International Journal of Engineering Research & Management Technology ISSN: 2348-4039
November-2015 Volume 2, Issue-6
Email: editor@ijermt.org www.ijermt.org

A STUDY OF DATA MINING TECHNIQUES FOR DETECTION OF MISCLASSIFICATION ERRORS IN MAILS

Satnam Singh
Ph.D Scholar
Computer Engineering
NIILM University
Kaithal, Haryana

Dr. Pankaj Kumar Verma
Professor
Department of CSE
NIILM University
Kaithal, Haryana

1. INTRODUCTION:

As we know that mail means sending a written document to a destination by post. The term e-mail is similar to this concept the difference is only that 'e' stands for electronic. So we can say that it is method of exchanging digital messages from source to destination. Exchange of messages from a single source to one or more destinations all around the world the internet. Email messages can be text files, graphics images and sound files. Email messages are usually encoded in the ASCII text. But now-a-days, the problem in the email is spam and security also. Text editor is included in the email systems to compose the messages. When one send the message to the on specified address then one can also send the same message to the several users and this is called broadcasting.

E-mail is very fast in comparison with the ordinary post and it is easy to use. Emails are fast and language used in emails is simple can be formal or informal. Message through email delivered at once. There is no paper work while using email. It contains friendly environment and can also have pictures, audio files, video files etc .There is also auto responders in email. Products can be advertised, so that companies can reach a lot of people and can advertise their product in a very short time. But having all these advantages emails have some disadvantages too like emails can carry viruses. Unknown and unwanted people can also send messages called spams. Through emails ones systems can get crashed. Mailbox may get flooded with emails after a certain time so one have to empty it from time to time.

1.1 STEPS OF EMAIL:

These are the basic steps of e-mail process:

- Message is sent by email client.
- Email server contacted to the recipients email server.
- Username's validity is checked by the email server.
- If valid username is typed, email is sent to the email server of the address.
- When the recipient sings in to his mailing account, he finds his email.
-

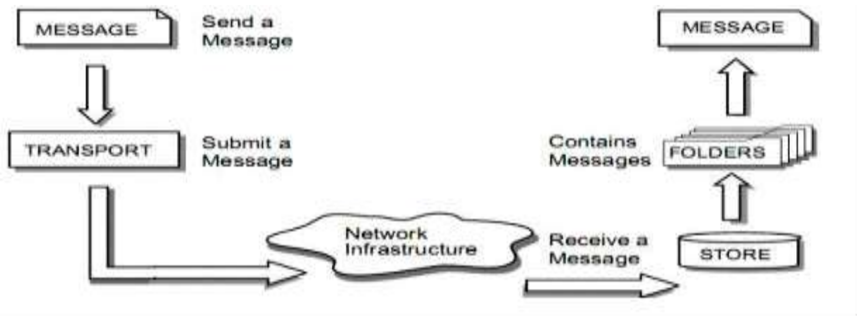

Figure 1: Working of email

1.2 EMAIL FILTERING:

Email filtering [1-2] is the processing of email to systematize it according to the exact criteria. Most often this refers to the automatic processing of incoming messages, but the term is also used to the involvement of human intelligence in addition to anti-spam techniques. Bayesian spam filtering is a statistical method of e-mail filtering. Bayesian spam filtering makes use for Naive Bayes classifier to make out spam e-mail. Work is classified by Bayesian to compare the use of tokens i.e typically words, or we can say irregularly other things, with spam and non-spam e-mails. Bayesian spam filtering is a extremely powerful technique for constricting with spam, that can adapt itself to the email needs of individual users, and gives low false positive spam finding rates that are generally acceptable to users.

