"Email Classification Using Enhanced AdaBoost Algorithm"

A Dissertation Proposal
submitted

By

Shobhika Rani

to

Department of Computer science and engineering

In partial fulfilment of the Requirement for the

Award of the Degree of

Master of Technology in Information Technology

Under the guidance of

Balwinder Kaur

(April 2015)

ABSTRACT

Email is one of the fastest, easiest and common forms of communication on the internet today. In the growing era of internet, there is rapid increase in the number of email users which increases Spam Emails in recent few years. Spam is unwanted mails which are sent in huge amount to anyone anywhere and are of no use to the recipient. Text classification using email classification presents various challenges because of the large number of the documents. The different mail classification algorithms are Naive Base, SVM, Neural Network, J48, and Adaptive Boosting Algorithm. Boosting is the machine learning method to improve the performance of any learning algorithm based on the concept of creating a high accurate prediction rule by combining various weak classifiers and non appropriate rules. It was first presented by Schapire and Freud. Further research on boosting techniques introduced a new boosting algorithm called AdaBoost Algorithm. In my work, I would replace the weak learner of AdaBoost with hybrid classifier that contains AdaBoost algorithm and are hybridized on the basis of average of their probabilities. And add more decision making conditions while calculating the class for model prediction i.e. on the basis of error rate adds more weight to class that will give better class for the prediction.

ACKNOWLEDGEMENT

It is with deep sense of reverence that I express my sincere thanks to my supervisor **Mrs. Balwinder Kaur** for their guidance, encouragement, help and useful suggestions. Their untiring efforts, methodical approach and individual help made it possible for me to complete this work in time.

I am also thankful to all my friends for their continuous motivation and help. Although it is not possible to name individual, I cannot forget my well wishers for their persistent support and cooperation.

(Shobhika)

DECLARATION

I hereby declare that the dissertation proposal entitled, Email Classification Using Enhanced AdaBoost Algorithm, submitted for the M.Tech Degree is entirely my original work and all ideas and references have been duly acknowledged. It does not contain any work for the award of any other degree or diploma.

Date: **Investigator**

Regn. No. 11300477

Certificate

This is to certify that Shobhika rani has completed MTech Dissertation titled Email classification using enhanced AdaBoost algorithm under my guidance and supervision. To the best of my knowledge, the present work is the result of her original investigation and study. No part of the dissertation has ever been submitted for any other degree or diploma.

The dissertation proposal is fit for the submission and the partial fulfilment of the conditions for the award of M.Tech Computer Science & Engg.

Date:

Signature of Advisor

Name:
UID:

Table of contents

Title	i
Abstract	ii
Acknowledgement	iii
Declaration	iv
Certificate	v
Table of contents	vi
List of figures	vii
CHAPTER 1 INTRODUCTION	1
1.1 Classification	4
1.2 Machine Learning	5
CHAPTER 2 LITERATURE REVIEW	10
CHAPTER 3 PRESENT WORK	15
3.1 Problem formulation	17
3.3 Scope of work	20
3.4 Objectives	21
3.5 Research methodology	22
CHAPTER 4 RESULTS AND DISCUSSIONS	23
CHAPTER 5 CONCLUSION AND FUTURE SCOPE	32
CHAPTER 6 REFERENCES	33
Websites	34

List of Figures

1.1 Data mining process 3

1.2 Data mining tasks 3

1.3 Machine learning technique 7

1.4 Learning process 7

1.5 Prediction process 8

2.1 Boosting representation 10

3.1 Adaboost Classifiers 16

3.2 Flowchart for the proposed scheme 17

3.3 Steps of research design 20

3.4 Flowchart of enhanced Adaboost Algorithm 22

Chapter1
Introduction

Data mining is referred to the process of gaining or extracting the knowledge which is relevant from various large and operational databases. The term data mining is also called the "Knowledge Discovery Process". Software of data mining is one of the major numbers of analytical tools for analyzing the business organizational data. It allows the users to analyze data from many different perspectives, classify it, and summarize the relationships found. Data mining uses information and facts from the past data for analyzing the results of a particular problem that may occur. Data mining aims to analyzing the data stored in data warehouses which is used to store the businesses data that is stored and gathered for being analyzed. The data in data warehouses can come from various parts of the business forms, industries and from the production to the management environment too. Managers in the organization use data mining for deciding the marketing strategies for their product analysis. They can use this data analysis for comparing among their competitors. Data mining interprets all the data into real time analysis framework which can be used to increase their sales, promoting any new product in market, or deleting any product that is not of any value to the company or an organization.

This field has its various application areas such as Financial and Banking Services, Consumer Products and utilities, Retail and marketing, Medical and Healthcare, Education Social Networking & Social Media. . Data mining is applicable on various kinds of data repositories such as data warehouses, relational databases, transactional databases, data streams, flat files and World Wide Web. Data mining is an essential step in the process of discovery of relevant knowledge.

The process of knowledge discovery or knowledge extraction is an iterative process and it contains the following steps:

1. Data cleaning involves cleaning of noisy data. It removes the noise and inconsistent, irrelevant data from databases.
2. Data integration where data from different multiple data sources are combined together and collected in one data store.

3. Data selection where the data which is relevant to the task under analysis are selected and retrieved from the database.
4. Data transformation where the data are transformed into the forms appropriate for the mining process by performing the summary or aggregation operations.
5. Data mining, the process where methods are applied to mine or extract important data patterns.
6. Pattern evaluation considers identifying the interesting patterns which represent knowledge based on the interestingness measures.
7. Knowledge presentation where the knowledge representation methods are used to present the mined and extracted knowledge to the database user.

Steps 1 to 4 are the steps which are used for pre-processing the data, where the data is processed prior to the mining so that and inconsistency, irrelevant or noisy data is removed from the database. This pre-processed data is passed to the data mining algorithms and techniques which produces an output in some forms of patterns. Data mining step interact with the user or a knowledge base. The patterns which are interesting and true are presented to the database user and can be stored as the new knowledge in the knowledge base. Data mining is the essential and most important step in knowledge discovery process because it mines the hidden patterns from the database which is important for the data evaluation and various data analysis tasks.

The main features of data mining are as follows:

- The Automatic discovery of the data patterns
- The Prediction of various expected outcomes
- The Creation of relevant information
- Concentration on large sets of databases

The process of data mining:

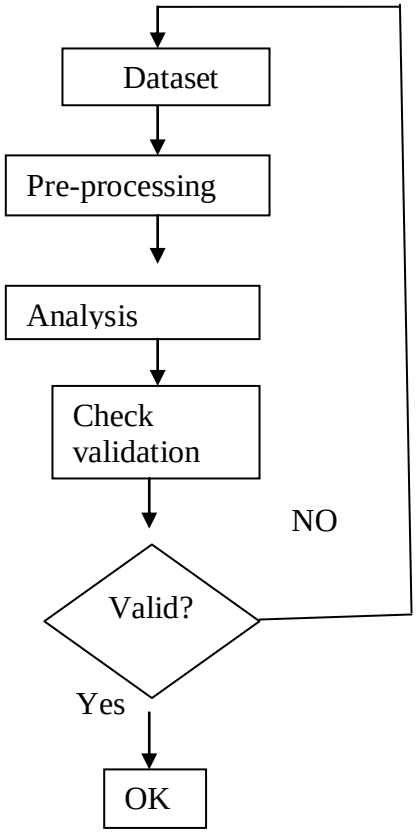


Fig. 1.1 Data Mining Process

Firstly, the input data is fed into to process , then data is pre processed to find any missing values , then the data is being analysed for the validation and is the data is valid then OK this data can be used for model generation otherwise the data is fed into the initial input phase again.

Data mining tasks:

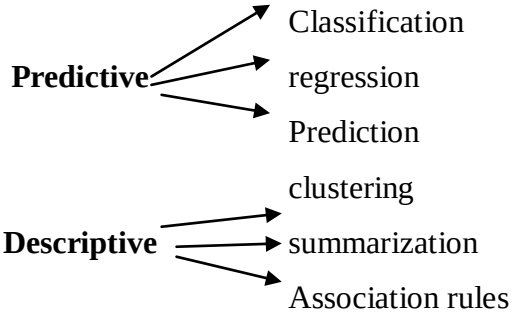


Fig. 1.2 Data mining tasks

Data mining applications are divided into four categories: Numerical prediction, association rules, clustering and classification.

Numerical Prediction

Numerical prediction is the form of supervised learning where we can predict the numerical values such as a company's shares or profit values. Numerical prediction is also known as regression. One of the popular way to numerical prediction is Neural Networks which is a complex technique based on human neurons. A neural network is being given different set of inputs and then it is used to predict different outputs.

Association Rules

Association rules are the form of unsupervised learning. Association rules are used when we want to use a training dataset to find the relationship that may exist between the values of variables in the form of rules. The most common type of application is "Market Basket Analysis". It analyse the purchasing behaviour of the market customers that help company to predict its future sales

Clustering

Clustering refers to grouping the datasets or objects which show similar behaviour or which have similar characteristics are in one cluster and objects showing different characteristics are grouped in different cluster.

Classification

Classification is one of the most important applications of data mining. It considers the tasks which occur in our everyday life. For example, we can classify students report as merit, pass or fail; a hospital may want to classify its patients into high, low or medium risk of having certain kind of illness and classify an email that whether it is legitimate or non legitimate email.

