Performance Evaluation for Different Classification Techniques of Spam mail using Weka

C. Neelavathi[1] Dr.S.M.Jagatheesan[2]
[1]Research Scholar [2]Associate Professor
[1,2]Department of Computer Science
[1,2]Gobi Arts & Science College, Gobichettipalayam, Tamilnadu, India

Abstract— Data mining is process of discovering new knowledge from database. Email is one of the essential factor of data communication. Spam emails are emails that the receiver does not wish to receive. Spam email is causing series problems for internet users, Internet service provider and the whole internet backbone network. In this work various spam email techniques are discussed. Using Spam email dataset Preprocessing and classification is workout in Weka Explorer. The dataset is taken from UCI repository. Multiple classification Techniques can be examined to get better result. By analyzing all these techniques, the Random Tree with Partition Membership Filter gives better Performance than others. Here Weka tool is used as a software tool to analyze result.

Key words: Data Mining, Spam Mail, Weka, Classification Algorithm, Filter and Random Tree

I. INTRODUCTION

Data mining [7] sometimes called data or knowledge discovery is the process of analyzing data from different perspectives and summarizing it into useful information. Data mining consists of number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it & summarize the relationship identified. Technically Data mining is used to finding correlations or patterns among dozens of field in large relational database. Data mining [6] involves data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. Data mining is categorized as medical Data mining, Spatial data mining, Sensor data mining, Visual data mining, Music data mining, Pattern data mining, Subject based data mining and Web mining.

II. SPAM EMAIL TECHNIQUES

Email spam [8] targets individual users with direct mail messages. Email spam lists are often created by scanning Usenet postings, stealing Internet mailing lists, or searching money out-of-pocket to receive. On top of that, it costs money for ISPs(Internet Service Provider) and online services to transmit spam, and these costs are transmitted directly to subscribers. Web mining can reveal security holes and weakness, for instance, Trojan horses that continually send page request to a server can cause the server to go slow. These can be identified by mining the user logs. Spam mining is a common technique used by email site in which links that generate spam are grouped together to classify received mail as spam. Most popular email sites have options for automatic spam blocking by mining the email headers, subject lines, content and attachments. Spam emails not only waste resources such as bandwidth, storage and computation power. In supervised spam blocking each user can identify links that generated spam. This information is stored on the servers, can come up with a global spam blocking list or separate list for users in different domain or areas.

A. Image spam Email:

Image spam [9], or Image-based spam, is a hiding/dishonest method in which the text of the message is stored as a GIF (Graphic Interchange Format) or JPEG (Joint Photographic Experts Group) image and displayed in the email. This prevents text based spam filters from detecting and blocking spam messages. Image spam was used in the "pump and dump" stocks. Often, image spam contains (filled with nonsense), computer-created text which simply annoys the reader. However, new technology in some programs tries to read the images by trying to find text in these images. They are not accurate (very close to the truth or true number), and sometimes filter out innocent images of products like a box that has words on it. A newer way of doing things, however, is to use a (full of life and energy/moving) GIF(Graphic Interchange Format) image that does not contain clear text in its initial frame, or to twist the shapes of letters in the image to avoid detection by OCR(Optical Character Recognition) tools.

B. Blank Spam Email:

Blank spam is spam missing a payload advertisement. Often the message body is missing completely, as well as the subject line. Still, it fits the definition of spam because of its nature as bulk and (without being requested) email. Blank spam may be started in different ways, either (on purpose) or accidentally.

C. Phishing Fraud Email:

Phishing misrepresentation email spam endeavors to trick email beneficiaries into believing that the message originated from another person (i.e. a known organization, brand, bank or money related foundation). Phishing spam is sent to trick recipients into visiting a website and provide financial information such as a credit card number or other sensitive data. This type of spam is mis-representative and potentially damaging to the recipient.

