A Review of Various Spam Classification Techniques

Er.Harpreet Kaur
M.Tech Scholar, Department of Computer Science & Engineering,
Amritsar College of Engineering & Technology,
Amritsar, Punjab

Er.Ajay Sharma
Associate Professor, Department of Computer Science & Engineering,
Amritsar College of Engineering & Technology,
Amritsar, Punjab

Abstract- Email has turn into one of the fastest as well as most economical mode of communication. It is an efficient means of communication as it save a bundle of time as well as capital this make it as a favourite means of communication in private as well as in professional e-mail. E-mails offer a way for internet user to simply spread information worldwide in a fraction of second. This continues growth in email users have resulted within the raising of unsolicited emails or junk emails well-known as email spam. Email spam is one of the main problems of the today's internet, Bringing economic harm to companies as well as frustrating human being. Spam are invading user without their authorization as well as filling their mail boxes. They consume more network capability as well as time in examination and deleting spam mail. The majority of the internet user are outspoken in their disdain for spam, while enough of them act in response to commercial offer that spam remains a feasible resource of earning to spammers. As most of the users wish to perform correct imagine avoiding and get rid of spam, they want understandable as well as easy procedure on how to perform. In this paper, we make an effort to study spam and various spam classification techniques that which are used to eliminate the spam mail.

Keywords - Data Mining, Data Mining Algorithm, Email, Spam Email Classification

I. INTRODUCTION

With the modern growth of information technology as well as computer science, high capacity information appears in our life. Advances within the storage machinery and exclusive development within applications like internet search as well as video retrieval have created a lot of high-volume data sets. Most of this data is store digitally into the electronic media, such as hard disk or other electronic media designed for automatic data analysis as well as retrieval techniques. In addition to the increase in the quantity of data, the diversity of existing data has also increased emails, blogs, business data, plus billion of web pages generate tetra bytes of fresh data every day. Many of these data streams are unstructured, adding to the complexity in analyzing them. That raise in both the volume as well as the diversity of data involve advances within the methodology to automatically recognize, process, plus précis the data. These data pools could be use to find a higher class of information than that gain from simple database inquiries. This is where Data Mining comes in.

Data mining is the process of discovering interesting patterns or rules from large amounts of data. Data Mining is the method of taking data as input and produce output as knowledge. The data sources contain databases, data warehouses, the Web, other information repositories, or data that is stream into the system dynamically.

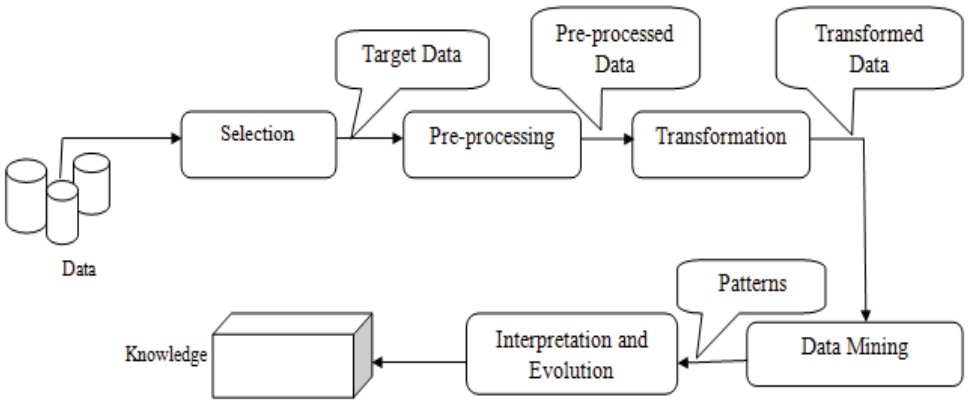

Figure 1. Steps of Data Mining Process

The procedure of Data Mining consists of iterative sequence of steps as follows:-

1. Selection: Selecting data related to the analysis task from the different data sources.
2. Pre-processing: Remove noise, inconsistent data and combining multiple data sources.
3. Transformation: Transformation is the process of transferring the data into appropriate forms so that it can use for data mining.
4. Data Mining: Select a data mining algorithm that is used to form the appropriate patterns or rules in the data or extract data patterns.
5. Interpretation/Evolution: Interpreting the patterns into knowledge by removing redundancy or unrelated patterns; translating the useful patterns into terms of human being understandable format.
6. Knowledge: Knowledge discovery within database is the procedure of discovery useful information and patterns into the data.