1.1 Email Classification:-

Today, Email has become the most common and an important part of our everyday's communication over the internet. This is found to be a quick and very economically good way to exchanging the information with each other. But, the users are annoyed with the unwanted emails they receive in their folders. These unwanted emails which are of no use to the recipient are called SPAM. These malicious spams consume the time, the network bandwidth and space area of the users. Studies say that approximately 95% of the emails received are Spam. To prevent these spam to enter the user email folders various filtration techniques are needed. Some of the filtration methods used is for email header part and others used are for body part of the email such as the filtering which is based upon the machine learning. There are various approaches for classifying email into spam or non spam. Experiment is being done to find the results of the spam filtering. I find Adaptive boosting i.e. AdaBoost algorithm an effective approach to solve the email spam problem.

Boosting is the machine learning method for improving the performance of any learning algorithm on the idea of creating a high accurate prediction rule by combining various weak classifiers and non appropriate rules. It was first presented by Schapire and Freud. Further research on boosting techniques introduced a new boosting algorithm called AdaBoost Algorithm.

In our research work, we propose Adaboost algorithm for the email filtering which is a machine learning algorithm with the focus on combining the weak classifiers and then demonstrating that this algorithm can be applied for improving the performance of spam filtering.

1.2 Machine Learning :-

Machine learning is an artificial intelligence branch in which a computer program learns from the experience and the performance if its performance measure improves with the experience. For e.g., we can train a machine learning system to distinguish between the spam and non spam messages in emails. After the learning process, we can use it then for classifying the emails received into different spam and non spam email folders.

Why machine learning is important-

1. Some tasks can be defined well only by examples.
2. Machine learning helps us to find the hidden correlations and relationships from the large amount of data.
3. Gives better results for prediction and model generation.
4. Environments may change sometimes.
5. Some problems with huge amount of knowledge too hard for humans to be described.

Example of designing a learning system:

- Description of the problem
- Selection of the training experience
- Selection of target function
- Selecting a representation algorithm
- Finally the design.

Machine learning algorithms types:

- Supervised learning
- Unsupervised learning
- Semi supervised learning

Our problem area comes under the unsupervised learning where the input data is fed when the output is unknown to us and our objective is to find the structure in the data. For example in cluster analysis, email classification comes under the unsupervised learning.

Machine learning technique:

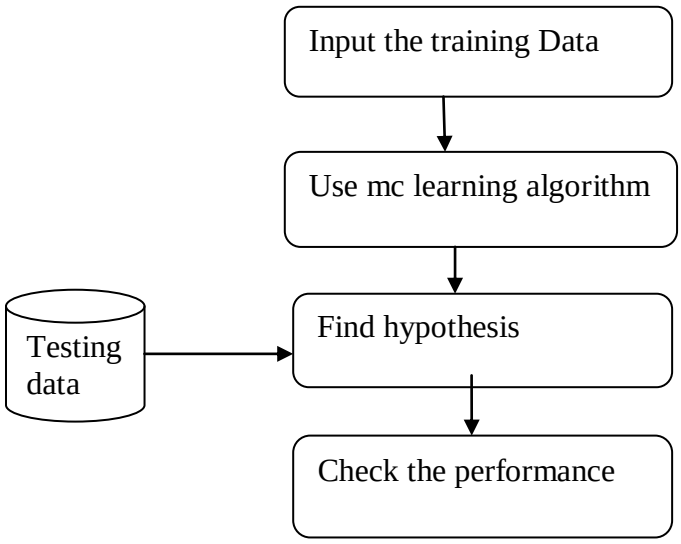


Fig. 1.3 Machine learning technique

Machine learning process:

1. Learning :

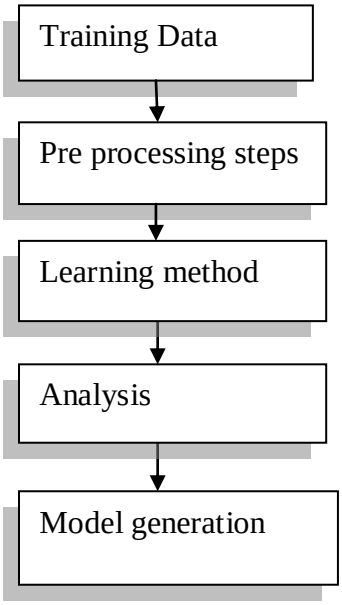


Fig. 1.4 learning process

In learning process, we have input training data, then pre processing steps are done , a learning method is applied for machine learning system, analysis of the system behaviour is done for the process of model generation.

2. Prediction:

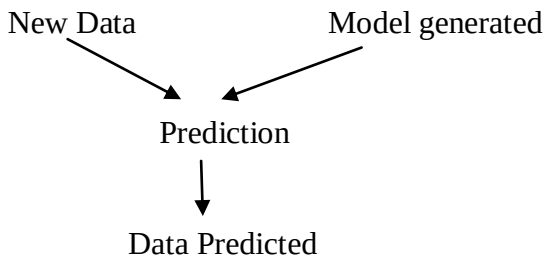


Fig. 1.5 Prediction process

In prediction, the new data and the model generated will be used for prediction behaviour and the outcome shall be the predicted data.

Adaboost algorithm is very important part of classification in machine learning algorithm. The algorithm takes the input a training dataset and numbers of iterations are performed on this dataset for classification purpose. Adaboost, is a method to improve the accuracy of any given machine learning algorithm, and used to solve the problem of object detection. The algorithm uses a method for updating the weights of base classifiers. The weights are changed by applying the different error rates.

Adaboost terms:

- h(x) – base learner
- H(x) – strong learner
- Weak learner
- Strong learner or classifier – a linear combination of various weak classifiers.

AdaBoost Features:

1. Programming of AdaBoost is easy and it gives better and quick results.
2. AdaBoost Works fine with other different machine learning algorithms.
3. AdaBoost works well with large number of training datasets.
4. The Weak Learners cannot be too complex or too simple.

Chapter 2

REVIEW OF LITERATURE

Sarwat Nizamani, Nasrullah Memon, Uffe Kock Will

This paper explains about the experiments used to detect the illegitimate or spam emails using boosting algorithms. An email is illegitimate if it is unwanted to the recipient. The email detection is divided into two parts: suspicious mail detection and spam mail detection. Boosting algorithms can enhance the accuracy of email classification algorithms. Different classification algorithms are applied as base learners. The classification algorithms are decision tree, naive bayes, SVM. Then AdaBoost Algorithm is experimented using the all above algorithms as weak learners of the boosting algorithms.

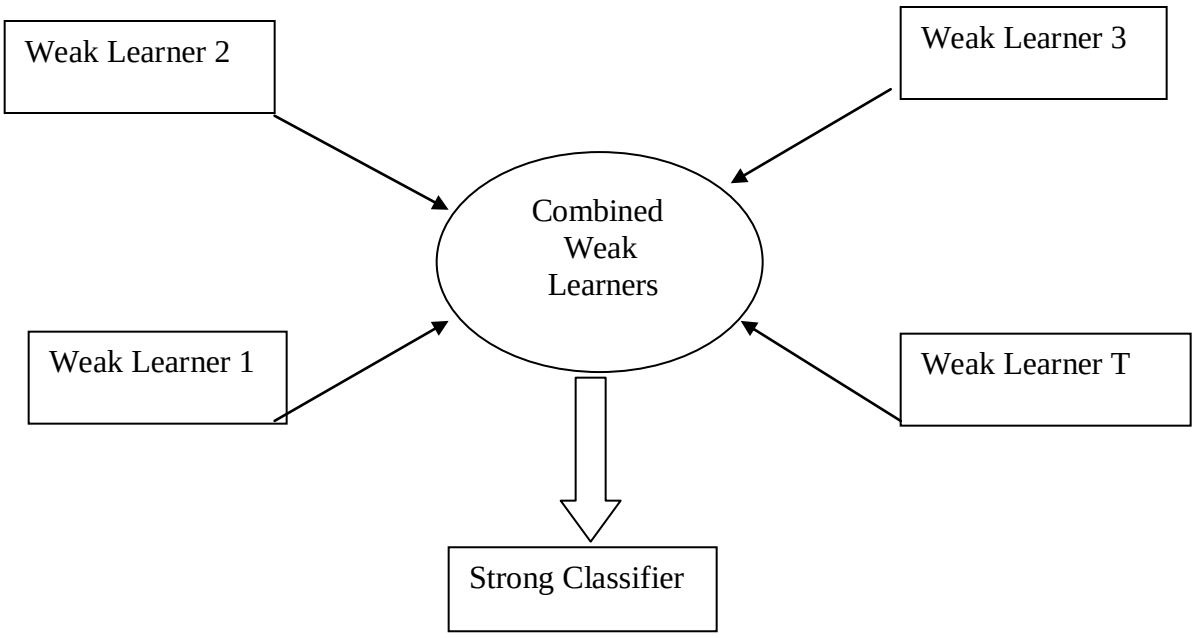


Fig. 2.1 Boosting Representation

Liao Shaowen, Chen Yong

AdaBoost algorithm is very important algorithm in the classification and machine learning algorithm, if there exist more n difficult instances in the training samples, and increasing number of the iterations performed then this algorithm leads to the degeneration. AdaBoost can be used in the problem of the face detection and many other applications. The paper defines the LWE-AdaBoost algorithm which helps in limiting the expanding weight sizes in the algorithm and also lessens the degeneration.

