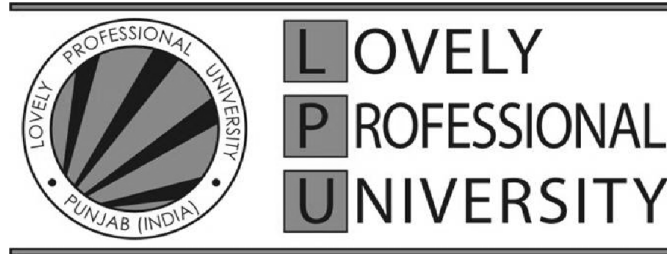


PREPROCESSING AND FILTERING OF NETWORK TRAFFIC CLASSIFICATION
USING ADABOOST ALGORITHM



**PREPROCESSING AND FILTERING OF NETWORK TRAFFIC
CLASSIFICATION USING ADABOOST ALGORITHM**

A Dissertation Submitted

By SUPRIYA KATAL

To

Department of Computer Science and Engineering

In Fulfilment of the Requirement for the

Award of the Degree of

Master of Technology in Computer Science and Engineering

Under the guidance of

Mr. HARDEEP SINGH

Assistant Professor

(May 2015)

PREPROCESSING AND FILTERING OF NETWORK TRAFFIC CLASSIFICATION USING ADABOOST ALGORITHM

PAC form



Discipline: School of Computer Science & Tech

PROJECT/DISSERTATION TOPIC APPROVAL PERFORMA

Name of student: Supriya Katal Registration No: 11211825
Batch: M.Tech (C.S.E) Roll No. RK221A23
Session: 2012-2014 Parent section: K2211
Details of Guide:
Name: Hardeep Singh (Sec) Designation: A.P
U.ID: 16869 Qualification: M.Tech (C.S.E)
Research Experience: 2 yrs

PROPOSED TOPICS

- Web data mining in Ontology
- Traffic network classification using machine learning
- Malware & security issues in data mining

Signature of Guide: Hardeep Singh
16869

*Guide should finally encircle one topic out of three proposed topics and put up for approval before Project Approval Committee (PAC)

*Original copy of this format after PAC approval will be retained by the student and must be attached in the Project/Dissertation synopsis and final report.

*One copy to be submitted to guide.

APPROVAL PAC CHAIRPERSON

Signature: [Signature]
17/04

PREPROCESSING AND FILTERING OF NETWORK TRAFFIC CLASSIFICATION
USING ADABOOST ALGORITHM

CERTIFICATE

This is to certify that **Supriya Katal** has completed M.Tech dissertation titled **Preprocessing and Filtering of Network Traffic Classification using 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 is fit for the submission and the fulfilment of the conditions for the award of M.Tech Computer Science & Engineering.

Date:

Signature of Advisor

Mr. Hardeep Singh

UID: 16869

PREPROCESSING AND FILTERING OF NETWORK TRAFFIC CLASSIFICATION
USING ADABOOST ALGORITHM

DECLARATION

I hereby declare that the dissertation entitled **Preprocessing and Filtering of Network Traffic Classification using 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: Supriya Katal

Reg. No. 11211825

PREPROCESSING AND FILTERING OF NETWORK TRAFFIC CLASSIFICATION USING ADABOOST ALGORITHM

ABSTRACT

The use of classification with machine learning algorithms is used in various approaches. In previous approach they had gone through correlation feature selection and we are using consistency feature selection. In previous approach there is no use of feature selection in Naive Bayes and the performance of proposed approach is more accurate than previous approach. The proposed algorithm used is Naive Bayes with feature selection, Bagging and Boosting. In our work we proposed preprocessing and filtering of data with the help of classification and statistical classifiers are used in our approach. At the end we compared the results and evaluate the parameters like accuracy, precision, recall, TP rate etc.

ACKNOWLEDGEMENT

I would like to present my deepest gratitude to **Astt. Prof. Hardeep Singh** for his guidance, advice, understanding and supervision throughout the development of this dissertation study. I would like to thank to the **Project Approval Committee members** for their valuable comments and discussions. I would also like to thank to **Lovely Professional University** for the support on academic studies and letting me involve in this study.