1.3 CLASSIFICATION:

Classification is a data mining function that assigns items in a collection to target categories or class. The goal of classification is to accurately predict the target class for each case in the data. For example, a medical researcher wants to analyze breast cancer data to predict which one of three specification treatments a patient should receive. In each of these examples, the data analysis task is classification, where a model or classifier is constructed to predict class (categorical) labels, such as 'safe' or 'risky' for the loan application data; 'yes', 'no' for the marketing data ;or 'treatment A', 'treatment B', 'treatment C' for the medical data. These categories can be represented by discrete values, where the ordering among values has no meaning.

Suppose that the marketing manager wants to predict how much a given customer will spend during a sale at All Electronics. This data analysis task is an example of numeric prediction, where the model constructed predicts a continuous-valued function, or is a stastical methodology that is most often used for numeric prediction; hence the two terms tend to be used synonymously, although other methods for numeric prediction exist. Classification [3] and numeric prediction are the two major types of prediction problems.

Data classification is a two step process, consisting of a learning step (where the classification model is constructed) and the classification step (where the model is used to predict class labels for given data).

➢ FIRST STEP:

In this step, a classifier is built describing a predetermined set of data classes or concepts. This is learning step, where a classification algorithm builds the classifier by analyzing or 'learning form' a training set made up of database tuples and their associated class labels. A class label attribute is discrete-valued and unordered. It is categorical in that each values serves as a category or class. The individual tuples making up the training set are referred to as training tuples and are randomly sampled from the database under analysis. In the context of classification, data tuples can be referred to as samples, examples, instances, data points or objects. The class label of each training tuple is provided; this step is also known as supervised learning (i.e the learning of the classifier is 'supervised' in that it is told to which class each training tuple belongs). It contrasts with unsupervised learning (or clustering), in which the class label of each training tuple is not known, and the number or set of classes to be learned may not be known in advance.

➢ SECOND STEP:

In this step of the classification, the model is used for classification. First, the predictive accuracy of the classifier is estimated. If we were to use the training set to measure the classifier's accuracy, this estimate would likely be optimistic, because the classifier tends to over fit the data. The accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified by classifier. The associated class labels of each test tuple are compared with the learned classifier's class prediction for that tuple. If the accuracy of the classifier is considered acceptable, the classifier can be used to classify future data tuples for which the class label is not known.

1.4 CLASSIFICATION ALGORITHMS:

➢ DECISION TREE:

A decision tree is a flowchart like tree structure, where each internal node (non-leaf node) denotes a test on a attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label. The topmost node in a tree is root node. A typical decision tree represents the concept buys computer, that is, it predicts whether a customer at All Electronics likely to purchase a computer. Internal nodes are denoted by rectangle, and leaf nodes are denoted by ovals.

Some decision tree algorithm produce only binary trees (where each internal node branches to exactly two other nodes) , where others can produce non binary trees[4].

➢ NAIVE BAYES:

The Naive Bayes algorithm is based on conditional probabilities. It uses Bayes Theorem, a formula that calculates a probability by counting the frequency of values and combinations of values in the historical data. Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred.

➢ SUPPORT VECTOR MACHINE:

It is a method of classification of both liner and non-liner data. In nutshell, an SVM is an algorithm that works as follows. It uses nonlinear mapping to transform the original training data into higher dimension. Within this new dimension, it searches for the linear optimal separating hyper plane (i.e., a ‘decision boundary’ separating the tuples of one class from another). With an appropriate nonlinear mapping to a sufficiently high dimension, data from two classes can always be separated by a hyper plane of the classes yield by individual trees (Random Forests is a trademark of Leo Breiman [5] and Adele Cutler for a troupe of choice trees).