Qiujie Li, Yaobin Mao

Boosting is a technique that is well used in machine learning algorithms. The issue rises with the imbalanced data classification is that imbalanced data issues a problem in the performance of various machine learning algorithms which lead to miss classification cost. This paper proposes the different boosting techniques for the data imbalanced such as IDBoosting. Imbalanced data classification had its various applications in fraud detection, medical diagnosis, intrusion detection.

Nikunj C. Oza

This paper tells about the Adaboost importance in ensemble machine learning algorithms. The algorithm construct a chain of the weak models where each model is made based on the performance of the last model on the input dataset. After the model is constructed, Adaboost the model is being tested over the training data to check if it is well learned. The paper assumes that the base learner is the weak learner. The weights of correctly classified instances are reduced and that of incorrectly classified samples is increased. The next model is generated with this newly comes out weight and the training samples.

Seongwook Youn and Dennis McLeod

Today In the growing internet word, the email is an important part of our daily communication. But the increasing size of these emails has posed a threat over the user security and user gets unwanted spam messages in their folders. This paper reviews the email data classification using four types of different classifiers namely Neural Network, SVM classifier, Naïve Bayesian Classifier, and J48 classifier. The experiments were performed based on the different data size . The outcome of the final classification should be 1 if it is found to be spam and 0 if it is non spam. The paper depicts that the simple J48 classifier, could be efficient for the dataset which could be classified as binary tree.

Charu C. Agarwal, ChengXiang Zhai

This paper tells about the survey of different text classification algorithms being used in data mining. Classification is an important part of the data mining and includes different fields in classification. The text mining has its many applications in which some of them are news filtering, document organization, opinion mining, and email classification and filtering of the email spams. Some of the text classification algorithms used are decision tree, rule based algorithms, neural networks, Bayesian classifiers and other classifiers. These text mining methods should be applied for managing the large number of elements whose frequency varies.

Mayank Pandey, Vadlamani Ravi

The millions of user across the world access the internet daily for communication and business Purposes in home or organizations. Phishing attack comes out to be a dangerous threat for financial as well as non financial organization. The existing email filtering techniques now a days tend to be ineffective to control the dangerous phishing attack. Hence, the paper defines a method using text and data mining techniques for the prediction of the attacks of phishing correctly.

Spam emails are unwanted email or bulk email, which contains link of phishing websites and sent to the mass of people who don’t need it or request it. These spams consume users’ precious time, disk space and cause frustration thereby reducing the bandwidth. According to a latest

study it is found that mostly 72% of the mails received are spam mails and the cost caused to the business companies due to these malicious mails cause approx. 1 billion dollar a year and this rate is still growing. So, these spam emails in folders need to be identified accurately and properly by the system.

Ali and Yang Xiang

This paper reviews the current approaches for blocking the spam and proposes a new spam classification method by using adaptive boosting algorithm. The Experiments are carried out to find the results of spam filtering. This paper finds that adaptive boosting algorithm is an effective approach to solve the email spam problem. The default method in WEKA tool such as Decision Stump is not the best associated algorithm to filter the email spams. After comparing Decision Stump, J48, and Naïve Bayes it concludes that J48 is the most suitable associated algorithm to filter the email spam with high true positive rate, low false positive rate and low computation time.

Boosting is a popular established method in the machine learning group for improving the performance of any learning algorithm. The boosting algorithms were first presented by Schapire and Freund. After their further research they introduced a new generation of boosting algorithm called Adaptive Boosting (AdaBoost) algorithm. AdaBoost.M1 and AdaBoost.M2 were then developed by Schapire and Freund from their AdaBoost algorithm. For the purpose of the binary classification problems, the two versions are equivalent, they are different only in the way they handle problems with more than two classes. AdaBoost.M1 has access to a learning algorithm (which is called as Weak Learner) which it calls repeatedly with probability distributions over the training set. Weal Learner calculates the hypothesis that tries to correctly classify all samples of the test data. Examples that are incorrectly classified are given greater weighting for the next pass. Finally, the adaptive boosting algorithm then combines all the hypotheses into one final hypothesis i.e. all weak learners are combined to form a strong classifier.

Peng Wu and Hui Zhao

This paper considers the introduction of AdaBoost Algorithm, an analysis and various aspects of this boosting algorithm. AdaBoost was first developed by Schapire and Freund in 1990. The AdaBoost Algorithm is self adaptive boosting technique which enhances the performance of the base learners or weak learners by creating the set of multiple classifiers. Specific learning algorithm is used to weak classifier and then calculated its error rate on different training sets. AdaBoost uses this error rate to find the distribution of training sets. A Greater weight is set to the incorrectly classified sample and weight is reduced if the sample is correctly classified. Finally a strong classifier is established by combining the different weights of the training sets. AdaBoost is an iterative algorithm which upgrades the weak classifiers at every iteration.

Chapter 3

PRESENT WORK

3.1 Problem formulation:

AdaBoost stands for adaptive boosting which is a boosting algorithm comes under the machine learning techniques. AdaBoost was developed by Fraud and Robert Schapire. The Adaboost has its benefit that it can be used and combined with any other machine learning algorithm to improve its performance. The algorithm is found to be sensitive in the presence of the noisy data and any outliers. It is adaptive in the sense that that algorithm learns from the performance of each subsequent weak classifier at each iteration being performed.

The inputs to the algorithm are:

1. Training dataset or instances as input.
2. The distribution on the all training instances.
3. A weak learner or base classifier to be used.
4. The number of iterations to be performed.

Boosting is the machine learning method for improving the performance of any learning algorithm on the idea of creating a high accurate prediction rule by combining various weak classifiers and non appropriate rules. It was first presented by Robert E. Schapire and Freud. Further research on boosting techniques introduced a new boosting algorithm called AdaBoost Algorithm.

A strong classifier is built with the combination of various base classifiers or learners and the final classification is done based on the votes of all the weak classifiers. The weight of the correctly classified samples is reduced and that of incorrectly classified instances should be increased to improve the performance in the model generation. The algorithm performs better and gives good results than that of random guessing. The algorithm yields better results when

performed with more number of iterations. However, the number of iterations to be performed is not pre fixed so it is decided as the algorithm steps grow.

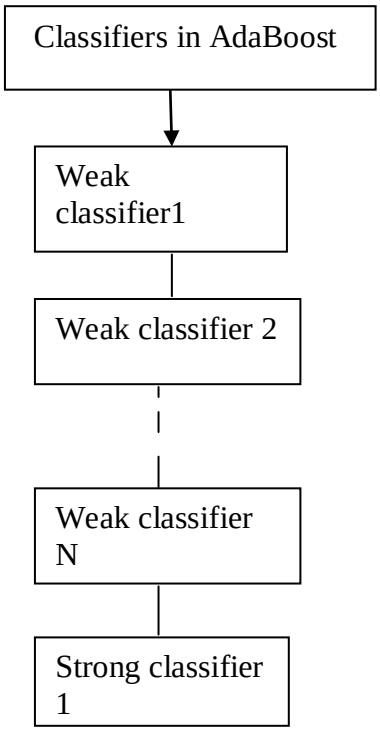


Fig. 3.1 Adaboost Classifiers

In our problem definition, the decision stump i.e. the base learner is replaced by one hybrid classifier that is combination of adaboost and naïve bayes and are hybridized based upon the average of their probabilities. The decision making conditions are applied when we calculate the class for model prediction. The input is training dataset which is collected with different emails which are spam and non spam and saved in .arff file. The input is first pre processed to find any missing value and for attribute selection using a filter and then the pre processed data is used for the classification purpose. We can set the conditions in filtration for numeric attributes, alphabetic attributes and same for lowercase letters or uppercase letters. The iterations are performed over the training data for the classification and at every step a class is being categorised as spam and non spam and weight is calculated for each class.

3.2 Flowchart for the proposed problem:

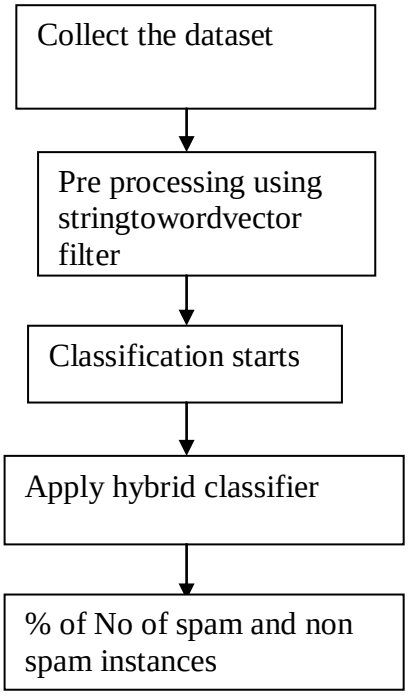


Fig. 3.2 Flowchart for the proposed scheme

After the classification, in enhanced adaboost algorithm, more number of correctly emails will be come out as compared to the previous algorithm so the percentage and efficiency of enhanced adaboost algorithm which gives less number of spam classes.