Supriya Katal

PREPROCESSING AND FILTERING OF NETWORK TRAFFIC CLASSIFICATION
USING ADABOOST ALGORITHM

TABLE OF CONTENTS

Chapter no.	Page no.
Chapter 1(Introduction).....	1-16
1.1 Data mining.....	1
1.2 Data mining goals.....	4
1.3 Technologies used.....	4
1.4 Applications of Data mining.....	6
1.5 Tools used in data mining.....	7
1.6 Classification.....	8
1.7 Network traffic classification.....	12
Chapter 2(Literature Review).....	17-32
Chapter 3(Present Work).....	33-42
3.1 Problem formulation.....	33
3.2 Objective.....	34
3.3 Research methodology.....	34
Chapter 4(Result and Discussions).....	43-50
Chapter 5(Conclusion and Future Scope).....	51
Chapter 6(References).....	52-54
Chapter 7(Appendix).....	55
7.1 Glossary.....	55
7.2 Publications.....	55

PREPROCESSING AND FILTERING OF NETWORK TRAFFIC CLASSIFICATION
USING ADABOOST ALGORITHM

TABLE OF FIGURES

Figure no.	Page no.
Figure 1.1 CRISP data mining process.....	4
Figure 1.2: Training data are analyzed by a classification algorithm.....	9
Figure 1.3: Concept of Decision Tree.....	10
Figure 1.4: Evolutions of protocols and classification techniques.....	14
Figure 3.1: Flowchart for previous approach.....	38
Figure 3.2: Flowchart for new approach.....	39
Figure 4.1: Netbeans IDE 8.0.....	43
Figure 4.2: Choose dataset to upload.....	44
Figure 4.3: Feature Selection.....	45
Figure 4.4: Dataset extracting best features.....	45
Figure 4.5: Naive Bayes using Feature Selection.....	46
Figure 4.6: Naive Bayes without Feature Selection.....	47
Figure 4.7: Perform mining using C 4.5.....	47
Figure 4.8: Perform mining using Bagging.....	48
Figure 4.9: MySQL server.....	49
Figure 4.10: Naive Bayes accuracy without Feature Selection.....	49
Figure 4.11: Naive Bayes accuracy with feature selection.....	50
Figure 4.12: Known and unknown Traffic.....	50

PREPROCESSING AND FILTERING OF NETWORK TRAFFIC CLASSIFICATION USING ADABOOST ALGORITHM

ABBREVIATIONS

SVM	Support Vector Machine
HTTP	Hypertext transfer protocol
P2P	Peer to Peer
QoS	Quality of Service
KDD	Knowledge Discovery of Data
OLAP	Online Analytical Processing
ROLAP	Relational Online Analytical Processing
MOLAP	Multidimensional Online Analytical Processing
HOLAP	Hybrid Online Analytical Processing
OLTP	Online Transaction Processing
WEKA	Waikato Environment for Knowledge Analysis
KNN	K-Nearest Neighbor
CFS	Consistency Feature Selection
TPR	True Positive Rate
FPR	False Positive Rate
FNR	False Negative Rate
TNR	True Negative Rate

CHAPTER 1

INTRODUCTION

1.1 Data Mining

The term Data Mining [1] is “knowledge mining from data”. The term knowledge mining does not mean that it mines huge amount of data. It is also refers to as knowledge discovery of data (KDD). The view of this database system and data warehousing communities is of three-tier system. This type of mining plays an essential role in the knowledge discovery process. It is the process of discovering interesting patterns from massive amounts of data or database. A Pattern makes the mining interesting or if it is valid then easily understood by us. In data mining knowledge obtained from patterns which are interested, data mining is a form of knowledge mining [2]. The process of Knowledge discovery contains the cleaning of data, integration of data, selection of data, and transformation of data, evaluate the patterns and present the knowledge. Basically, Data mining is used for to process the information and obtained the knowledge or interesting patterns from the large amount of data (or information). The data is very important for bussiness growth. An organisation uses the data mining process for mine the information (or data) and generates the knowledge from it. This knowlegde is important for business growth and an oraganisation in decision making process. It includes:

- i. Data cleaning
- ii. Data integration
- iii. Data selection
- iv. Data transformation
- v. Pattern evaluation
- vi. Knowledge presentation

Data cleaning is to remove unnecessary errors and the data which is not important to us. When we combine various different data forms into one then it is said to be data integration. In many business intelligence fields the two term data cleaning and data integration refer to as preprocessing step and the outcome of this is stored in warehouses. When we perform any analysis task on data then it meant to be selected data. For

PREPROCESSING AND FILTERING OF NETWORK TRAFFIC CLASSIFICATION USING ADABOOST ALGORITHM

performing transformation we apply summary or aggregation operations. To represent pattern we deal with true knowledge which is based on interesting measures. Knowledge presentation used to perform and present mined knowledge to users.