D. Emotional Spam Email:

This type of spam attempts to trick email recipients by playing on their emotions. This is done primarily in two ways; either by raising the hopes of the recipient with suggestions of monetary winnings (i.e. lottery or International Money Transfer scams) or by playing on their sentimental nature (illness, romance, etc.). A good example of this is the Microsoft Global Email Lottery Spam. Emotional Spam email is meant to envoke an emotional response in the recipient and entice them to take action and eventually send money to the spammer either to “claim a prize” or to provide relief to the sender who is supposedly

suffering somehow. These fraudulent scam emails might seem obvious to some; but they fool enough people to make it worthwhile for spammers to continue sending them.

E. Virus and Malware Email:

This type of email spam will include an attachment or a link to a file that will trigger some virus of malware to install to the recipients computer when the attachment is opened or the link visited. Sometimes they are sent automatically from one computer to others using scripts in the virus that cause the recipient computer to forward it on to other people in their address book without the recipient's knowledge. The objective of virus spam is to spread computer viruses across networks via email.

F. Replica Watch, Stock and Weight Loss Email:

This type of spam is very common. People are trying to purchase prescribed medicines, replica watches or stock of some kind of products through email. This type of spam doesn't pretend to be anything but an unsolicited attempt to get people to click and people do (one in every 12,500,000 spam messages that are sent!). The objective of Replica Watches, Stock and Weight Loss spam is to market products and services to a large audience in hopes of getting a percentage of the millions who receive it to click and convert. Stock spam has a secondary objective in that they try to drive a stock's price up and then the sender sells it off shortly after the stock climbs as a result. This type of spam burdens are affected computer resources and networks. As a result, it costs companies and individual's time, productivity and money for anti-spam technology required to reduce it. These types of spam mails are analyzed by using Weka tool.

III. WEKA TOOL

Weka[10] is a machine learning software tool. It can be developed by University of Waikato in Newzealand. This is a GUI tool that allows executing algorithms which gives better results than other tools. For beginners it is easy to use. Weka tool is applicable for those computers such as linux or Mac, Windows. User can choose one of the Explorer, Experimenter, Knowledge flow and Simple CLI.

A. Weka GUI Chooser

- Simple CLI - provides users without a graphic interface option the ability to execute commands from a terminal window.
- Explorer - the graphical interface used to conduct experimentation on raw data.
- Experimenter - this option allows users to conduct different experimental variations on data sets and perform statistical manipulation.
- Knowledge Flow - basically the same functionality as Explorer with drag and drop functionality.

B. Weka Explorer

Weka Explorer consists of six tabs:

- Preprocess - used to choose the data file to be used by the application.
- Classify - used to test and train different learning schemes on the preprocessed data file under experimentation.
- Cluster - used to apply different tools that identify clusters within the data file.
- Association - used to apply different rules to the data file that identify association within the data.
- Select attributes - used to apply different rules to reveal changes based on selected attributes inclusion or exclusion from the experiment.
- Visualize - used to see what the various manipulation produced on the data set in a 2D format, in scatter plot and bar graph output.

C. Preprocessing

In order to experiment with the application the dataset needs to be presented to WEKA in a format that the program understands. There are rules for the type of data that WEKA will accept. There are three options for presenting data into the program.

- Open File - allows for the user to select files residing on the local machine or recorded medium.
- Open URL - provides a mechanism to locate a file or data source from a different location specified by the user.
- Open Database - allows the user to retrieve files or data from a database source provided by the user.There are restrictions on the type of data that can be accepted into the program. Originally the software was designed to import only ARFF(Attribute Relation File format) files, newer versions allow different file types such as CSV(Comma Seperated Value) and serialized instance formats.

D. Classify

The user has the option of applying many different algorithms to the data set that would in theory produce a representation of the information used to make observation easier. It is difficult to identify which of the options would provide the best output for the experiment. The best approach is to independently apply a mixture of the available choices and see what yields something close to the desired results. There are several options to be selected inside of the classify[3] tab. Test option gives the user the choice of using four different test mode scenarios on the data set:

- Use training set
- Supplied training set
- Cross validation
- Split percentage

There is the option of applying any or all of the modes to produce results that can be compared by the user.

E. Cluster

The Cluster tab opens the process that is used to identify commonalties or clusters of occurrences within the data set and produce information for the user to analyze.

F. Associate

The associate tab opens a window to select the options for associations within the data set. The user selects one of the choices and presses start to yield the results.

