Spam Email Detection on Data Mining: A Review

Elifenesh Yitagesu Desta
Lecturer,Department of Computer Science, College of Computing, Madda Walabu University
POBox 247, Bale Robe, Ethiopia

Abstract

As we know email is an effective tool for communication and it is the fastest way to send information from one place to another and it saves time and also cost. But the email is affected by attacks which include spam mails. Spam is unwanted email or it is bulk data that is flooding the internet with many duplication of similar message, in an attempt to force the email on people who would not otherwise choose to receive it. To address the growing of spam email on the internet the interest of spam filtering also grow accordingly. In this paper we review various spam detection technics. We are use the technics with feature selection algorithm and without feature selection algorithm and apply all the classifier of data mining tool. In this study we analyze the classifier algorithm using two different data mining tools those are WEKA and TANAGRA. Data mining is the discovery of knowledge from the large database and it is the technique of finding out new patterns in a huge data sets. Both data mining tool use different classification algorithms like K-Nearest Neighbor (K-NN), Naïve Bayes (NB) and others. Then finally, the best classifier for email spam is identified based on the accuracy of the algorithm on each data mining tools.

Keywords: Classifier, Feature selection, Spam E-mail.
DOI: 10.7176/JIEA/9-2-01
Publication date: April 30th 2019

Introduction

Electronic mail (e-mail) is now a day it is an effective way of communication and it provide a way for internet user to easily transfer the information all over the world and it saves time and cost as well. So this makes it as a favorite means of communication. But there is a case when the email are affected by attacks. Occasionally we receive e-mail from unknown source and also e-mail comprised of contents which is no important to the user. These kind of unwanted or bulk mails are known as spam mails. Spam email is a part of electronic spam involving nearly identical messages sent to various recipients by email and also it is the practice of frequently sending unwanted message or bulk data in a large quantity to some email accounts.

To say the email is Spam email that meets the following three criteria:

1. Anonymity: the address and identity of the sender are concealed
2. Mass mailing: the email is sent to large group of people
3. Unsolicited: the email is not requested by recipients.

Spam in emails is one of the most complex problems in email services. A lot of work has been done on spam filtering and the ways of evaluation and comparison on different filtering methods .Most of the spam filters are based on machine learning classification techniques

In data mining several classification algorithms are used for classification of spam email. Which are extensively utilize and analyze out of which support vector machine, Naive Bayes, Decision tree, neural network classifiers, Random forest and Random tree are well known classifiers. Initially we experiment on entire data set which consists of total 58 attributes and total number of instances is 4601. We apply above mentioned algorithm one by one on the data set and check the result and it is retrieved from the study that out of all these classifiers Random Forest and Random Tree works well and gives accuracy better than other classifiers in detection of spam mails. In order to compare the result that classifiers works well with some attributes selected or not, then we apply Feature selection algorithm on the same dataset (the algorithm we used here is Best First Search algorithm) and apply the same classifiers with features selected. As we compared with all classifier Random tree shows better result in accuracy. (megha rathi, vikas pareek, 2013)

We address the issue of anti-spam filtering with the aid of machine learning. We examine supervised learning methods, which learn to identify spam e-mail after receiving training on messages that have been manually classified as spam or non-spam (hereafter *legitimate*). (Ion androutsopoulos, georgios pailouras, vagelis karkaletsis, georgios sakkis, constantine D.spyropoulos and panagiotis stamatopoulos)

We use the data mining tools or the machine learning to classify the emails or detect the junk email those tools are WEKA tool and TANAGRA tool.

Data Mining

Data mining is the process of finding anomalies, patterns and correlations within large data sets to predict outcomes. Data Mining is the discovery of knowledge from the large database. It is a technique that attempts to find out new patterns in huge data sets. It is combination of various fields like Artificial Intelligence, Machine Learning, statistics, and Database systems.