The data sources from where data mining mines data are database, data warehouse, data repositories. In data mining drill down and roll up plays an important role for representing a dimensional data. Data which we mine from other sources are valuable and data warehouse keep all the data repositories very well maintained. Metadata is the data about data like data in dictionary. The one word represents many meaning. The structured and unstructuredness of data is to maintain semantic knowledge of data and to keep data beneficial for further use. We can say data mining is a repository of huge amount of data which is valuable and have knowledge of each specific term. For n-dimensional space or cube we do drill down and roll up, OLAP and OLTP servers are also used in data mining. Further [2] OLAP is of three parts: ROLAP, MOLAP, HOLAP servers.

1.1.1 Data Mining Process

Basically the data mining used for to mine the data by using the mining process and obtained the patterns and stores it into the database and these patterns are used for making the decisions of an organization [17]. Data mining is a form of knowledge which is used for making the decisions for an organization.

Data mining involves the following phases:

- i. Problem definition
- ii. Data exploration
- iii. Data preparation
- iv. Modeling
- v. Evaluation
- vi. Deployment

i. Problem Definition

Firstly understands the business problem for starting the data mining project. The business expert's work together to define the objectives of the business and converted it into problem definition. To define the exact problem means to prove the hypothesis true and Data mining tools are not used in this phase.

PREPROCESSING AND FILTERING OF NETWORK TRAFFIC CLASSIFICATION USING ADABOOST ALGORITHM

ii. Data Exploration

The domain experts collect, describe, and explore the data and also check the quality of the data. To explore the data mining process we need an exact meaning to our knowledge and how we can mine data for further holdings so that we are able to take expert advice for exploring the mined data for future reference.

iii. Data Preparation

In data preparation phase, the data models are used for modeling process, the domain experts build the data models. Data mining functions accept the data in format so, domain experts collect, cleanse and format the data and also create new modeling attributes. In this phase, the tools are used for selecting the records, tables and attributes.

iv. Modeling

In modeling phase, same type of data mining problem can uses the different mining functions. Experts select and apply the various kinds of mining functions. In this phase, modeling and evaluation two phases are coupled to gather, both are used to change the parameters and achieved the optimal values.

v. Evaluation

In evaluation phase, experts evaluate the models. Model satisfies their expectations; they go to the next phase. If the model does not satisfy, go to back modeling phase and change the parameters and achieve the optimal values. In this phase, experts evaluate the models and results.

vi. Deployment

In deployment phase, mining results are used by experts and results are exported into database tables. E.g. Spreadsheets. The data sources from where data mining mines data are database, data warehouse, data repositories. The structured and unstructuredness of data is to maintain semantic knowledge of data and to keep data beneficial for further use. We can say data mining is a repository of huge amount of data which is valuable and have knowledge of each specific term.