3.3 Scope of the work:

As the organisational databases size is increasing and the volume of the data being stored in databases is increasing so there is need of some techniques to summarize these data, identify true interesting trends and patterns from these databases, and act upon the outputs. Data mining helps in uncovering the hidden patterns of the data. Outputs found through the data mining process are economical and crucial to the business organisations who aim at competitive advantages in the market. Data mining rely on the system databases to take out input data and the data stored in repositories can be noisy, inconsistent and irrelevant. Data mining has its various applications in which email classification is important. Email is primary the common method of communication today over the internet and emails can be spam or non spam. Spam emails are sent to the recipients in bulk and are unwanted to receiver. These types of spams are very serious and illegal to the recipients. There are different classification algorithms of email and I found AdaBoost algorithm is an effective approach to solve the email spam problem. It comes under the boosting techniques. Adaboost is fast, easy and used in many areas and it requires no prior knowledge about the weak classifier.

3.4 Objectives:

1. The decision stump i.e. the base learner is replaced by one hybrid classifier that is combination of adaboost and naïve bayes and are hybridized based upon the average of their probabilities.

2. The decision making conditions are applied when we calculate the class for model prediction.

3.5 Research methodology:

Steps defined for the design are:

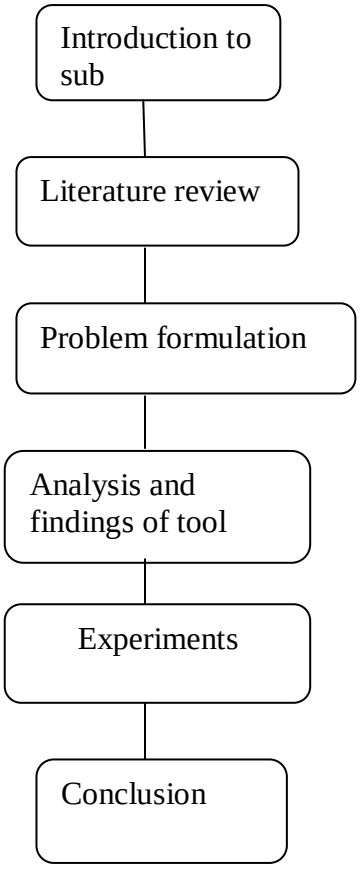


Fig. 3.3 Steps of research design

The introduction to subject is defined, then the review of literature is done thoroughly to study about the problem. Analysis is done for the research work and present work being done in problem formulation and tool is decided for the implementation work. Experiments are gathered through the implementation and finally conclusion is get for the defined research problem.

Methodology for the present work:

Model generation

1) Assign equal weight to each training instance initially.
2) For *t* iterations:

   Apply learning algorithm to the weighted dataset, and store the result
3) Compute model's error rate *e* on the weighted dataset
4) If *e* = 0 or *e* >= 0.5

   Then we terminate the model generation
5) For each training instance in dataset:

   e=0 means sample is correctly classified and e=1 means sample is incorrectly classified. In the next round of iteration in algorithm, adaboost pay more attention to the incorrectly classified sample for meeting the idea of upgrading weights.
6) If sample is classified correctly by the model then:

   Multiply training instance's weight by *e*/(1-*e*)
7) Normalize the weight of all instances.

For Classification

1) Assign weight = 0 to all classes
2) For each of the *t* (or less) models:

   For the class to which this model predicts

   Add –log *e*/ (1-*e*) to this class's weight
3) Return class with the highest weight.

The implementation part is done in eclipse and the interfacing of the code is performed through java netbeans.

Flowchart for enhanced adaboost algorithm:

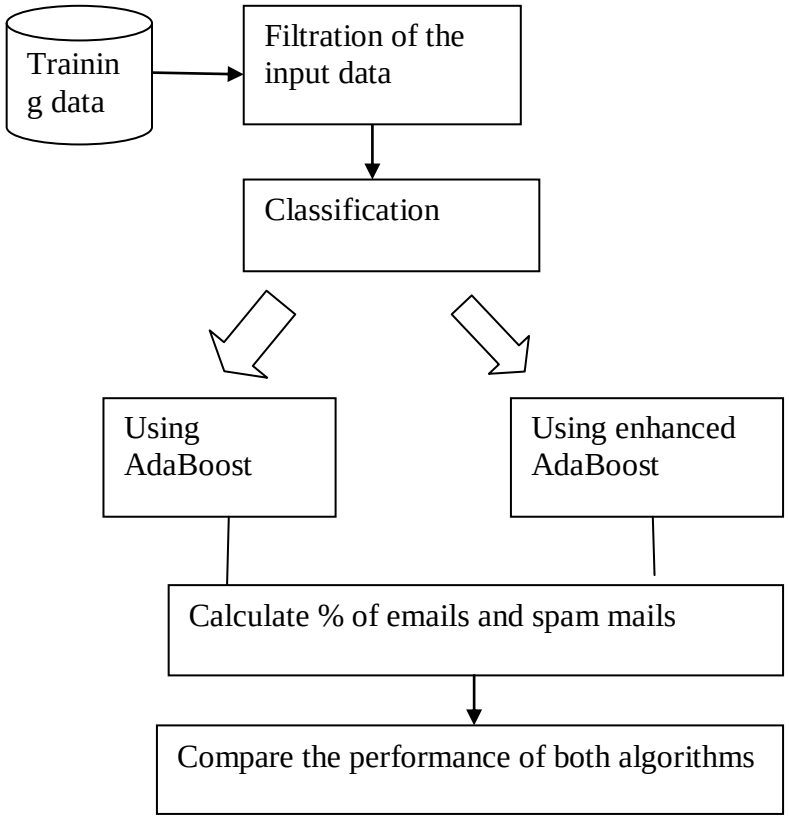


Fig. 3.4 Flowchart of enhanced adaboost algorithm

Classification is done by the adaboost and enhanced adaboost algorithm and then the percentage of correctly and incorrectly emails are calculated from each and the results found from both the algorithm are compared for the analysis purpose. After the complete process, we found that our problem definition of enhanced adaboost algorithm give better results and increases the efficiency or percentage rate by 87% as compared to previous one.

Chapter 4
RESULTS AND DISCUSSION

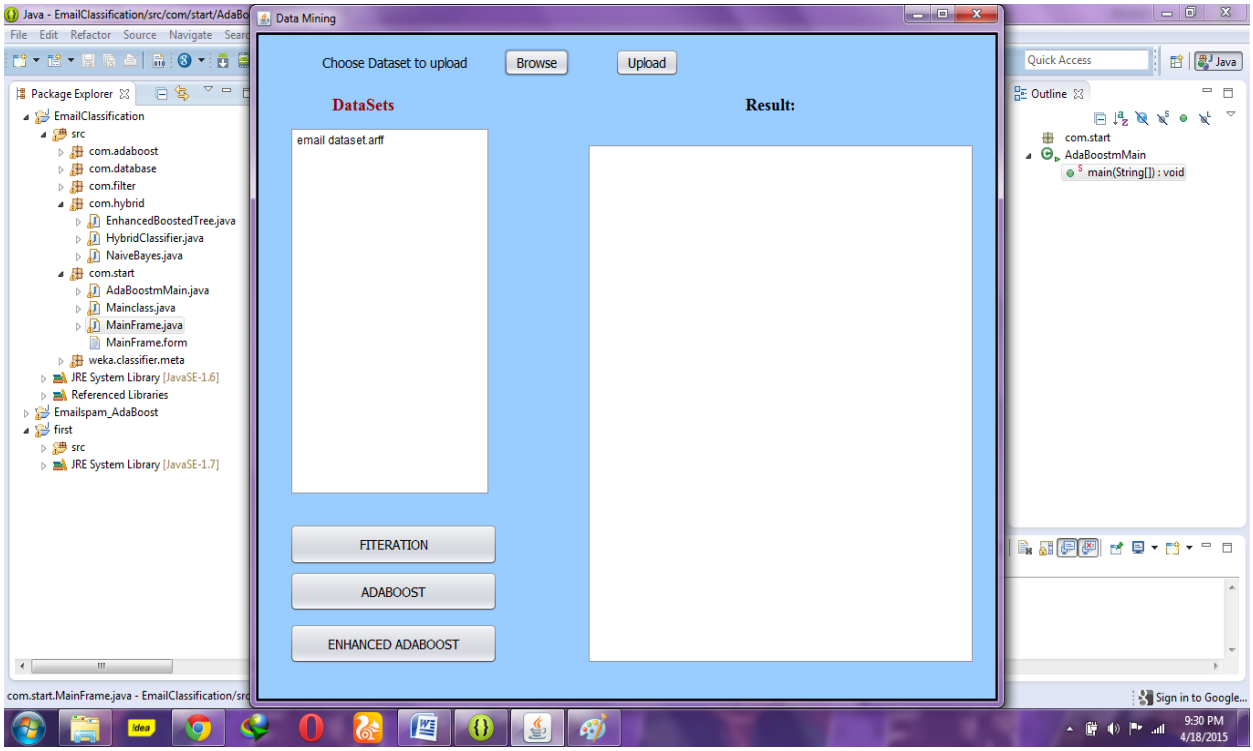
In our enhanced AdaBoost algorithm, we apply more decision making rules if the error rate comes between e=1 to e=4 while calculating the class i.e. on the basis of the error rate we add more weight to the class that will give better class for the email spam prediction.

