IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 6, June 2016

Spam Mail Detection using Classification

Parhat Parveen[1], Prof. Gambhir Halse[2]

Student, Department of CSE, KLE Dr. M S Sheshgiri College of Engg & Tech., Belgaum, India[1]

Assoc. Professor, Computer Science &Engg, KLE Dr. M S Sheshgiri College of Engg & Tech., Belgaum, India[2]

Abstract: We are going to explore on how the elaboration of web results in all communication and transactions which are all taking place through web based tool known as email and these days all people are making all their general or business transactions through emails so email acts as an active machine for communication so the user essential time and amount of cash spent on bandwidth will be saved. Spam is an alternative way of an electronic messaging system to send a large number of messages to the user inbox. Here we are going to experiment many data mining techniques to the dataset of spam in an attempt to search the most suitable classifier to email classification as spam and non-spam. Here we are going to check the performance of many classifiers with the use of feature selection algorithm and we found that in the result analysis part the Naïve Bayes classifier provides finer accuracy with respect to other two classifiers such as support vector machine and J48 and we can also see that time taken for Naïve Bayes classifier is lesser than other two classifiers which means that Naïve Bayes classifier is the best classifier among the other two classifier which are used for classifying the spam mails.

Keywords: Classifier, Feature selection, Emails, Spam mails.

I.INTRODUCTION

A. Data Mining

Data mining is described as the technique of fetching the data from very large data sets or it is the art of mining or collecting the knowledge from the data the fetched data can be used for exploration of science and control of production, retention of customer and detection of fraud and for analysis of market .It is combination of fields such as database systems, statics, machine learning and artificial intelligence. The main motto behind data mining is to fetch all related facts from the very big data set and converting it to an intended meaning such that it can be used further. We analyze the data by different perspectives using data mining and is combined or grouped to the helpful information.

Data mining is collection of the information such as business transactions data related to scientific calculations, medical data, details, personal data, satellite sensing digital media, virtual worlds, and text related all information and email messages, which may also contain video or surveillance information and picture related data. Data mining consists of checking of data which has been stored in the data warehouse .The important methods of data mining involve regression, classification and clustering. Data mining is referred as knowledge discovery in database, and methods of data mining includes:

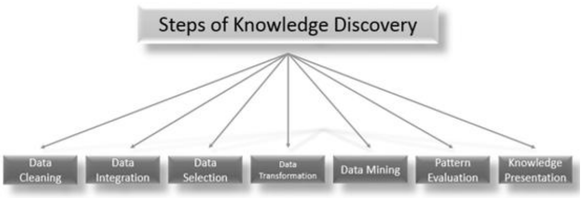

Fig.1.Steps of Knowledge Discovery

- Cleaning of Data: It is the step in which the data which is not relevant and noisy data is discarded by the collection.
- Integration of Data: In this step the data from the multiple sources such as heterogeneous aggregated to a single source.
- Selection of Data: In this process the data which is suitable for the analysis is taken into consideration and extracted from the data collection.
- Transformation of Data: This step is referred as consolidation of data and in which the data is converted to most suitable structure for the mining process.
- Data mining: It is critical process in which intelligent methods are used in an order to fetch patterns which are useful.
- Evaluation in Pattern: In this process the patterns which are interesting they have been based on the given measures identified.
- Representation of knowledge: It is the last step and in which the knowledge which has been discovered illustrated before the user.

B. Machine Learning Techniques:

Machine learning is a study of neural network and it gives the capacity for computers to study and provides brief description of the program being learned and it concentrates on the progress of computer programs that will result itself to originate and modify when applied to the new data.

The machine learning technique is very much similar to the data mining technique. In machine learning without fetching data from large data set here we are going to utilize data to recognize patterns of data and thereby sets working of program appropriately.

II.LITERATURE SURVEY

In 2004 [1]Nie N, Simpser A, Stepanikova I, and Zheng L they proposed that 10 days in a year will get waste in dealing with spam mails and it costs billions of dollars for loss of bandwidth to providing service to suppliers. In 2009 [2] Almeida T, Yama kami A, Almeida J they have been implemented that it is necessary to characterize the mails as spam mails or legitimate mails but the success ratio for the machine learning algorithm is very high for classification. In 1998 [3]Vapnik V N they have been found that as the support vector machine does the classification which in turn optimally distinguishes data into two categories the way by building N dimension hyper plane .SVM is machine learning method used for classification and regression. In 2005 [4] Ian H, Witten and Eibe Frank they have been proposed that the Naïve Bayes method is constructed on the basis of Bayesian concepts.