PREPROCESSING AND FILTERING OF NETWORK TRAFFIC CLASSIFICATION USING ADABOOST ALGORITHM

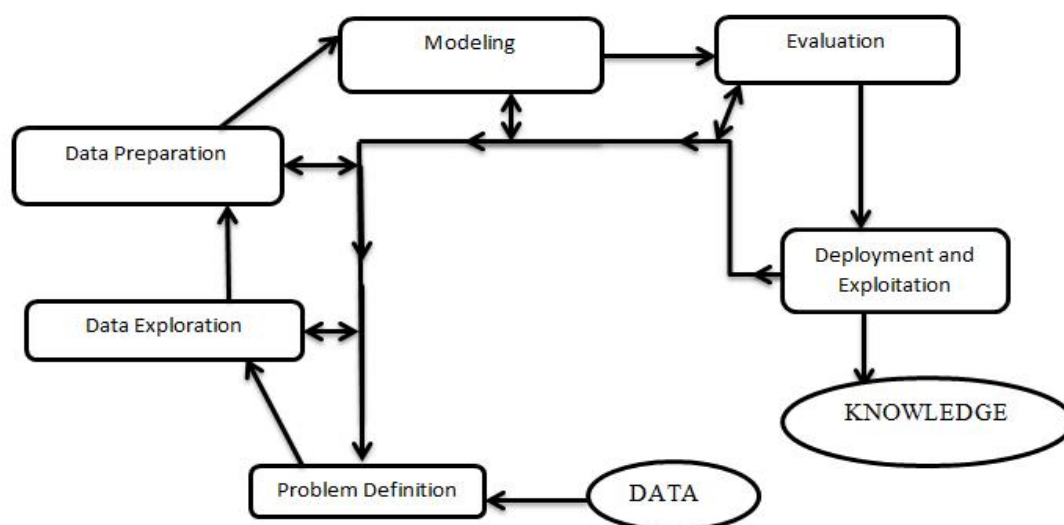


Figure 1.1 CRISP Data Mining Process (Cross Industry Standard Process for Data Mining [2])

1.2 Data Mining Goals

Data mining is Process of discovering and fetching the valuable information from the large amount of data. For business, this valuable information is very useful in decision making process. In company, Data mining can also be used to identify patterns from the organization data. For example: If customers buy product *A* and product *B*, and they wants to buy the product *C* as well? For this, Data mining is used for identification of the goals used for same group of customers. So, Data mining tools ease and automate the process of discovering this kind of information from large stores of data.

1.3 Technologies Used

The various technologies used in data mining are:

- a) Machine learning
- b) Statistics
- c) Database system
- d) Data warehouse
- e) Pattern recognition
- f) Visualization
- g) Information retrieval

PREPROCESSING AND FILTERING OF NETWORK TRAFFIC CLASSIFICATION USING ADABOOST ALGORITHM

a) Machine learning

This means how [1] computer can learn or improve their performance based on data. In machine the basic concept is to deal with computer so that they can automatically learn to capture pattern and from those achieved patterns can make intelligent decisions which is based on data related to the pattern. Machine learning is a fast growing discipline. It is of further three types:

- Supervised machine learning
- Unsupervised machine learning
- Semi-supervised machine learning

In supervised learning, it deals with classification of data. It deals with classification of data. It always comes from training dataset which is of labeled. In unsupervised learning, deal with clustering and not labeled set. We use cluster to form classes in data. Semi-supervised learning falls under both labeled and unlabeled.

b) Statistics

In data mining refer to as flow diagram or flow or work in which we can show our improvement in the form of graphs and other related methods or applications.

c) Database System

It deals with creation maintenance and organization of database systems uses many methods for the creation and main of Database such as indexing, query optimization, query processing and accessing methods. . In many business intelligence fields the two term data cleaning and data integration refer to as preprocessing step and the outcome of this is stored in warehouses. When we perform any analysis task on data then it meant to be selected data. For performing transformation we apply summary or aggregation operations. To represent pattern we deal with true knowledge which is based on interesting measures. Knowledge presentation used to perform and present mined knowledge to users. To show the data is structured or unstructured.

d) Data warehouse

Data warehouse means to gather data from multiple sources like external and to perform OLAP operations in multidimensional database. The structured and unstructuredness of data is to maintain semantic knowledge of data and to keep data beneficial for further use.

PREPROCESSING AND FILTERING OF NETWORK TRAFFIC CLASSIFICATION USING ADABOOST ALGORITHM

We can say data mining is a repository of huge amount of data which is valuable and have knowledge of each specific term. For n-dimensional space or cube we do drill down and roll up, OLAP and OLTP servers are also used in data mining. Further OLAP is of three parts: ROLAP, MOLAP, HOLAP servers.

e) Information Retrieval

Information Retrieval is to search for knowledge from internal sources which reside on web and to generate the bag of words in the document.

f) Visualization

It is use to visualize things for discovery of new mined knowledge so that gathered knowledge can help in reviewing data for success data mining research.

1.4 Applications of Data Mining

There are various areas in which data mining is required such as

1.4.1 Financial data analysis

1.4.2 Telecommunication Industry

1.4.3 Retail industry

1.4.4 Other scientific applications

1.4.1 Finance Data Analysis

In banks and finance industry data being used is reliable and it is having a good product. So, for data analysis uses the data mining in systematic way. It includes the various steps such as:

- i. For multidimensional data mining and data analysis needs to design and construct the various centers for mined data.
- ii. The customer credit policy analyzes the prediction of loan payment.
- iii. For targeted marketing, clustering and classification of customers.
- iv. Detects the money thafterts and other finance crimes.