G. Select Attributes

The next tab is used to select the specific attributes used for the calculation process. By default all of the available attributes are used in the evaluation of the data set. This is useful if some of the attributes are of a different form such as alphanumeric data that could alter the results.

H. Visualization

The last tab in the window is the visualization tab. There are a few options to manipulate the view for the identification of subsets or to separate the data points on the plot. Polyline can be used to segment different values for additional visualization clarity on the plot. This is useful when there are many data points represented on the graph. Rectangle tool is helpful to select instances within the graph for copying or clarification. Using Polygon tool, users can connect points to segregate information and isolate points for reference.

IV. RELATED WORK

A. Supervised learning and Cross Validation

Supervised Learning: Supervised learning[5] is perhaps the most frequently used mining/learning technique in both practical data mining and Web mining. It is also called classification, which aims to learn a classification function (called a classifier) from data that are labeled with pre-defined classes or categories. The resulting classifier is then applied to classify future data instances into these classes. Due to the fact that the data instances used for learning (called the training data) are labeled with pre-defined classes, the method is called supervised learning.

B. Cross-Validation:

When the data set is small, the n-fold cross-validation method is very commonly used. In this method, the available data is partitioned into n equal-size disjoint subsets. Each subset is then used as the test set and the remaining n - 1 subsets are combined as the training set to learn a classifier. This procedure is then run n times, which gives n accuracies. The final estimated accuracy of learning from this data set is the average of the n accuracies. 10-fold and 5-fold cross-validations are often used.

C. Decision Tree Classifier:

A Tree is a convenient way to break a large dataset into smaller ones. By representing a learning set to the root and each interior node, the data at the leaves can often be analyzed very simply. For example a classifier to predict the likelihood that a credit card transaction is fraudulent may use an interior node to divide a training data set into two sets, depending upon whether or not five or fewer transactions were processed during the previous hour. After a serious of each leaf can be labeled fraud/no-fraud by using a simple majority vote. Tree based classifier were independently in information theory, statistics, pattern recognition and machine learning. Random Tree[1] algorithm is used to analyze dataset.

D. Naive Bayes Classifier:

Naive Bayes[4] classifier is a simple probabilistic classifier based on applying Bayes' true idea with strong independence ideas. A more descriptive term for the hidden chance model would be "independent feature model". Multinomial Naive Bayes classifier is used to analyze dataset. Multinomial Naive Bayes classifier is a specific instance of a Naive Bayes classifier which uses a multinomial distribution for each of the features which implements the naive Bayes algorithm for multinomially distributed data.

E. Rule Classifier:

J48 algorithm is open source algorithm in weka data mining tool. Weka classifier package has its own version of JRip classifier known as j48 or j48graft it is used in weka platform. J48 algorithm is an optimized implementation of JRip classifier. This classifier is experimented in this study with the parameters. For experiment JRip[4] algorithm is used.

F. Function Classifier:

Function classifier consists of Linear Regression, logistic, Functions Logistic, RBnFnetwork, etc. For experiment SGD Algorithm is used to analyze dataset [2].

G. Meta Classifier:

Meta Classifiers are Bagging, AdaBoostM1, LogitBoost, Multiclass classifier, CVParameterSelection and Filtered Classifier. The Filtered Classifier is an easy way of filtering data on the fly. It removes the necessity of filtering the data before the classifier can be trained. Also, the data need not be passed through the trained filter again at prediction time.

H. Lazy Classifier:

Lazy classifier methods are IB 1, IBK, K-Star, LWL and LBR. For experiment, K-star algorithm is used. K-Star is an instance-based classifier that is the class of a test instance is based upon the class of those training instances similar to it, as determined by some similarity function. It differs from other instance-based learners in that it uses an entropy-based distance function. These various classifiers are experimented and the results are shown below.

V. EXPERIMENTAL RESULTS

In this survey the accuracy of six data mining technique is compared by using the Email spam dataset[2]. Different performance matrix such as TP rate, FP rate and Precision, Recall, F-measure and ROC area are reported. In supervised machine learning technique partition membership filtering technique is used.