2. LITERATURE REVIEW:

Sahami Mehran et al (1998)[1] In addressing the growing problem of junk Email on the internet a method is examined for the automated construction of filters to eliminate such unwanted messages from user’s mail stream. By casting this problem in decision theoretic framework it can be able to make used of probabilistic learning methods in conjunction with a notion of differential misclassification cost to produce filters which are especially appropriate for the nuances of this task. while this may appear at first to be straight forward text classification problem, much more accurate filters can be produced to show that by considering domain specific features of this problem in addition to the raw text of email messages, finally it can be shown that all efficiency of such filters in a real world using scenario arguing that this technology is mature enough for deployment

Konstantinos V. Chandrinos et al. (2000)[6] In the proposed research, to detect the spam Naive Bayesian is trained automatically. This approach is tested on a large collection of personal email messages which are made publically available in ‘encrypted’ from contributing towards standard benchmarks. Appropriate Cost sensitive measures are introduced. In this approach Naive Bayesian filter is compared to see the performance, to filter which is part of widely used email reader. In this approach filtering/routing, text categorization, test collection keywords are used. In conclusion, it concluded after experiment results that cost sensitive evaluation suggests that neither the Naive Bayesian nor the keyword-based filter perform well enough to be used.

Xiaoming JIN et al. (2003)[7] This paper presents an index structure for partially clustered datasets which constitute a large portion of data stored in current information systems. The goal was to make the index respond efficiently to both clustered and uniform data in one database and to perform queries on it without losing precision and recall. This index structure improves the query efficiency in the following ways:

Hovold Johan (2004)[8] in this research, the use of the naive bayes classifier as the basis for personalized spam filters is explored. According to this paper, the several machine learning algorithms are explored already, they were included variants of naive bayes ,but in this proposal the author used word position based attribute vectors, through which very good results are given when they tested on several publically available corpora . These gave many results like , frequent and infrequent words are removed respectively and of course they used mutual information. After proposing this paper, finally an efficient weighting scheme is introduced. So the conclusion of this proposed paper was the complexity of the algorithm and of course the simplicity of the algorithm.

Yang Song et al. (2009)[9] In this paper, Several improvements to the NB classifier have been proposed which is well suited to applications requiring high precision, such as spam filtering. A term-weighting function based on the correlation measure is introduced, which was demonstrated to perform very well on its own and as well as alternative to the typical multiplicative aggregation of several term weighting functions. The problem of feature sparsity is addressed for short documents a class dependent. To expand their attribute vectors CF technique was proposed. Improvement of the classifier performance in the low-FP region is shown. Finally, a novel two-stage NB cascade was introduced. This combines the ability to tackle the potential non-linearity of the decision boundary with an algorithm that jointly optimizes the decision thresholds of the terminal components of the.

M. Basavaraju et al. (2012)[10] In this paper, an email clustering method is proposed and implemented to efficient detect the spam mails. The proposed technique includes the distance between all of the attributes of an email. The proposed technique is implemented using open source technology in C language; ling spam corpus dataset was selected for the experiment. Different performance measures such as the precision, recall, specificity & the accuracy, etc. were observed. K-means clustering algorithm works well for smaller data sets. BIRCH with K-NNC is the best combination as it works better with large data sets. In BIRCH clustering, decisions made without scanning the whole data &BIRCH utilizes local information (each clustering decision is made without scanning all data points). BIRCH is a better clustering algorithm requiring a single scan of the entire data set thus saving time. The work presented in this paper can be further extended & can be tested with different algorithms and varying size of large data sets.

Rachana Mishra el at. (2014)[11] This paper showed classification of spam mail and solving various problem s related to web space. This paper also showed measures parameters which are helpful to reduce the spam or junk mail. Many machine learning algorithm are used to classify the spam and legitimate mail. This paper proposed the best classifier and better classification approach using different data mining tools using bench mark data set. The data set consists of 9324 records and 500 attributes used for training and testing to build the model. In this paper a procedure that can help eliminate unsolicited commercial e-mail,viruses,torjans and worms as well as frauds perpetrated electronically and other undesired and troublesome e-mail. This paper showed analyzing of different supervised classifiers technique using different data mining tools such as weka, rapid miner and support vector machine[10][11][21].