1.4.2 Telecom Industrialization

PREPROCESSING AND FILTERING OF NETWORK TRAFFIC CLASSIFICATION USING ADABOOST ALGORITHM

The telecom organization is the most emerged and popular field which provides the various services such as Internet messenger, pager, fax and cellular phone etc. The telecommunication industry is rapidly expanding because of the development of new computer and communication services [5]. In the telecommunication industry, data mining is used to telecommunications industry to identify the patterns and also helps to catch the fraud activities. So, data mining is improves the telecommunication as the multidimensional analysis and association the data, and the sequential patterns analysis. For telecommunication data analysis uses the visualization tools to identify the unusual patterns and the fraudulent pattern.

1.4.3 Retail Industry

In the retail industry field includes the purchasing history of customer, sales, consumption and services. Data mining has the ability to collects the information from various fields, the Data Mining in Retail Industry helps to customer to identifying the buying patterns and trends. In the retail industry, the use of data mining is that to improve the good customer satisfaction and quality of customer service. The examples that are related to the retail industry as below:

- i. Construct the design of the data warehouses.
- ii. Multidimensional analysis of the customers, sales, time, products and region.
- iii. Product recommendation.
- iv. Customer Satisfaction

1.4.4 Other Scientific Applications

In the Scientific domain the big amount of data collected or large amount of data sets are generated such as geosciences, astronomy etc. The datasets are established due to the fast numeric simulations in the fields such as ecosystem modeling, climate and fluid dynamics, chemical engineering etc. In the field of scientific application the data mining are used to discover the interesting patterns.

1.5 Tools Used In Data Mining

In Data Mining different types of tools are available. Most common tools of data mining are classified into three categories such as:

PREPROCESSING AND FILTERING OF NETWORK TRAFFIC CLASSIFICATION USING ADABOOST ALGORITHM

a) Text Mining Tools

b) Dashboards Data Mining Tools

c) Traditional Data Mining Tools

a) Text Mining Tools

First tool of data mining is text mining tool. This tool is used to mine the information or data from different kind of text and also scan the contents then convert the data into a compatible format, and the Scanned content/data can be unstructured or structured.

b) Dashboards Data Mining Tools

The third common tool of data mining is dashboards. This tool is used to monitor the database data and this tool is installed in computers, to monitor the data and provides the information about the data updates. These types of functionality such as monitor and provide the information about the any change of data, it makes dashboards easy to use and manage the data.

c) Traditional Data Mining Tools

Second tool of data mining is traditional data mining tool. This tool is used to establish the data patterns of the many companies [5]; it can be done with the many complex algorithms and techniques. On the desktop, this tool is installed to monitor the data, and also help to capture the information/data from the outside databases.

1.6 Classification

- Extracts models
- Predict class labels
- Classification methods proposed by researchers in machine learning, pattern recognition and statistics.
- Memory resident

Classification [1] is two step processes: In first step, we build a classification model based on previous data. In the second step, we determine if the models accuracy is acceptable, and if so, we use the model to classify new data. Also we can say, training set and learning set. The training data and classification algorithm suggesting classification rules like if name is Bill lee then loan_decision is risky. In this classification we can also

PREPROCESSING AND FILTERING OF NETWORK TRAFFIC CLASSIFICATION USING ADABOOST ALGORITHM

suggest that the age of the person is youth, senior or middle_aged. Mainly, the classification used to extract models from their predicted class labels.