Fig. 1: Performance Measure

In Figure 1 Performance measure of True Positive (TP) rate, False Positive rate(FP), Precision, Recall is defined. Comparing with Random Tree algorithm with other algorithms it gives low false positive rate from the chart.

Random Tree algorithm gives accuracy rate of 97.08 %, kappa statistics 0.938, Time Taken to build model 0.56 secwhich establishes best result than other algorithms. JRip algorithm gives accuracy rate of 96.47%, kappa statistics 0.926, Time Taken to build model 33.62sec.Filtered Classifier gives accuracy rate of 93.87%, kappa statistics 0.873, Time Taken to build model 2.75sec. K-Star algorithm gives accuracy rate of 97.04 %, kappa statistics 0.937, Time Taken to build model 0 sec.

SGD gives accuracy rate of 97.04%, kappa statistics 0.937, Time Taken to build model 5.44sec, mean absolute error of 0.029 %, root mean squared error 0.171%, relative absolute error of 6.189%. Multinomial algorithm gives accuracy rate of 92.80%, kappa statistics 0.849, Time Taken to build model 0.02 sec, mean absolute error of 0.077 %, root mean squared error 0.258%, relative absolute error of 16.30%.

Performance Algorithm	Accuracy	Kappa Statistics	Time Taken to Build model	Mean Absolute Error	Root mean Squared error	Relative Absolute Error
Random Tree	97.08	0.938	0.56	0.05	0.16	11.36
JRip	96.47	0.926	33.62	0.063	0.182	13.38
Filtered Classifier	93.87	0.873	2.75	0.104	0.230	21.83
K-star	97.04	0.937	0	0.054	0.169	11.37
SGD	97.04	0.937	5.44	0.029	0.171	6.189
Multinomial	92.80	0.849	0.02	0.077	0.258	16.30

Table 1: Performance Comparision

The Table 1 and Figure 2 describes the accuracy rate, Kappa Statistics[6], Time Taken to bulid model, Mean Absolute error, Root mean squared error, Relative absolute error. From the representation Random Tree is better than all other algorithms.

Fig. 2: Line Chart Representation of algorithms.

VI. Conclusion

This proposed method provides the analytical result of six selected classification algorithm, in Weka Tool. The Performance of each of six algorithms can be improved, if the dataset is preprocessed using Partition membership Filter. The result shows the best classifier algorithm (i.e) Random Tree classifier, for UCI Spambase dataset. The Random Tree classsifier generates the best outcome in terms of accuracy, kappa statistics and less error rate. This analysis is more beneficial part of web mining.

References

[1] Christina, "A Study on Email Spam Filtering Techniques", International Journal of Computer Applications (0975 –8887) Volume 12–No.1, December 2010, pp.07-09.
[2] J. S Kong, B. A. Rezaei, N. Sarshar, V. P. Roychowd-hury and P. O. Boykin, "Collaborative Spam Filtering Using E-Mail Networks," IEEE Computer Society on Computer, Vol. 39, No. 8, 2006, pp. 67-73.
[3] P. Mishra, N Padhy, R Panigrahi ," The Survey of Data Mining Applications And Feature Scope", International Journal of Scientific & Engineering Research Volume 4, Issue3, March-2013 8,ISSN 2229-5518.
[4] Pang Ning Tan, "Introduction to data mining", Addision Wesley Publication,2002
[5] R. Kishore Kumar, G. Poonkuzhali, P. Sudhakar "Comparative Study on Email Spam Classifier using Data Mining Techniques", IMECS2012, p539-544 .
[6] S. Yirong, and J. Jing," Improving the Performance of Naive Bayes for Text Classification", CS224N Spring 2003.
[7] Samuel Kovac, "Suitability analysis of data mining tools and methods", Masaryk University Faculty of Informatics.
[8] Spam email dataset available online at: http://archive.ics.uci.edu/m1/datasets/Spambase.
[9] Supriya S.shinde "Improving spam mail filtering using classification algorithms with discretization filter", IJETCAS2010.
[10] Weka is a collection of machine learning algorithms for data mining tasks. http://www.cs.waikato.ac.nz/ml/weka/downloading.html.