3. PROBLEM FORMULATION:

As we know that emails are easy to use. Emails are fast and language used in emails is simple can be formal or informal. Message through email delivered at once. There is no paper work while using email. It contains friendly environment and can also have pictures, audio files, video files etc .There is also auto responders in email. Products can be advertised, so that companies can reach a lot of people and can advertise their product in a very short time. But having all these advantages emails have some disadvantages too like emails can carry viruses. Unknown and unwanted people can also send messages called spams. Through emails ones systems can get crashed. Mailbox may get flooded with emails after a certain time so one has to empty it from time to time.

4. PROPOSED WORK:

My Proposed research is for the less error prone classification by reducing the misclassification. Misclassification is defined as when legitimate emails are categorized as junk emails or vice versa. Cost of misclassifying legitimate emails as junk is much higher than the cost of junk mails as legitimate mails.

Remedies can be found by Classification schemes which will save our time and data, by categorizing between spam and non-spam.

In case of Linear Discriminant Analysis, there are training data and sample data. The observations with known class labels are known as training data. There are sample data on which we will be using the training data sets. Then we will be computing the reconstitution error which is the misclassification error (the proportion of misclassified observations) on the training set. We will also compute the confusion matrix on the training set. A confusion matrix contains information about known class labels and predicted class labels. Generally speaking, the (i,j) element in the confusion matrix is the number of samples whose known class label is class i and whose predicted class is j. The diagonal elements which would be represented in graph will correctly classified observations. For some data sets, the regions for the various classes are not well separated by lines. When that is the case, linear Discriminant analysis is not appropriate. Instead, you can try quadratic Discriminant analysis (QDA) for our data. Decision trees can handle both categorical and numerical data. For the decision tree algorithm, the cross-validation error estimate is significantly larger than the reconstitution error. This shows that the generated tree over fits the training set. In other words, this is a tree that classifies the original training set well, but the structure of the tree is sensitive to this particular training set so that its performance on new data is likely to degrade. It is often possible to find a simpler tree that performs better than a more complex tree on new data.

5. OBJECTIVE:

1. Accessing and categorizing the UCI repository on email filtering
2. Implement Linear and Quadratic Discriminant analysis.
3. Implement Naive Bayes Algorithm
4. Implement decision tree algorithm
5. Finding out the misclassification error

6. RESEARCH METHODOLOGY:

Previously, spam classification is done on different classification algorithm and it was found that Random Forest algorithm is best suitable for the same. But there are some disadvantages of Random Forest algorithm [37][40]. These are:

i. Large number of trees may make the algorithm slow for real-time prediction.
ii. It is not suitable for less number of dataset due to longer execution time.
iii. Hard to understand

As our dataset is already filtered, we will not need to create large number of trees. So from the angle of dataset, decision tree best suits our research. It has the following advantages:

i. Easy to interpret and explain
ii. Lesser execution time over random forest

6.1 METHODOLOGY:

As the base of our research is to use parallel algorithms, so we will be implementing the linear and quadratic Discriminant analysis, Naïve Bay's algorithm and decision tree, so we will be implementing the standard decision tree algorithm.

- Classification using linear distribution
- Classification plotted using Quadratic Distribution
- Classification using Naive Bayes Gaussian distribution
- Classification using Naive Bayes Kernel distribution
- Classification using decision tree classifier

6.2 TOOLS USED FOR CALCULATION AND VISUALIZATION:

➢ MATLAB (12.0):

MATLAB is widely used in all areas of applied mathematics, in education and research at universities, and in the industry. MATLAB stands for Matrix Laboratory and the software is built up around vectors and matrices. This makes the software particularly useful for linear algebra but MATLAB is also a great tool for solving algebraic and differential equations and for numerical integration. MATLAB has powerful graphic tools and can produce nice pictures in both 2D and 3D. It is also a programming language, and is one of the easiest programming languages for writing mathematical programs. MATLAB also has some tool boxes useful for signal processing, image processing, optimization, etc.