A. Data mining application-

Data mining application are widely used in:

- Business transactions
- Medical and personal data
- Surveillance video and picture
- Satellite sensing
- Games
- CAD and software engineering data
- Text reports and memos (e-mail messages)

B. Data mining algorithms & techniques-

Data mining involves various algorithms and techniques to achieve appropriate patterns. All the algorithms are used to inspect the data as well as define the model which is nearby to the features of the data being examined. The model which is used to define the appropriate patterns can be predictive or descriptive.

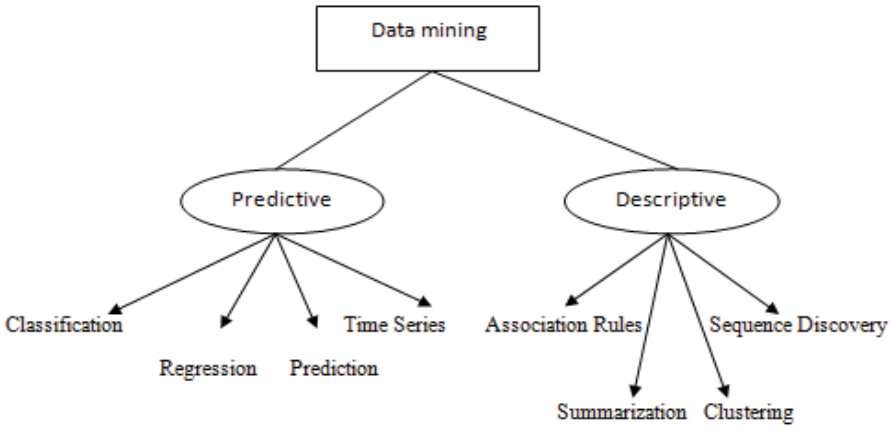

Figure 2. Data Mining Models and Tasks

Clustering - Create groups similar data together into clusters

a) unsupervised learning
b) segmentation
c) partitioning

– Similar to one another within the data are in the same cluster.
– Dissimilar objects within the data are in different clusters.
– Good clustering is defined as:-
  a) High intra-class similarity.
  b) Low inter-class similarity.

Different types of clustering methods are Partitioning Methods; Density Based Methods; Model-Based Methods and Grid-Based Methods.

Association Rules - Association as well as relationship is usually in the direction of locate frequent item set detection along with large data set. Association rules algorithm require to create rules by means of confidence value less than one.

Summarization - Maps data into subsets with associate simple descriptions i.e. Characterization and Generalization.

Time series - It is used to:

a) Predict future value.
b) Determine similar patterns over time.
c) Classify behaviour.
d) Example: - stock market.

Regression - Regression is required to map a data item to real valued predication variables.

Prediction - Perform inference on data to make predication.

Classification - Classification is used to assigning a class label to a set of unclassified data.

a) Supervised learning (classification)
  - The set of possible classes is known in advance.
  - The input data, also define as training set which is consists of number of records and each record has multiple attributes and features.
  - All the records are tagged as a class label.

- The aim of classification is to analyze the input data plus develop an ideal report or model for all classes using the features exist into the data.
- This model is also used to classify the test data for which the class descriptions are not known in advance.

b) Unsupervised learning (clustering)
Set of possible classes is not known in advance. After classification we can make an effort to assign a name to that class. Unsupervised classification is also known as clustering.

C. Email-

Most of the human being on the earth used the emails. There are about more than 3 billion email accounts, approximately half of the world residents. They are expected to achieve 4 billion in 2015. Now the days, kids are also allowed to have their own email accounts under the supervision of their parents.

D. Spam classification-

Spam emails are one of the most complex problems. Spam mails are defined as unwanted, unsolicited emails. This type of emails are sent for either marketing purpose, scams, hoaxes and are not intended for specific receiver. In 2009, 97% approximately of emails are classified as spam so because of this reason many of the research papers are studied and analyzed the emails focused on classification of mails into spam or not. However, the effort between the spammer and spam detection tool is continuous to overcome the technique developed by the other by creating a new way.

Email spam classifier is not only depend upon the classifying the accurate spam email as spam but also classifying non-spam emails as non-spam or normal. Both the conditions are used to define the quality its classifications or prediction. Four predictions are used to evaluate the quality of email. TP (True Positive) define that the spam detection tool predicts that the email is spam and it is truly spam. TN (True Negative) define that the tool predicts that the email is normal and truly it was not a spam. FP (False Positive) define that the tool predicts that the email is spam but truly it was a good email. FN (False Negative) define that the tool predicts that the email is normal but truly it was a spam. As such prefect detection system has some value TP 100%, TN 100%, FP 0%, FN 0%. But in reality such situation are impossible and impractical. TP and FP are complement of each other i.e. total should be 100% and same thing is applied on TN and FN.