Following outcomes are expected from the research work:

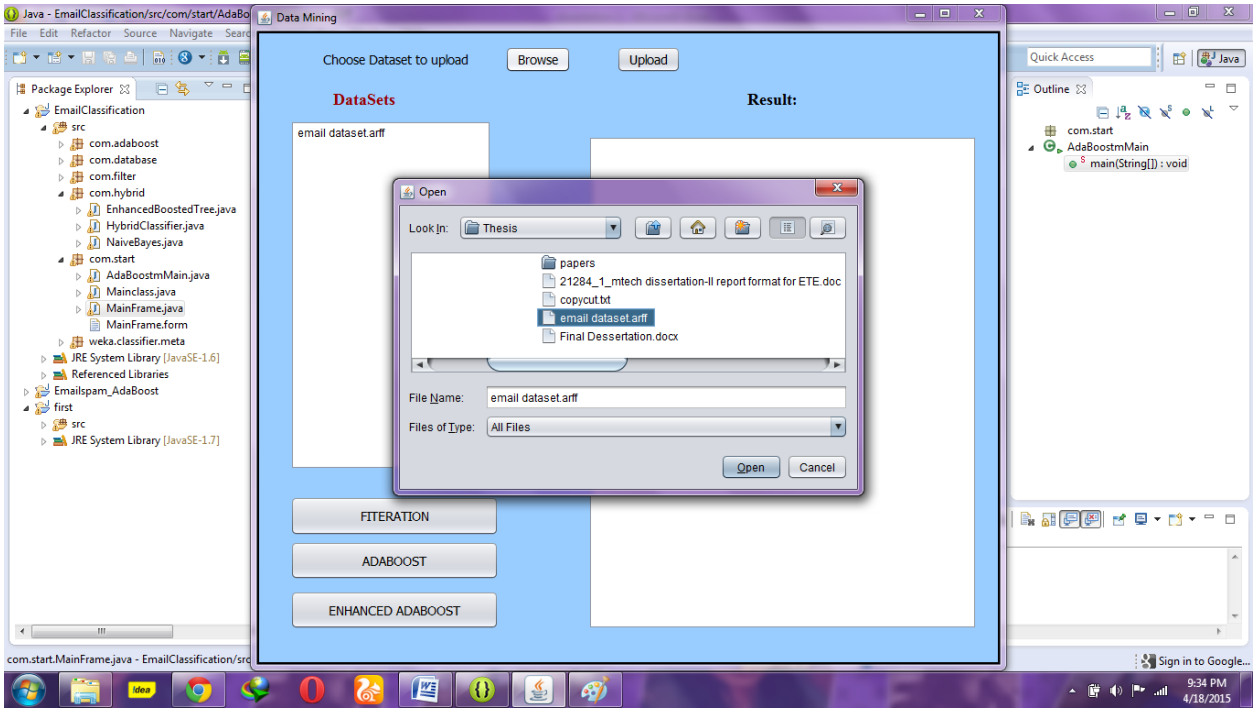
1. Replace the weak learner of AdaBoost with hybrid classifier that contains AdaBoost algorithm and are hybridized on the basis of average of their probabilities.
2. Add more decision making conditions while calculating the class for model prediction i.e. on the basis of error rate adds more weight to class that will give better class for the prediction.

Experimental results:

1. The interface made for the experiments: This is made in the net beans various buttons are depicted for the purpose of the algorithm being designed.

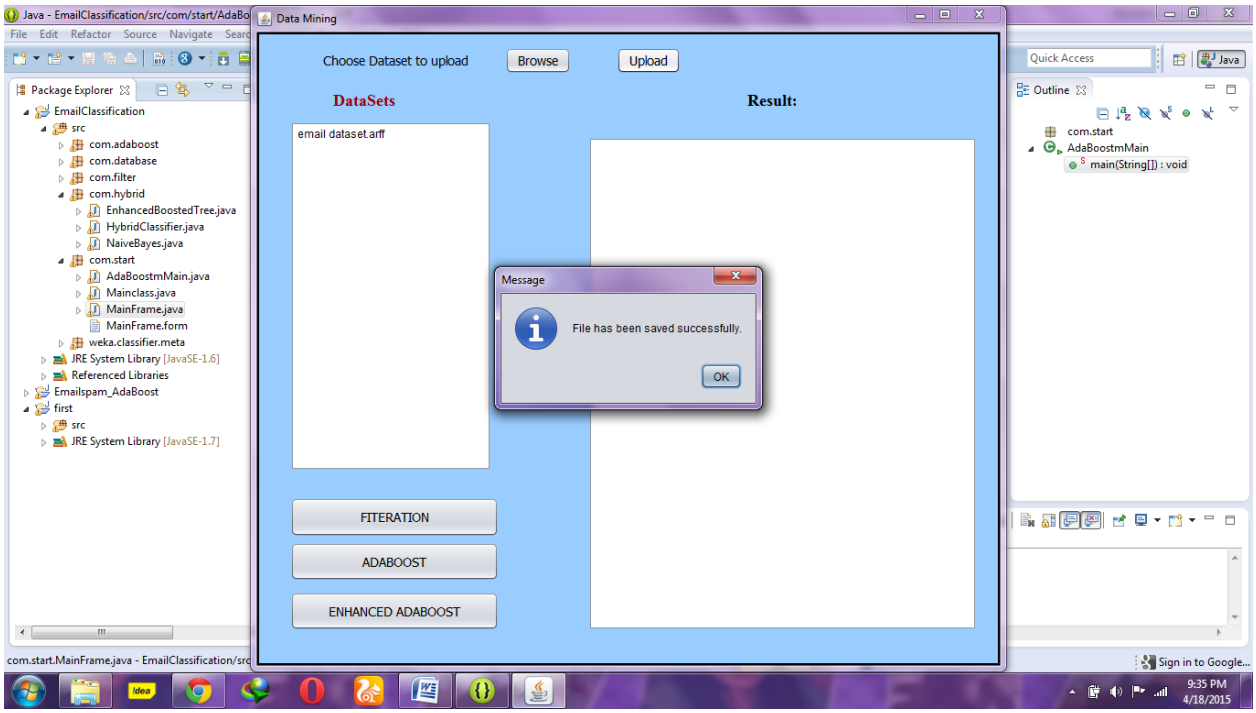


2. Browse the training dataset:



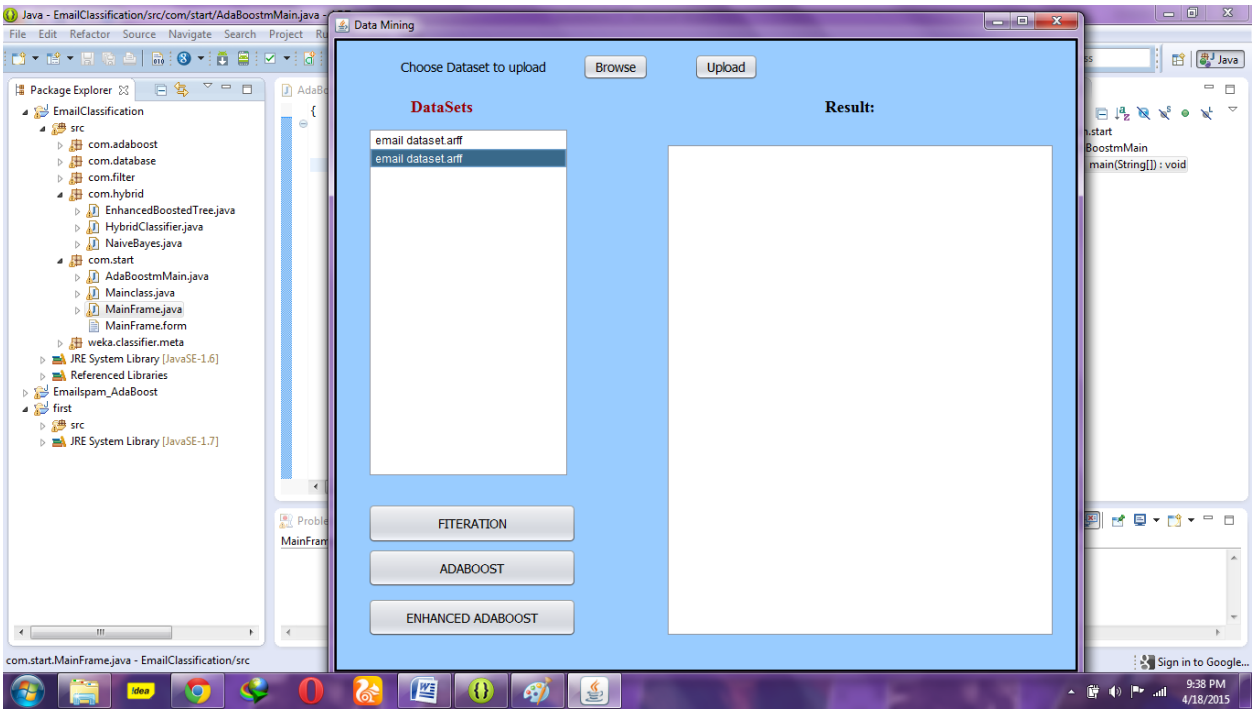
The email dataset has been collected from various sites and is saved in excel sheet in .arff format.