➢ WEKA:

Weka (pronounced to rhyme with Mecca) is a workbench that contains a collection of visualization tools and algorithms for data analysis and predictive, together with graphical user interfaces for easy access to this functionality. The original non-Java version of Weka was aTCL/TK front-end to (mostly third-party) modeling algorithms implemented in other programming languages, plus data preprocessing utilities in C, and a Make file-based system for running machine learning experiments. This original version was primarily designed as a tool for analyzing data from agricultural domains, but the more recent fully Java-based version (Weka 3), for which development started in 1997, is now used in many different application areas, in particular for educational purposes and research.

7. CONCLUSION:

This research is to classify the filtered data. The main purpose of this research is to reduce the error rate of the data and to improve the accuracy. In the previous techniques of classification there may be some misclassification. But in this research the problem of misclassification is reduced. The work presented by this research is the classification techniques. Therefore, it’s a good enterprise solution for filtering. This will optimize the system performance and make some improvements on the previous algorithm. This will give the better results from the previous one.

In this paper the filtered mails are further filtered to measure the misclassification using different data mining techniques. This paper shows that the decision tree is the best classifier. It is easy to interpret and explain the executives. In comparison to random forests are time efficient. Decision tree requires relatively less effort from users for data preparation. For proper visualization and calculation, weka tool and MATLAB has been used.

Decision trees can handle both categorical and numerical data. For the decision tree algorithm, the cross-validation error estimate is significantly larger than the resubstitution error. This shows that the generated tree over fits the training set. In other words, this is a tree that classifies the original training set well, but the structure of the tree is sensitive to this particular training set so that its performance on new data is likely to degrade. It is often possible to find a simpler tree that performs better than a more complex tree on new data.

REFERENCE:

1. Sahami, M., Dumais, S., Heckerman, D., & Horvitz, E. “A Bayesian approach to filtering junk e-mail. In Learning for Text Categorization:” the 1998 workshop (Vol. 62, pp. 98-105).
2. Androutsopoulos, I., Paliouras, G., Karkaletsis, V., Sakkis, G., Spyropoulos, C. D., & Stamatopoulos, P. “Learning to filter spam e-mail: A comparison of a naive bayesian and a memory-based approach.arXiv preprint cs/0009009”(2000).
3. Jason D.M.Rennie “Practical Concerns Surrounding The Application Of Text Classification To The Problem Of Mail Filtering”, International Journal of Computer Applications (2000).
4. Ian H. Witten, Eibe Frank, “Data Mining – Practical Mahine Learning Tools and Techniques” 2nd Edition, Elsevier, (2005).
5. Han, J., Kamber, M., & Pei, J.“Data mining: concepts and techniques.” Morgan Kaufmann (2000).
6. Konstantious V. Chandrinos, Constantine D.spyropoulos “To detect the spam Naïve Bayesian is trained automatically” (2000).
7. Jin, X., Wang, L., Lu, Y., & Shi, C. “MC-tree: Dynamic index structure for partially clustered multi-dimensional database” Tsinghua Science and Technology, 8(2), 174-180”.

8. Hovold, Johan. "Naive bayes spam filtering using word-position-based attributes" In Proceedings of the 2nd Conference on Email and Anti-Spam (CEAS July-2004)."
9. Song, Y, Kołcz, A, & Giles, C. L. "Better Naive Bayes classification for high precision spam detection. Software: Practice and Experience", 39(11), 1003-1024 (2009)".
10. Basavaraju, M., & Prabhakar, R. (2012). "A novel method of spam mail detection using text based clustering approach" International Journal of Computer Applications, 5(4).
11. Rachna mishra,Ramjeevan Singh Thakur "An efficient approach for supervised learning algorithms using different data mining tools for spam categorization" Journal IEEE 2014.
12. Jehad Ali,Rehanullah khan,Nasir Ahmad,Imran Maqsood(Sep 2012)".Random Forests and Decision trees .Journal IJCSI".
13. Prajwala T R(2015)".A comparative study on decision tree and random forest using R tool.International journal of advanced research in computer and communication engineering.(vol.4 196-1).