In 1994 [5]Freitag D and Caruana R.A. they have been implemented that as the feature selection algorithm is regarded as solution obtained from many calculations and it is inspired by fixed set constraints of relevance and the features which are not relevant are not considered for induction and it is not necessary that all features which are relevant are being applied for induction. In 1997 Langley P and Blum A.L they have been proposed the characterization of the feature selection algorithm and characterization would be a search problem in the sense that search organization generation of successor, evaluation measure [6, 7, and 8].

III.PROPOSED WORK

Here spam mails are detected with the help of many classifiers. Firstly many classifiers are applied for the main purpose of spam mail classification and the results are tested based on the accuracy performance related to each classifier.It has been discovered that with Feature Selection algorithm, we can see a remarkable improvement in the classifiers accuracy compared previous results. The classifiers used for the spam mail classification are 1) SVM (Support Vector Machine) 2) Naïve Bayes 3) J48.

IV. SYSTEM DESIGN

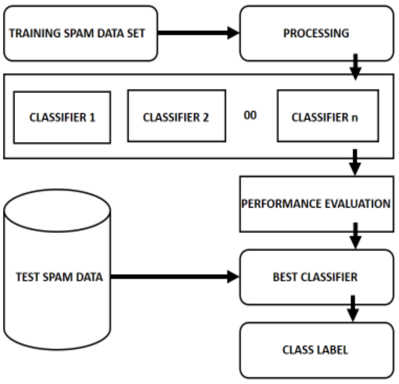

Fig.2.Overall Proposed System Architecture

The above diagram depicts architecture of the proposed system and here firstly in the given Architectural diagram we are going to train the spam data set and later processing has been applied as the real world data comprised of errors it is very important to mine the data in order to get better outcomes from the given data set and the data should in the pre-processed format before using classifier to the data set and is comprised of data cleaning, integration, transformation and it is very much important to normalize the whole data set (normalization is the process where database is structured in systematic way of tables and results should be in an unambiguous format) before using any methods of data mining in order to get better outcomes.

V. IMPLEMENTATION

The classifiers used in the implementation are

- Naïve Bayes
- SVM (Support vector machine)
- J48

A. Naive Bayes:

The benefit of using of Naïve Bayes classifier is that it needs small quantity of training data to evaluate variables which is needed for characterizing. The Bayesian grouping method has a hypothesis, to prove the dataset related to some specific category.

The Naïve Bayes method is constructed on the basis of Bayesian concept as a result it is a fast classifier. Let P (M) is the probability of event M and P (M/N) is the probability of event M which will going to occur where event N has occurred before. The Bayesian Theorem is given by

$P\ (M/N) = P\ (N/M)\ P\ (M)\ /P\ (N)$

B. Support Vector Machine:

SVM algorithm to large extent utilized for email classification. It is easy to implement with the more accurate outcomes. Support vector requires an input data set and produce the prediction output which can be further used to distinguish the data into two separate classes. A SVM does sorting of mails by building a hyper plane which should have N-dimension that sorts the data into two parts and SVM is mostly applied for sorting and association. Two hyper planes are built in order to characterize the mails.

C. J48:

J48 is a type of previous ID3 algorithm and the training data set is experimented with the help of J48 algorithm using WEKA then the outcome will be pictured in decision tree. It is also known as C4.5 algorithm in WEKA. The outcome of this classifier will be non-binary tree with the help of estimate known as gain ratio decision tree to build, the root node will be taken based on the value of information gain, if it is bigger it will be taken into considerations, and the data set is being characterized based on the values of root node. For all sub nodes information gain being evaluated separately and this procedure is recalculated till when all the prediction is done. It deals with decision tree pruning and it is an extended version of ID3 algorithm and I will take into consideration missed attributes values, and attribute values which are continuous .J48 deals with both continuous and discrete valued attributes.J48 also deals with post pruning methodologies such as substitution of sub tree and creation of sub tree.

VI. RESULTS

In the above graph we can see that the accuracy of Naïve Bayes Algorithm is 76%, and for SVM is 62% and for J48 is 54% so Naïve Bayes Classifier provides better result than other two classifiers and we can also see that time taken in Naïve Bayes Classifier is lesser than other two, as a result Naïve Bayes classifier is best classifier for spam mail classification.