3. File has been uploaded on the screen:

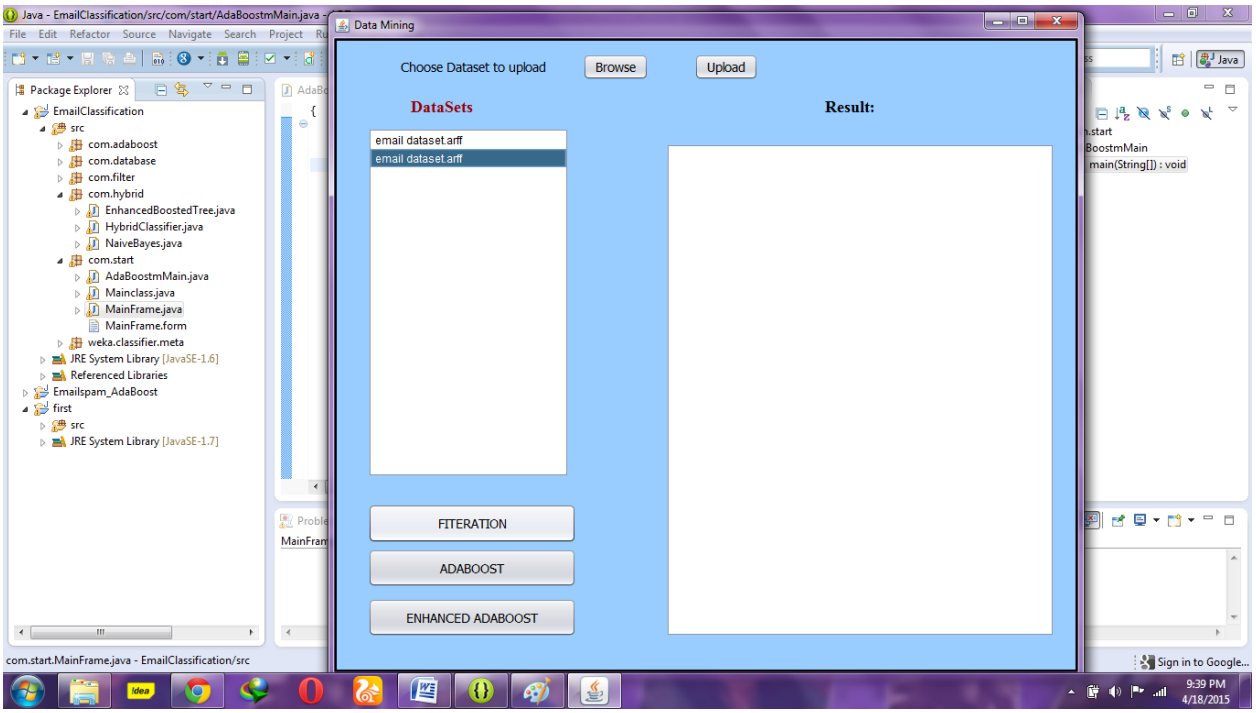


After selecting the particular location of folder, the file has been submitted successfully for the training.

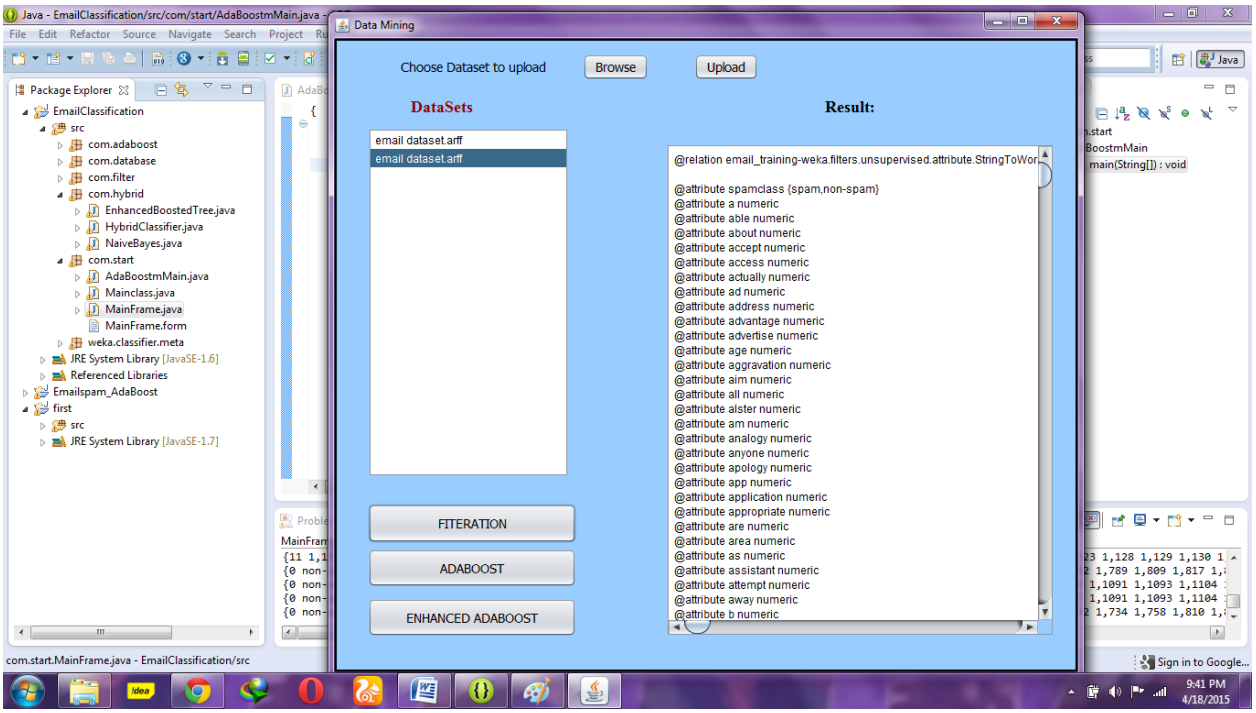
4. Select the uploaded dataset to open it:



