but the result has been that this change has caused the accuracy of the algorithm to increase manifold. It has not compromised the ability of the classifiers as well. The name of this technique is apposite because the decision trees that we chose here are sensitive to rotation of the feature axes. The independently trained decision trees makes a major part of this technique.

E. J48

J48 algorithm is basically an implementation of C4.5 algorithm in Java. It is used in the data mining tool WEKA.

In WEKA J48[9] can be accessed via the decision trees option under the Classify tab. Training sets can be imported in WEKA[10] in the ARFF format. Then by using filtering methods such as Ranker method redundant features can be identified and subsequently removed. J48 is an evolution of ID3 algorithm. ID3 works only on nominal attributes whereas J48 works on nominal as well as numeric attributes. Like ID3, J48 works on the concept of entropy. At each node of the tree, J48 chooses the attribute that most effectively splits its set of samples into subsets enriched in one class or another. The splitting criterion is the difference in entropy. J48 creates pruned decision trees. The above figure represents an example of the decision tree that will be constructed while using the J48 algorithm. This figure represents the decision tree created during implementation in WEKA. After applying the J48 algorithm to a dataset in WEKA, the window that appears on clicking the visualize option, will give the above result. For applying J48 algorithm the data set in arff format must have the target attribute as spam and not spam. J48 will produce a decision tree which will describe the conditions for a message to be spam. Then using this logic, detection of spam message can be done.

III. COMPARISON

The below table represents a comparison amongst the above mentioned five techniques:

<table>
<thead>
<tr>
<th>Technique</th>
<th>Classes</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
</table>
| ZeroR          | Numeric class, Date class, Nominal class, Missing class values, Binary class | 1. Simplest of all classification algorithms.  
2. Easy to implement. | 1. Relies only on the target attribute.  
2. Ignores all other attributes in the dataset. |
| Bagging        | Nominal class, Missing class values, Date class, Numeric class, Binary class | 1. Uses a specific or major attribute for a node split in a decision tree.  
2. Performs classification by voting. | 1. Difficult to imply combining rules other than voting.  
2. It degrades the performance of certain stable methods. |
| Random Forest  | Nominal class, Binary class, Missing class values | 1. Selects the splitting of nodes based on a random number of attributes.  
2. Helps in scaling, improving the interconnection amongst the features, reducing the noise. | 1. Atleast one instance is required to apply the algorithm as compared to J48 which can work with 0 number of instances. |
| Rotation Forest| Nominal class, Missing class values, Binary class | 1. It chooses a particular feature from a subset of randomly chosen features.  
2. Reduces the complexity of the decision tree. | 1. It is not possible every time to come up with the accurate feature based on random selection. |
| J48            | Nominal class, Missing class values, Binary class | 1. It creates pruned decision trees to reduce the complexity.  
2. It handles numeric attributes which are not considered by its predecessor ID3 algorithm. | 1. Lower confidence factor incurs more pruning. |
IV. PROPOSAL

After having done a comprehensive study of all the five techniques for spam detection during computer network intrusion, we propose the best technique to be Rotation Forest. The very first technique considered was ZeroR. It considered only the target attribute. It had practically no predictable power. Since, it could only classify the target class based on the number of occurrences, it serves only as a benchmark for comparing other classification algorithms. The second technique discussed is Bagging. It creates bootstrap samples and performs classification with the help of majority voting. But, it is difficult to imply combining rules other than voting. It also degrades the performance of certain stable methods. The next technique which is Random Forest is truly reaching up to the task. It helps in selecting the splitting of nodes based on a random number of attributes. It also helps in scaling, improving the inter-connection amongst the features and reducing the noise. But it also has the disadvantage of requiring at least a single instance to be implemented. The J48 algorithm creates pruned decision tree, to reduce the complexity. The low confidence factor creates the problem of more pruning or cutting the decision trees. But the Rotation Forest technique is the best to be considered amongst all the above mentioned ones. It selects a feature from a subset of randomly selected features. It reduces the complexity manifold due to this approach.

V. CONCLUSION

The purpose of this paper was to make a comparative study of the five important techniques involved in spam detection. The major goal was to come up with a single technique that would be most beneficial and would provide the highest efficiency versus the remaining. The initial technique reviewed was concerned with providing a benchmark for the other techniques. The other four techniques had some great advantages to rule out the slight disadvantages that they possessed. The J48, Random Forest, Rotation Forest and Bagging were the techniques that provided major advantages and applications that can be successfully used in real life scenarios for spam detection, personally as well as commercially. For future works, a spam database can be used for testing and checking experimental results to determine the best possible technique.

REFERENCES

[9] Shailendra Sahu, B M Mehtre, Network Intrusion Detection System Using J48 Decision Tree