There are some challenges which are faced by the email detection system:-

1. If it is restricted through various spam-detection roles, TP can go high, however at the account of getting many false alarms. On the other side, very lean rules may get TN high but at the account of FN.
2. Speed of spam detection system.

In addition to spam based classification, some papers are conducted on the research of spam classification based on the different aspects like folder classification or automatic classification priority base filtering of email messages, e-mails and contacts clustering etc.

Following are the some focuses in the research of email analysis:-

1. Email analysis under text categorization. Different algorithms are used.
2. The major research is based on the classification of spam and non-spam of emails.
3. Some classifications are based on the gender of the sender.
4. Classification also used automatically assign emails to predefined folders.
5. Emails can also classify into interesting as well as uninteresting e-mails.
6. Features are extracting from the contents of title, body or other metadata.
7. Clustering of emails is used to cluster the emails into different subjects and folders.
8. Time information is also used in some of research papers to classify the emails.
9. Some are classified on the base of similar threads or subjects.

II. RELATED WORK

We select some papers which are based on the spam and non-spam classification. Those papers are:-

Kh.Ahmed (2007) paper talked about that spam emails is a major problem which financially damage the companies and annoying individual user. Among all the approaches, filtering is one of the most important techniques to stop the spam. Spam emails are the unsolicited or unwanted or junk mails that are sent to the number of recipients which neither are not required by them. The main task of the spam filter is to filter out all the spam emails automatically from the user's mail stream. These unsolicited mails create lot of problems like bandwidth wastage, consume much space in disk, IT security problems, consume user time etc.

Abdullah, Al-Kabhi (2011) shows that spam is a serious problem. Authors conducted the survey to define the current status of spam mails in KSA. Paper also present the statistics gathered from stakeholders highlights their issues and the measures taken by them to manage spam in their network.

Gunal et al. (2006) presents two different ways to choose the most discriminative feature for spam email detection. Common Vector Approach (CVA) is used to develop the feature choice method which provides considerable reduction on the number of feature without changing the recognition rates.

Rathi et al. (2013) examine the performance of a variety of classifiers with feature choice algorithm and without feature choice algorithm. It shows the improved result with feature choice algorithm which declared that Random tree is a best classifier for spam classification at the rate of 99.72%.

Kumar et al. (2012) examine the various classification algorithms that are applied on the dataset. And at last best classifier is defined depends on the error rate, precision and recall.

Elssied et al. (2014) proposed the hybrid technique of SVM and K-means that use the spam base dataset to judge the feasibility of the technique.

Kumar et al.(2015) proposed a spam detection design which consists of Vector Quantization (VQ) method that remove redundancy from the training and testing data. And after the removal of redundancy from the data, it send to the particle Swarm Optimization (PSO) that mines the optimum features for the classification. Finally, the chosen features are used by the Probabilistic Neural Network (PNN) for the classification of spam with lot of accuracy and precision.

Table – 1 Comparison Table

Ref. No.	Authors	Year	Technique	Feature	limitation
[5]	**Rathi et al.**	**2013**	**SVM (support vector machine), Naive Bayes, Best-First feature selection Algorithm, random tree, Random forest, j48, Bayes Net**	**Analyze the performance of various classifiers on the bases of by means of feature selection algorithm and without feature selection algorithm.**	**The result of classifiers is not accurate without feature selection algorithm.**
[7]	**Elssied et al.**	**2014**	**SVM (support vector machine), k-means**	**Improved the accuracy, reduce the false positive and time cost of the classifier.**	**Not compare the results with different evolutions and opposition based learning as feature selection.**

[9]	**Alsmadi et al.**	**2015**	**SVM (support vector machine), k-means**	**Improve accuracy rate.**	**Unsupervised filter is not used as a pre-processing tool.**
[10]	**Sharma et al.**	**2015**	**Principal Component Analysis (PCA), Correlation Feature Selection Techniques (CFS)**	**Develop feature selection and feature reduction technique**	**These filtering techniques are only used at the client level for classification of spam and ham emails.**
[11]	**Intrajit et al.**	**2012**	**Fuzzy Relational Classification- MOS Entropy Based Weight Matrix and Fuzzy Partition Matrix**	**Feature entropy based on image quality classification.**	**Enhance the quality of the image by parametric learning methods.**
[12]	**Dhillon et al.**	**2011**	**k-means, Vector Space Model**	**Efficient clustering of very large document on the bases of time and memory.**	**Efficiency and scalability problem.**
[13]	**Bekkerman, Ron.**	**2004**	**Benchmark experiments, Naive Bayes, SVM, Maximum Entropy**	**Automatic characterization of emails into folders.**	**Accuracy of obtaining folders are relatively low.**
[14]	**Aradhya et al.**	**2014**	**Gabor filter, k-means**	**Detection of text in images and videos using texture features and sharp edges of input images.**	**Discontinuities along the edges or curves in the images are not represented.**
[15]	**Idris et al.**	**2015**	**Negative Selection Algorithm, particle Swarm Optimization**	**Generate the improved email spam detector generation depends upon the bases of negative selection algorithm rather than random generation of detector.**	**Accuracy problem.**
[17]	**Sasaki et al.**	**2005**	**Vector Space Model, K-mean clustering**	**Spam detection using text clustering.**	**Clustering is not bases on the dynamic updating.**