5. Filtration is chosen for the pre processing step:

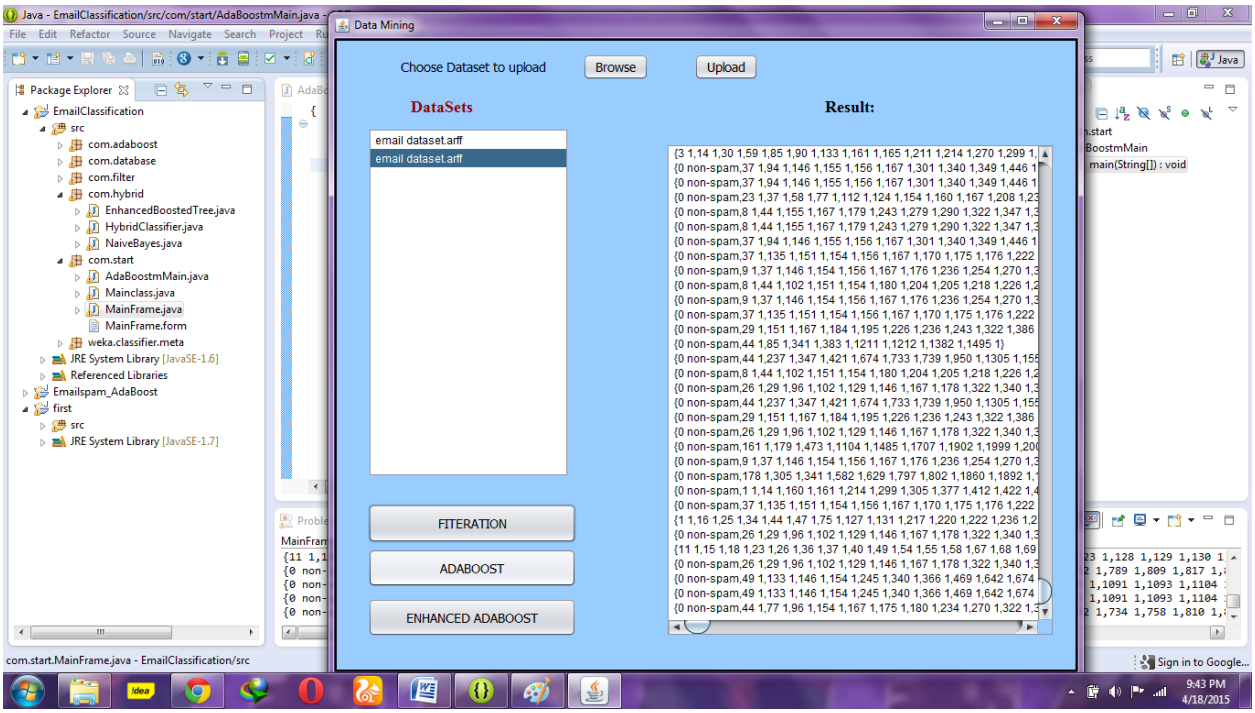


6. Filtration performed:

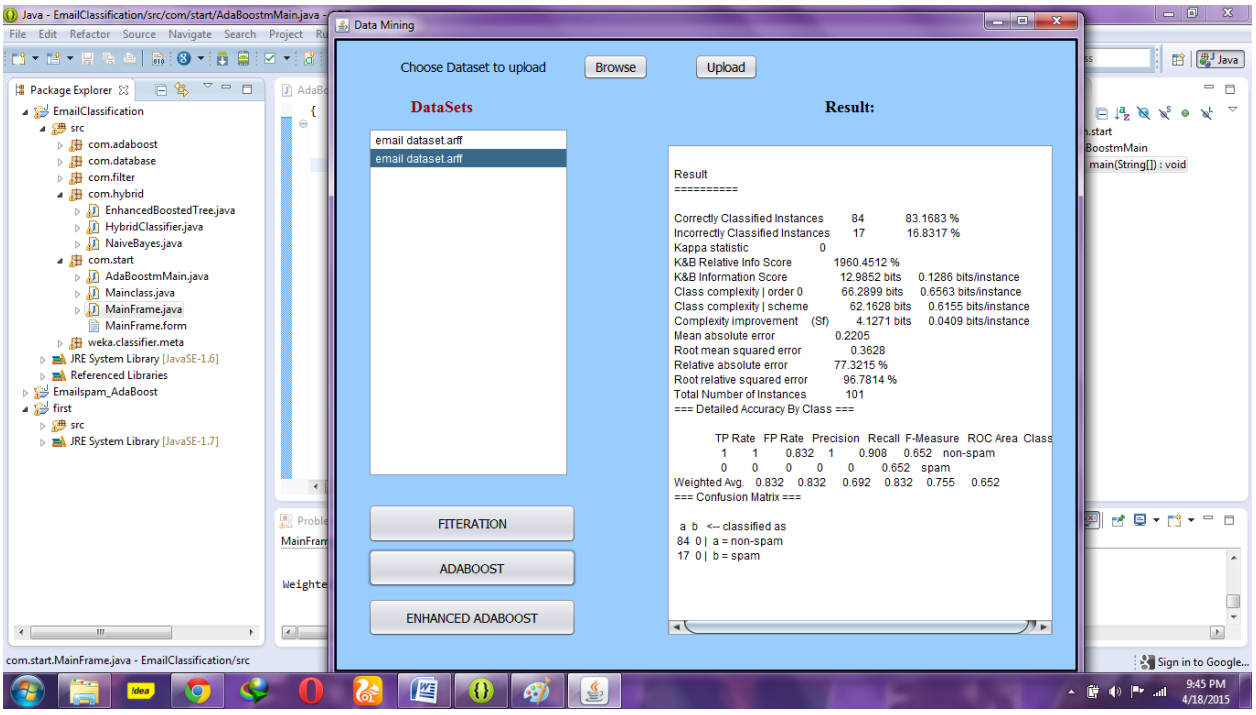


Here the rows of the training data is converted into words for attribute selection and at each line the word is shown from the total no. of instances taken.

7. AdaBoost Algorithm is selected for the calculation of weights and decision stump is used as classifier for the model generation:

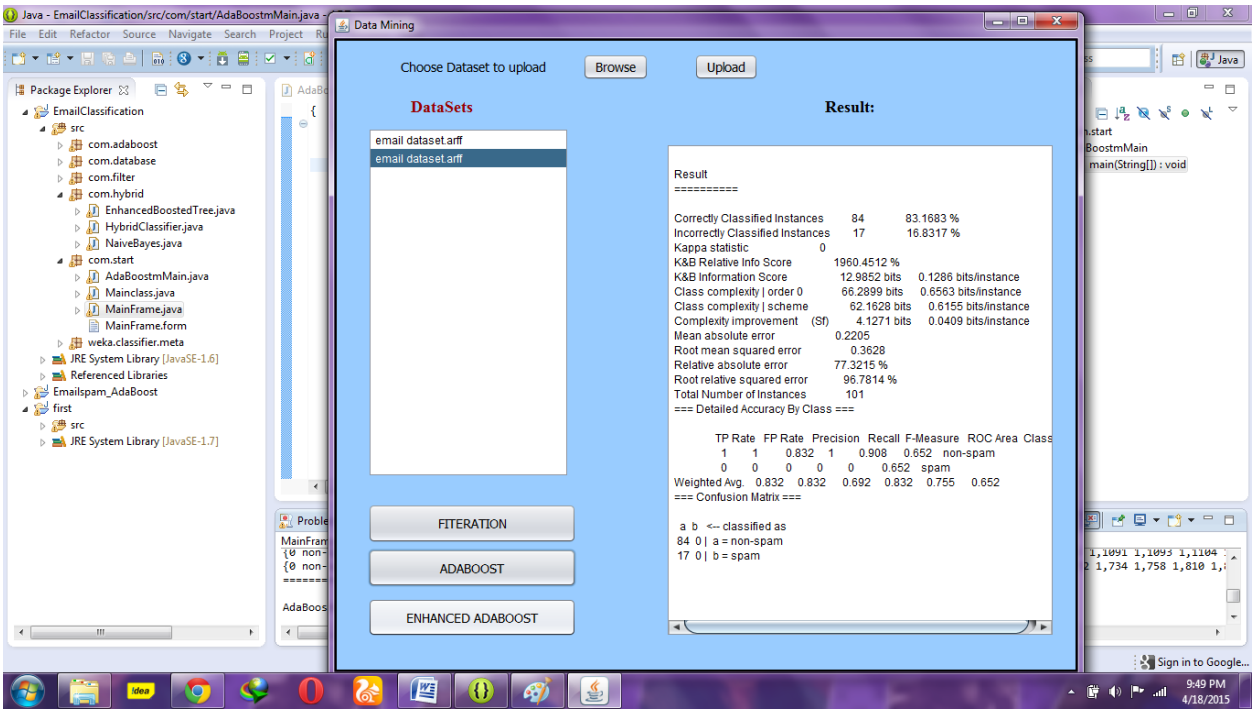


8. Results of Adaboost:

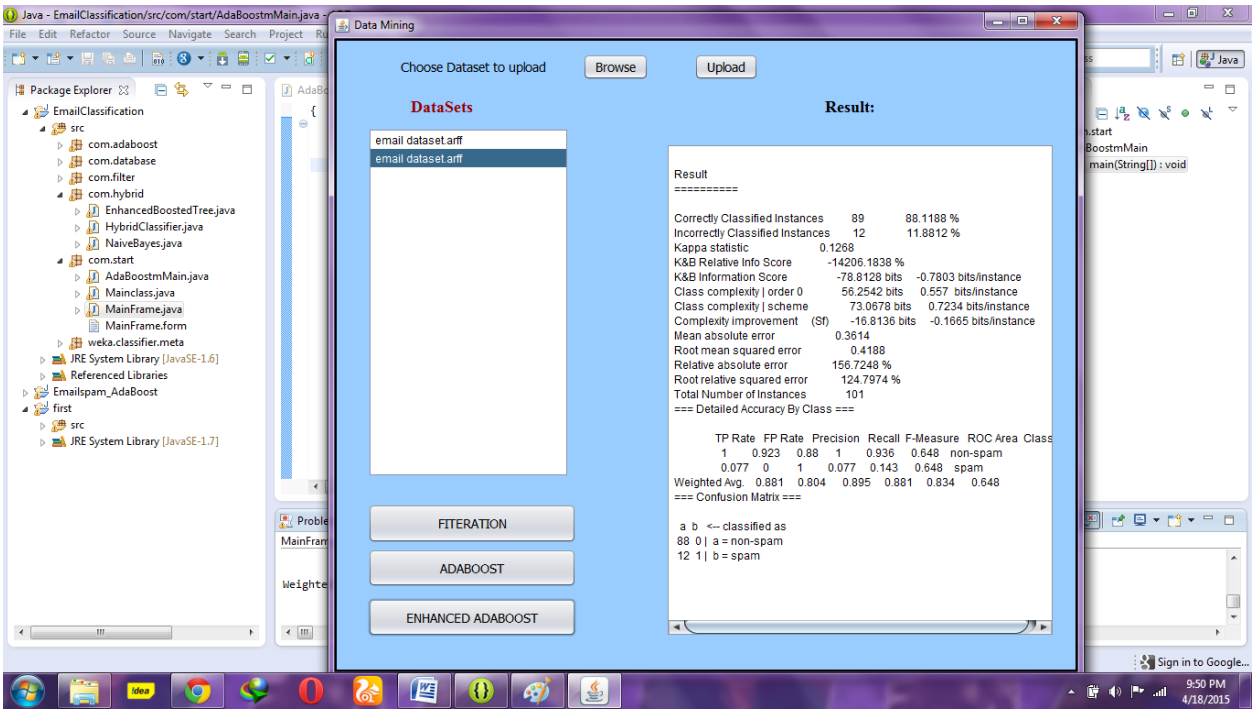


Correctly classified instances are 83.16 % and incorrectly classified instances are 16.83%. Total number of the instances are 101 .