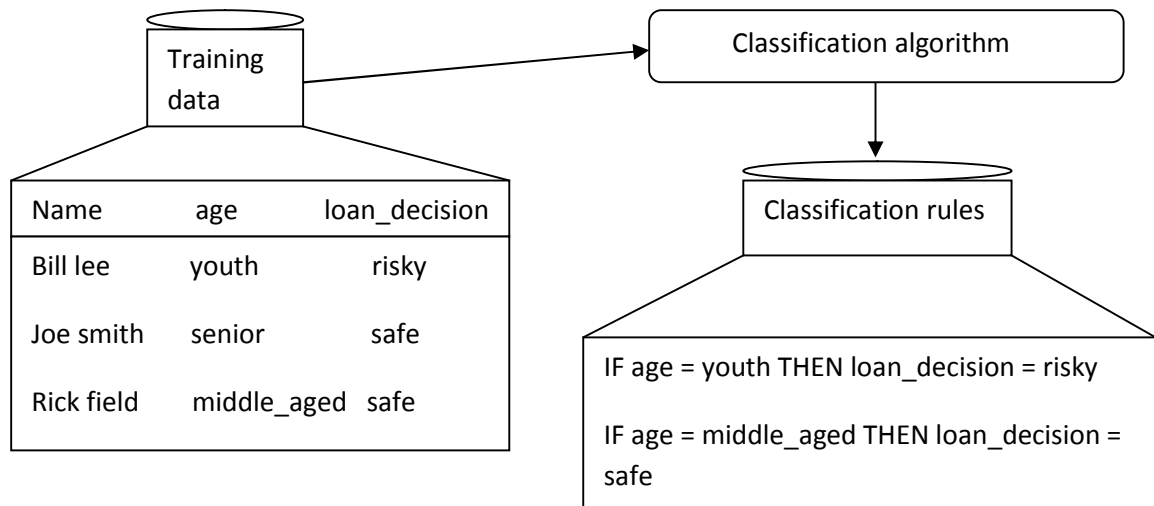


Figure 1.2: Training Data are analyzed by a Classification Algorithm [1]

The tuples which are individual and are of training set refer to as training tuple and another name for this is supervised learning. And the opposite, in which training set are not known refer to as unsupervised learning. We can perform classification on the basis of three factors:

- Accuracy
- Precision
- Recall

Some of the algorithm used in classification is:

- a) Decision tree
- b) Naive bayes theorem
- c) Rule-based classification
- d) Support vector machines
- e) Lazy learners
- f) Genetic algorithm

- a) Decision tree

PREPROCESSING AND FILTERING OF NETWORK TRAFFIC CLASSIFICATION USING ADABOOST ALGORITHM

It is the learning of class-labeled training tuples. [1] It is a flowchart like tree structure. In which node represents an attribute, branch represents an outcome and leaf node represents a class label. Root node is topmost node in tree. In figure 1.3, it shows customer wants to buy iphone and this tree shows customer will buy or not.

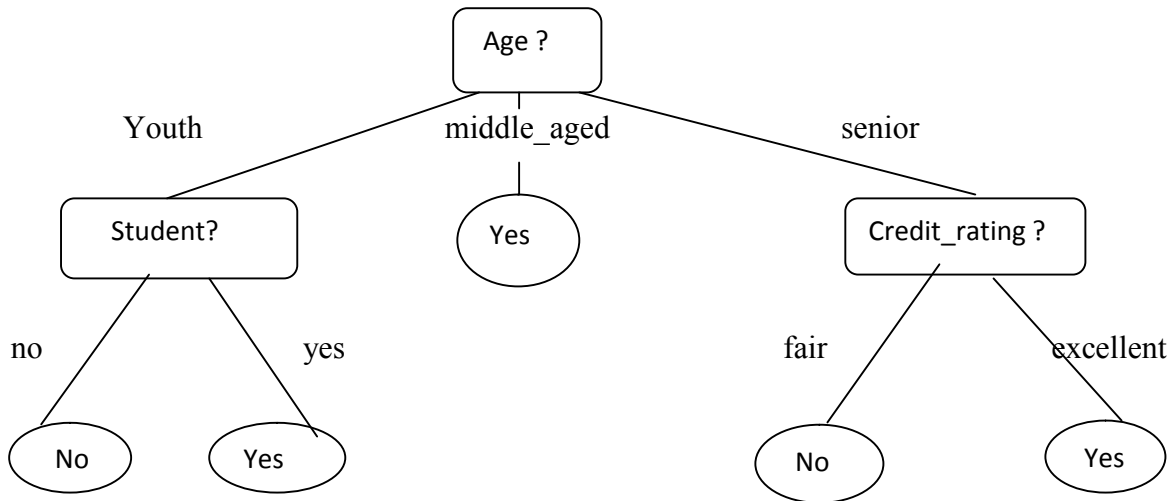


Figure 1.3: Concept of Decision Tree [1]

Internal nodes denote rectangle and leaf nodes denote ovals [1]. Some decision tree algorithm produce only binary tree (where each node branches to exactly two other nodes), whereas others can produce non binary trees.