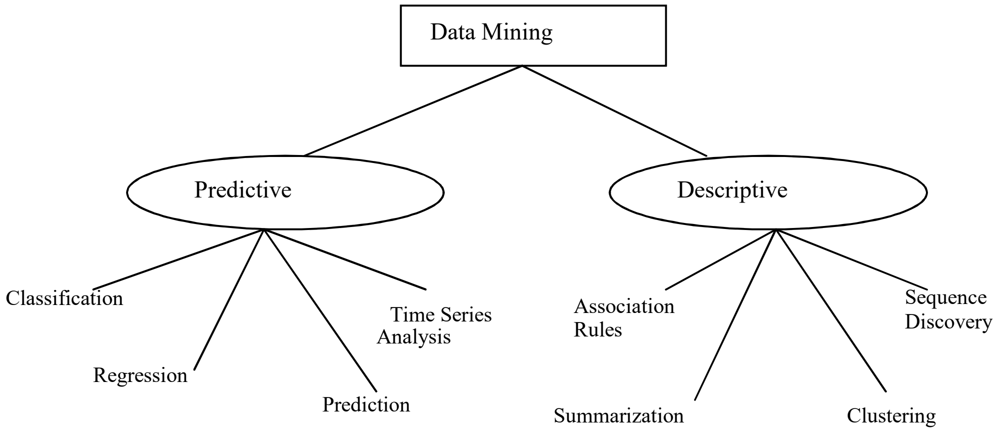

Figure 1: Data mining model (megha rathi, vikas pareek, 2013)

The main objective of data mining approach is to extract information from a data set and transform it into and understandable form for further use. The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously undetermined interesting patterns. Data Mining is the process of analyzing data from different perspective and summarizing it into useful information and this information can be used to increase revenue, cut costs, for classification, prediction etc. It is the process of finding correlations in large relational databases. (megha rathi, vikas pareek, 2013)

Spam Dataset

The spam dataset was taken from UCI machine learning repository and was created by Mark Hopkins, Erik Reeber, George Forman, and Jaap Suermondt. Hewlett-Packard Labs. This dataset contains 4601 instances and 58 attributes (57 continuous input attribute and 1 nominal class label target attribute). (UCI Machine Learning Repository, n.d.)

Machine Learning Methods Performance

This section presents an overview of three different methods on how to detect spam mails what kind of algorithm is used and what are those experiments are performed.

In the first paper we review the data mining algorithms that used for spam mail detection and the experiments which is performed by those algorithms; those algorithms are:

Support vector machines are supervised learning models with associated learning models that analyze data and are mainly used for classification purpose. Support vector machine (SVM) takes a set of input data and output the prediction that data lies in one of the two categories i.e. it classify the data into two possible classes. Given a set of training examples, each marked as belonging to one of the two classes, an SVM training algorithm build a model that assign new data in one class or the other.

A Naive Bayes classifier is a simple probabilistic classifier with strong independence assumptions. In simple terms, a Naive Bayes classifier assumes that the presence/absence of a particular feature of a class is unrelated to the presence/absence of any other feature, given the class variable depending on the nature of probability model

A Decision tree is a classification method that results in a flow-chart like tree structure where each node denotes a test on attribute value and each branch represents an outcome of the test. The tree leaves represents the classes. Decision tree is model that is both predictive and descriptive; it represents relationships found in training data. The tree consists of zero or more internal nodes and one or more leaf nodes with each internal node being a decision node having two or more child nodes

Feature Selection

Feature Selection also known as feature reduction, attribute selection is the technique of selecting a subset of relevant features for building the learning models. Feature selection is very important step in analyzing the data, by removing irrelevant and redundant features from the data. Feature selection overall improves the performance of learning model by:

1) Alleviating the effect of curse of dimensionality.
2) Enhancing generalization capability.
3) Speeding up learning process.
4) Improving model interpretability.

Feature Selection helps in gaining the better understanding of the data by telling which are the important attributes or features and how they are related with each other. It is the process of selecting a subset of Spam Mail Detection through Data Mining

In the first experiment, to validate the proposed scheme for spam mail detection we conduct several experiments. The mail goal is finding the best classifier whose accuracy is better than the rest of classifiers. Spam base data set is used for experiment which consist 57 attributes with one target attribute in discrete format. From the classification algorithm 8 classification algorithms like Navïe Bayes, Bayesian Net, Support Vector Machine (SVM), Function Tree (FT), J48 Random Forest, Ra.ndom Tree and Simple cart are applied one by one on the dataset. From the algorithms Random forest achieves highest accuracy that is 94.82%. In the second level highest accuracy is achieved by function tree the accuracy is 93.34% and so on. So from this study it is found that tree like classifier performs well in case of classification of spam mails.