III. CONCLUSION

Emails are used on both the personal as well as professional levels and it can also be considered as official documents in communication between users. Email's data mining and analysis can be conducted for several purposes such as spam detection and classification, subject classification, and so on. In this paper, A Review has shown that the use of unsupervised filtering to filter the input data set is ignored by the most of the existing researchers. The use of hybridization of data mining techniques is ignored in order to improve the accuracy rate further for Detection of fraudulent emails. Most of the existing techniques are limited to some significant features of emails therefore utilising more features may provide more significant results.

REFERENCES

[1] Zaïane, Osmar R. "Introduction to data mining." (1999).
[2] Khorsi, Ahmed. "An overview of content-based spam filtering techniques."*Informatica* 31, no. 3 (2007).
[3] Al-Kadhi, Mishaal Abdullah. "Assessment of the status of spam in the Kingdom of Saudi Arabia." *Journal of King Saud University-Computer and Information Sciences* 23, no. 2 (2011): 45-58.
[4] Günal, Serkan, Semih Ergin, M. Bilginer Gülmezoğlu, and Ö. Nezih Gerek. "On feature extraction for spam e-mail detection." In *Multimedia content representation, classification and security*, pp. 635-642. Springer Berlin Heidelberg, 2006..
[5] Rathi, Megha, and Vikas Pareek. "Spam Mail Detection through Data Mining-A Comparative Performance Analysis." *International Journal of Modern Education and Computer Science* 5, no. 12 (2013): 31.
[6] Kumar, R. Kishore, G. Poonkuzhali, and P. Sudhakar. "Comparative study on email spam classifier using data mining techniques." In *Proceedings of the International MultiConference of Engineers and Computer Scientists*, vol. 1, pp. 14-16. 2012.
[7] Elssied, Nadir Omer Fadl, Othman Ibrahim, and Waheeb Abu-Ulbeh. "AN IMPROVED OF SPAM E-MAIL CLASSIFICATION MECHANISM USING K-MEANS CLUSTERING." *Journal of Theoretical & Applied Information Technology* 60, no. 3 (2014).
[8] Kumar, S., and S. Arumugam. "A Probabilistic Neural Network Based Classification of Spam Mails Using Particle Swarm Optimization Feature Selection." *Middle-East Journal of Scientific Research* 23, no. 5 (2015): 874-879.
[9] Alsmadi, Izzat, and Ikdam Alhami. "Clustering and classification of email contents." *Journal of King Saud University-Computer and Information Sciences* 27, no. 1 (2015): 46-57.
[10] Sharma, Amit Kumar, and Renuka Yadav. "Spam Mails Filtering Using Different Classifiers with Feature Selection and Reduction Technique." In*Communication Systems and Network Technologies (CSNT), 2015 Fifth International Conference on*, pp. 1089-1093. IEEE, 2015.
[11] De, Indrajit, and Jaya Sil. "Entropy based fuzzy classification of images on quality assessment." *Journal of King Saud University-Computer and Information Sciences* 24, no. 2 (2012): 165-173.
[12] Dhillon, Inderjit S., James Fan, and Yuqiang Guan. "Efficient clustering of very large document collections." In *Data mining for scientific and engineering applications*, pp. 357-381. Springer US, 2001.
[13] Bekkerman, Ron. "Automatic categorization of email into folders: Benchmark experiments on Enron and SRI corpora." (2004).
[14] Aradhya, VN Manjunath, and M. S. Pavithra. "A comprehensive of transforms, Gabor filter and k-means clustering for text detection in images and video." *Applied Computing and Informatics* (2014).
[15] Idris, I., and A. Selamat. "A Swarm Negative Selection Algorithm for Email Spam Detection." *J Comput Eng Inf Technol 4* 1 (2015): 2.
[16] Sasaki, Minoru, and Hiroyuki Shinnou. "Spam detection using text clustering." In *Cyberworlds, 2005. International Conference on*, pp. 4-pp. IEEE, 2005.