b) Naive bayes theorem

Naive bayes are known as statistical classifiers. [1] They can predict class membership probabilities such as the probability that a given row belongs to a particular instance. It assumes that the effect of class is independent of the values of the other attributes. This assumption is called class-conditional independence. Bayes' theorem is useful in that it provides a way of calculating the posterior probability, $P(H|X)$. Bayes' theorem is [1] :

$$P(H|X) = \frac{P(X|H) P(H)}{P(X)}$$

where $P(H)$ is prior probability.

c) Rule-based classification

[1] It is represented by IF- THEN rules. An IF-THEN rule is an expression of the form:

PREPROCESSING AND FILTERING OF NETWORK TRAFFIC CLASSIFICATION USING ADABOOST ALGORITHM

IF condition THEN conclusion

For example,

R1: IF age = youth AND student = yes THEN buys_computer = yes

d) Support Vector Machine

It is a method [1] for the classification of both linear and non-linear data. It uses a non-linear mapping to transform the original training data into a higher dimension. Within this new dimension, it searches for the linear optimal separating hyperplane. In this hyperplane data from two classes can be separated. We can find SVM using support vectors. The first paper for SVM is given by Vladimir Vapnik and Colleagues Bernhard Boser and Isabelle Guyon. SVM used for numeric prediction and classification. They applied in areas like handwritten digit recognition, object recognition and speaker identification. In SVM, we use graph to represent separate classes.

A [1] separating hyperplane can be written as:

$$W.X + b = 0$$

e) Lazy Learner (K-nearest neighbor classifier)

This method is used when large training sets are available. They are based on learning by analogy. Every tuple we store in n-dimensional space. K-nearest neighbor searches for unknown tuple. Then K training tuples are “nearest neighbor” of unknown tuple. For calculating Euclidean distance [1] between two tuples, say

$$\text{Dist}(X_1, X_2) = \sqrt{\sum_{i=1}^n (x_{1i} - x_{2i})^2}$$

f) Genetic Algorithm

By randomly generated rules an initial population is created. [1] Each rule can be represented by a string of bits. In this fittest as well as offspring rules are generated. Offspring are created by applying genetic operators such as crossover and mutation. In crossover substrings from pairs of rules are swapped to form new pairs of rules. In mutation, randomly selected bits in a rule's string are inverted. Genetic algorithm is easily parallelizable and has been used for classification as well as other optimization problems. Given a set of objects, each of which belongs to a known class, and each of which has a known vector of variables, our aim is to construct a rule which will allow us to assign future objects to a class, given only the vectors of variables describing the future objects.

1.7 Network Traffic Classification

Computer networks became very important for our life. Many people use them in everyday life and many companies need them for their business. Unfortunately, there is also an effort to misuse the network in order to thwart illegal distribution of copyrighted works, send fraudulent messages, attack other clients, etc. These activities can be classified as threats which may harm users, companies or artists hence there must be a way of protecting them. Each of the treat requires a different approach to deal with it. The defence can be for example properly configured network firewall, updated operating system and applications, law restrictions or network traffic inspection and blocking unwanted applications.

Traffic classification is an automatic process for generating traffic according to different parameters into a number of traffic classes. In network traffic data is wrapped in packets and each packet contains control information and user data. For UDP size limit of packet is 64kb and for TCP there is no size limit. Communication which is provided by the socket that is between two computers using TCP and UDP protocol. Due to increase in internet growth as well as emerging protocols and applications, the internet bandwidth has to be promoted. The more increase in internet, more bandwidth is requires and more the use of security threats. Traffic classification deals with traditional method and statistical methods. In traditional methods port-based and payload-based methods come which now a day is not used. To overcome their problems we came across flow-based statistical method and the issues or challenges which is generating are of selection of flow statistical features and uniqueness proof of flow statistical features. So from these issues we went through many research papers and recognize the problem of Preprocessing and Filtering of network traffic classification. We are taking supervised learning because it works in classification.

Traffic classification has been emerging day by day from past few years. It is widely used in networks, including intrusion detection, security and research. Many of the protocols and proposed applications have been investigated and developed by using machine learning algorithm. The growth of today's internet is giving popularity to the needs of research. Many emerging tools have been proposed by many of the researchers. This is the hot area of research in emerging field of networks. With the use of machine learning algorithms which have been used in data mining fields taken in convergence.