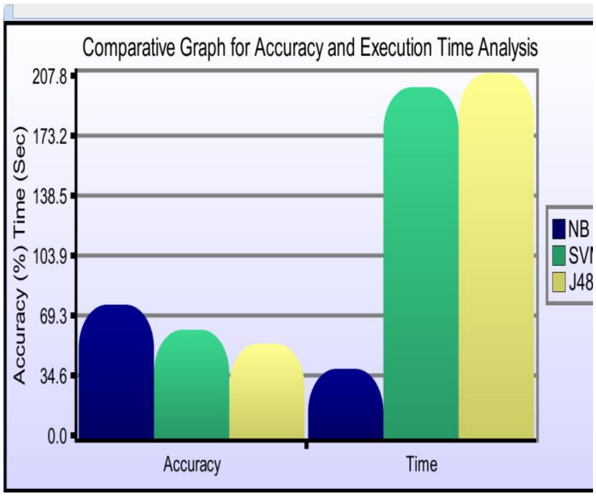

Fig.3.Conclusion and Future Work

VII.CONCLUSION AND FUTURE WORK

Using feature selection the performance analysis in case of Naïve Bayes classifier is highest that is 76% and in case of SVM it is about 62% and in case of J48 performance is about 54%. In the performance analysis graph we can also see that time taken for Naïve Bayes classifier is lesser than SVM and J48, it signifies that Naïve Bayes classifier is the best classifier among the other two classifier.

In the future work using streaming mechanism we can filter the messages as soon as it arrives in the user inbox and thereby eliminates the spam mails, and with the use of high quality filters we can achieve much more accuracy and thereby we can save the user time and cost in such way that without wasting much of network bandwidth. We can also use the classifiers which are much more highly accurate than which we have used in our project to get good performance result in classifying the spam mails.

REFERENCES

[1] "A Study of Spam E-mail classification using Feature Selectionpackage", R. Parimala, Dr. R. Nallaswamy, National Institute of Technology, Global Journal of Computer Science and Technology, Volume 11 Issue 7 Version 1.0 May 2011.

[2] "Comparative Study on Email Spam Classifier using Data Mining Techniques", R. Kishore Kumar, G. Poonkuzhali, P. Sudhakar, Member, IAENG, Proceedings of the International Multiconference of Engineers and Computer Scientists 2012 Vol I, IMECS 2012, March- 14-16, Hong Kong.

[3] "Machine Learning Methods for Spam E-mail Classification", W.A. Awad and S.M. ELseuofi, International Journal of Computer Applications (0975 – 8887) Volume 16– No.1, February 2011.

[4] "Email Spam Filtering using Supervised Machine Learning Techniques", Christina,S.Karpagavalli, G.Suganya, (IJCSE) International Journal on Computer Science and EngineeringVol.02, No. 09, 2010, 3126-3129.

[5] "Email Classification Using Data Reduction Method", Rafiqul Islam and Yang Xiang, Member IEEE, School of Information Technology Deakin University, Burwood 3125, Victoria, Australia.

[6] "Spam Classification based on Supervised Learning using Machine Learning Techniques",Ms.DKarthikaRenuka,Dr.T.Hamsapriya,Mr.M.RajaChakkaravarthi, Ms. P. Lakshmi Surya, 978-1-61284-764-1/11/$26.00 ©2011 IEEE.

[7] "An Empirical Performance Comparison of Machine Learning Methods for Spam E-mail Categorization", Chih-Chin Lai, Ming-Chi Tsai, Proceedings of the Fourth International aConference on Hybrid Intelligent Systems (HIS'04) 0-7695-2291-2/04 $ 20.00 IEEE.

[8] "Feature Subset Selection: A Correlation Based Filter Approach", Hall, M. A., Smith, L. A.,1997, International Conference on Neural Information Processing and Intelligent Information Systems, Springer, p855-858.

[9] "Exploiting Partial Decision Trees for Feature Subset Selection in email Categorization", Helmut Berger, Dieter Merkl, Michael Dittenbach, SAC'06 April 2327, 2006, Dijon, France Copyright 2006 ACM 1595931082/06/0004.

BIOGRAPHIES

Parhat Parveen Buddannavar is a MTech student in the Department of Computer Science & Engineering, KLE DR. M S Sheshgiri College of Engineering Belagavi. She completed her Bachelor of Engineering in Computer science & Engineering from Gogte College of Engineering and Technology Belagavi.

Prof. Gambhir Halse is currently working as Associate Professor in Computer Science &Engg, KLE Dr. M S Sheshgiri college of Engg& Tech, Belgaum. He did his Bachelors of Engg in Computer Science &Engg from Gulbarga university, Gulbarga in the year 1996, did his ME in computer Science &Engg from Shivaji University, Kolhapur in 2005 and pursuing his Ph.D in Computer science &Engg from VTU, Belgaum. His research interests are Data mining and pattern recognition.

DOI 10.17148/IJARCCE.2016.5674