9. Enhanced adaboost algorithm:



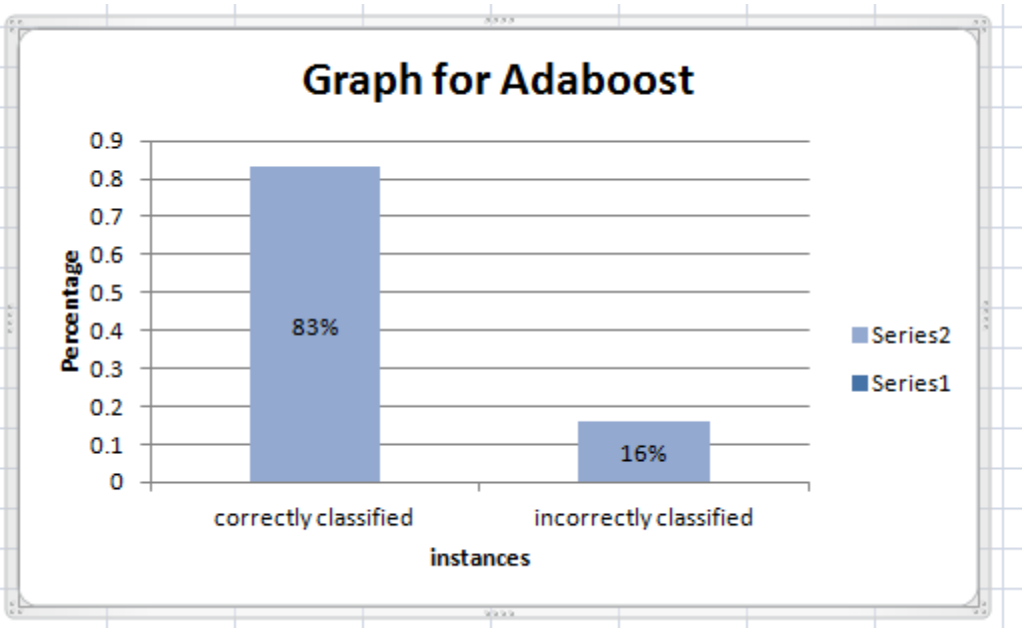
10. Calculation of result using enhanced adaboost algorithm:



Using enhanced adaboost, the number of correctly classified email instances have been increased and improved to 88% when compared to standard adaboost algorithm. And percentage of the incorrectly classified instances has been decreased in enhanced adaboost algorithm to 12%.

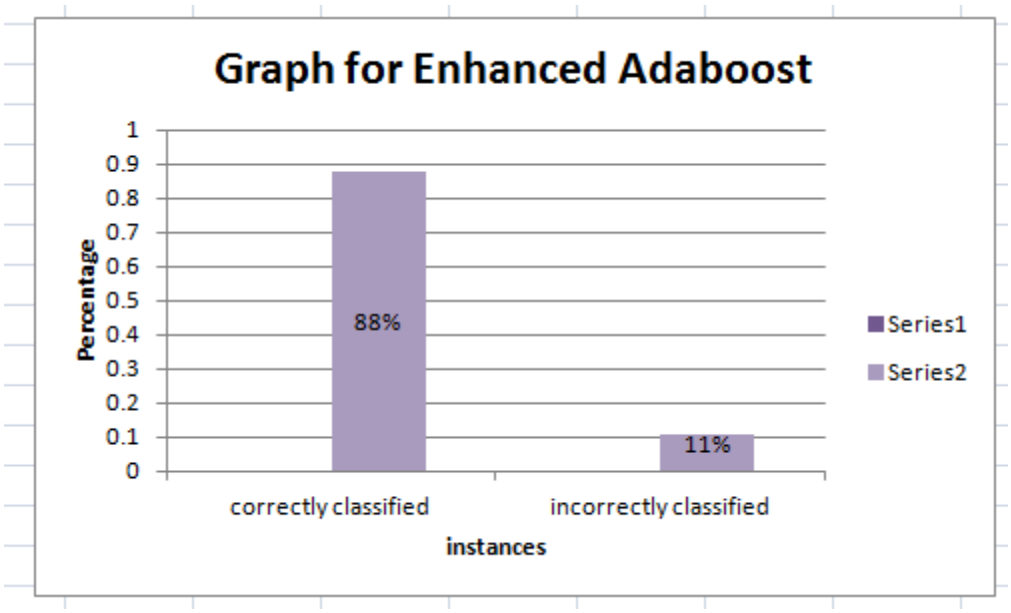
Graphs for the outcomes of the implementation:

1. Percentage of spam and non spams using Adaboost:

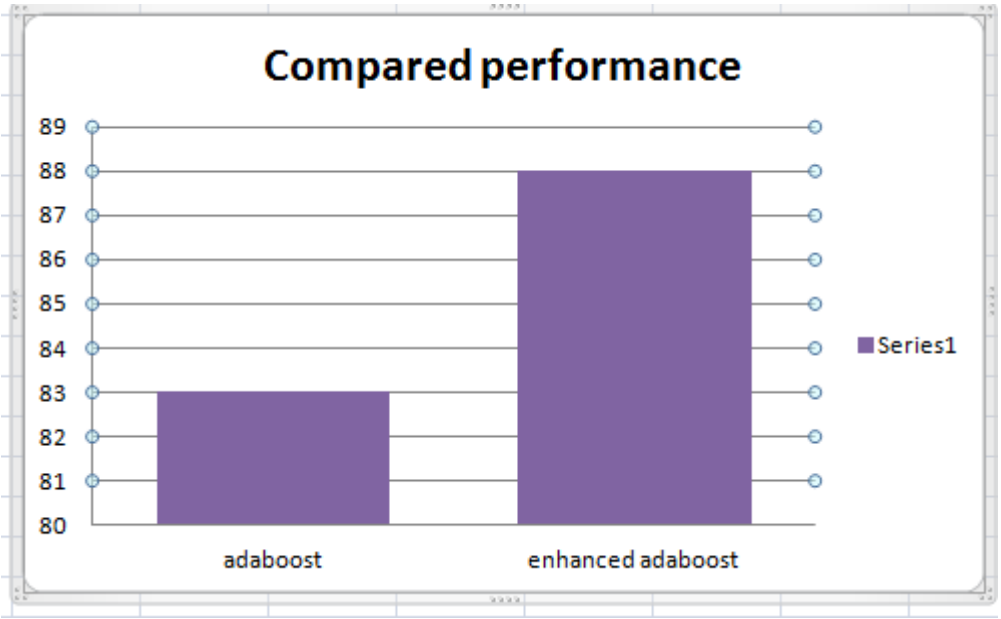


The percentage shown in graph for correctly classified instances i.e. non spam class is 83% and that of Spams is 16 %

2. Percentage of spam and non spam using enhanced adaboost:



3. Compared performance of both the algorithms:



The performance results of both the algorithms have been compared and we find that enhanced adaboost algorithm gives more accurate instances and less spam mails as compared to the adaboost.

Chapter 5

CONCLUSION AND FUTURE SCOPE

AdaBoost algorithm is the best representative algorithm of the boosting family. It does not need any prior knowledge about the weak classifier. A weak classifier are the classifiers that have some guess on how to predict the right labels, but not as much as strong classifiers predict the right labels such as Naive Bayes, Neural Networks or SVM. It classifies with the smallest error on any probability distribution and set of training samples.

We have replaced the weak learner i.e. decision stump of AdaBoost with the hybrid classifier that contains AdaBoost algorithm and Naive Bayes which are hybridized on the basis of average of their probabilities.

Also we have applied certain decision making conditions in calculating the error while calculating the weight for the instances. The enhanced Adaboost have given better results and its efficiency and percentage is more than standard Adaboost algorithm

In future, if someone wants to carry out further research on Adaboost algotihm then more decision making conditions can be applied for the weight of the training instances and other classifier can also be applied for the improvement in the algorithm.

Chapter 6

REFERENCES

Research Papers

[a] Ali Shawkat ABM and Xiang Yang (2007) "Spam Classification Using Adaptive Boosting Algorithm", School of computer sciences, Central Queensland University, Rockhampton, Australia.

[b] Aggarwal C.C. and Zhai C.X. (2012) "A survey of text classification algorithms" pg. 163

[c] E. Schapire Robert (2013) "Explaining AdaBoost".

[d] Ferreira J. Artur and Figueiredo A.T. M´ario (2012) "Boosting Algorithms- A review of methods, theory and applications" pg. 35.

[e] McLeod Dennis and Youn SeongWook "A comparative study for email classification", University of southern California, Los Angeles, USA.

[f] Nizamani Sarwat, Memon Nasrullah, Will Kock Uffe (2011) "Detection of illegitimate Emails Using Boosting Algorithm", Wiil (ed.), Counterterrorism and Open Source Intelligence, Lecture Notes in Social Networks 2, U.K.

[g] Pandey Mayank, Peng Vadlamani Wu and Hui Zhao (2011) "Some Analysis and research of the AdaBoost Algorithm", Computing centre, Henan University, Kaifeng, China.

[h] Ravi (2013) "Text and data mining to detect Phishing websites and spam emails" pg. 559

[i] Seongwook Youn and Dennis McLeod,University of Southern California "A comparative study for email classification"

[j] Zhao Bo (2014) "Target Monitoring On Face Detection Based On Improved AdaBoost Algorithm", the Third Research Institute Of Ministry of Public Security, China.

[k] Windeatt T. and Rolli F. (2003) “Boosting with averaged weight vectors” pg. 15-24

Websites:

http://www.slideshare.net/aman3001/machine-learning-with-ada-boost

http://machinelearningsoftware.org/can-machine-learning-give-investigative-journalism-the-scoop/