In the second experiment before applying the classification algorithms first applied Best-first-feature selection algorithm for selecting a subset of features from the given data set. 58 attributes was present in the present in the given dataset. But after applying Best-First-algorithm on the given data total 15 attributes are selected then we apply classifiers on this reduced data set for the detection of spam mail. The highest accuracy is achieved by Random Tree classifier and in the second place Random Forest also score the second highest accuracy. (megha rathi, vikas pareek, 2013)

In (R.Kishore Kumar, G.Poonkuzhali,P.Sudhakar, member,LAENG, 2012) study spam dataset is analyzed using the datamining tool is called TANAGRA to explore the efficient classifier for email spam classification. Hear applied feature construction and feature selection is done to extract the relevant features. Then various classification algorithms are applied over this dataset and cross validation is done for each of these classifiers. This study also use the spam base dataset that is downloaded from the UCI machine learning repository in the form of text file This dataset contains 57 input attributes of continuous format and 1 target attribute in discrete format .the experimental result and performance evaluation is Then feature construction is done for feature transformation. Since the training dataset contains all the input attributes as continuous and target attribute as discrete, the following four feature selection algorithms namely, Fisher filtering, ReliefF, Runs Filtering and Step disc are executed on this dataset for retrieving relevant features. Classification algorithms such as Naive Bayes continuous, ID3 ,K-NN, multilayer perceptron, CSVC, Linear discriminant analysis, CS-MC4, Rnd tree, PLS-LDA, PLS-DA etc., are applied to each of the above filtering algorithms. Runs filtering and Step disc feature selection algorithms almost provide the same result. From the results, the Rnd tree classification is considered as a best classifier, as it produced 99% accuracy through fisher filtering feature selection. (R.Kishore Kumar, G.Poonkuzhali,P.Sudhakar, member,LAENG, 2012)

Tanagra is a free suite of machine learning software for research and academic purposes developed by Ricco Rakotomalala at the Lumière University Lyon 2, France.[1] Tanagra supports several standard data mining tasks such as: Visualization, Descriptive statistics, Instance selection, feature selection, feature construction, regression, factor analysis, clustering, classification and association rule learning. (Tanagra(Machine learning), 2018)

In (Vahid Nasir1 · Sepideh Nourian1 · Stavros Avramidis1 · Julie Cool1, 2018) the performance of artificial neural networks (ANN), support vector machines (SVM), and naïve Bayes (NB) classifiers for thermowood classification was evaluated and compared. The results showed that mechanical attributes such as dynamic modulus of elasticity obtained from the stress wave timer test and wood hardness account for the least suitable features, whereas color measurement provided an accurate classification. Both SVM and naïve Bayes model showed significantly higher performance than ANN because the latter requires a higher number of tuned and optimized parameters

CONCLUSION

Spam in emails is one of the most complex problems in email services. A lot of work has been done on spam filtering and the ways of evaluation and comparison on different filtering methods .Most of the spam filters are based on machine learning classification techniques. Email spam detection has received a remarkable attention by majority of the people as it helps to identify the unwanted information and threats. In data mining several classification algorithms are used for classification of spam email. Which are extensively utilize and analyze out of which support vector machine, Naive Bayes and Decision tree. Therefore, most of the researchers pay attention in finding the best classifier for detecting unsolicited mail. From the achieved results, fisher filtering and runs filtering feature selection algorithms performs better classification for many classifiers.

References

Ion androutsopoulos, georgios pailouras, vagelis karkaletsis, georgios sakkis, constantine D.spyropoulos and panagiotis stamatopoulos. (n.d.). Learning to Filter Spam E-Mail: A Comparison of a Naive Bayesian and a Memory-Based Approach.

megha rathi, vikas pareek. (2013). spam mail detection through data mining- a comparative performance analysis.

international journal of modern education and computer science , 31-39.

R.Kishore Kumar, G.Poonkuzhali,P.Sudhakar, member,LAENG. (2012). Comparative Study on Email Spam Classifier using Data Mining Techniques. *Proceeding of the international multi conference of Engineering and Computer Scientists*.

Tanagra(Machine learning). (2018, december 24). Retrieved from Wikipedia The Free Encyclopedia: https://en.wikipedia.org/wiki/Tanagra_(machine_learning)

UCI Machine Learning Repository. (n.d.). Retrieved from Spambase Dataset: http://archive.ics.uci.edu/ml/datasets/Spambase

Vahid Nasir1 · Sepideh Nourian1 · Stavros Avramidis1 · Julie Cool1. (2018). Classification of thermally treated wood using machine learning techniques. *Wood Science and Technology*.