PREPROCESSING AND FILTERING OF NETWORK TRAFFIC CLASSIFICATION USING ADABOOST ALGORITHM

Data mining is knowledge discovery of data. It includes many features like clustering, classification, outlier detection, noise detection etc. But in our field we had combined the research area with networking so that an better or optimal solution could be found. In past year research fields, traditional method such as port-based method and payload based method had been taken place but unfortunately, these methods gained no success due to some problems.

Port based method assign port numbers to network traffic given by IANA. This technique was successful because many of the applications used fixed port numbers but some applications like P2P assign random port numbers to traffic so due to this problem port based method is not effective. The limitations of port based method has overcome by payload based but it also gains no popularity in network traffic classification because it can't opt encrypted traffic. So the limitations of both overcome by statistical method for generating flow based method. Now a days, this technique is achieving great success in the filed of network traffic. In my work also I am using statistical technique with the help of machine learning algorithm

Traffic classification [3] has extensively researched in recent years and many techniques have been proposed including Flow-Based technique, Host-Based technique and Graph-Based technique. Some of them were under research but many of them had achieved great success in the area of research. Now due to enhancement in today's internet new applications came and they become sophisticated so with respect to these emerging technologies many of the issue and challenges have been raised and faced by the researcher. Traffic classification has been extensively examined in recent years, as it is widely used in network management, design, security, advertising and research. In the past few years, the traffic classification techniques have been evolved along with the development of Internet protocols and applications, and many approaches have been investigated, proposed and developed. Nowadays, the ever increasing network bandwidth, the constantly sophisticated applications and the growth incentives to confuse classification systems to avoid filtering or blocking are among the reasons that traffic classification remains one of the hot areas in network research. Now a day, flow-based techniques are in great progress. In flow-based, we light on feature selection and in machine learning, we named it as variable selection or attribute selection. It contains many redundant or irrelevant data and it extract new features from their currently selected data set.

PREPROCESSING AND FILTERING OF NETWORK TRAFFIC CLASSIFICATION USING ADABOOST ALGORITHM

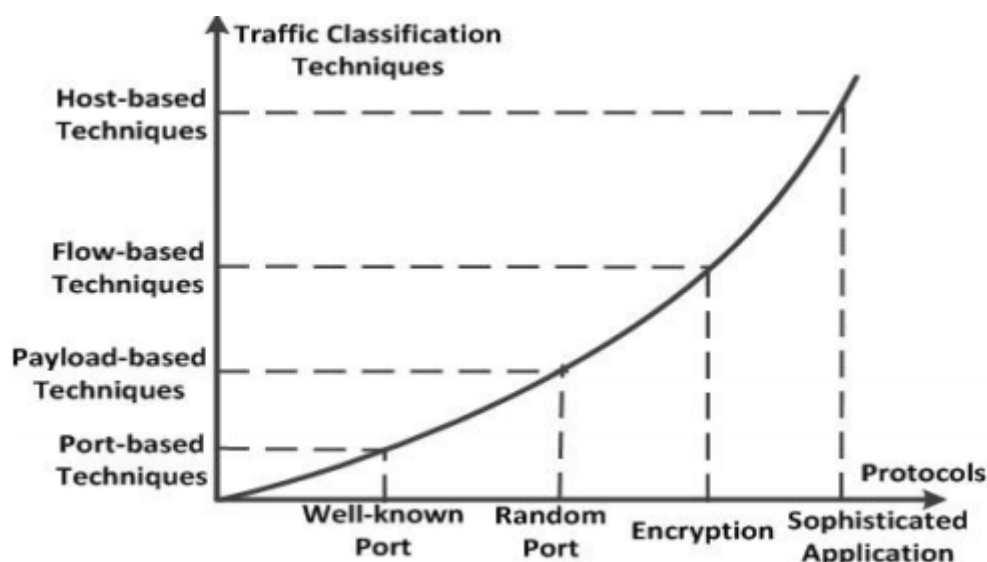


Figure 1.4: Evolutions of Protocols and Classification Techniques [3]

Previously, we use P2P, VoIP, and Bit Torrent and now new applications had taken place like Google Talk, Facebook, and Cloud computing, big data, Hadoop and yahoo messenger for extracting feature data set from traffic flow. Our work came under Statistical Classification, which

- Relies on attributes such as byte frequencies, packet size and the inter-arrival time.
- Fast technique.
- Used to detect unknown applications.

1.7.1 Traffic Classes

There are three traffic classes which are commonly used:

- i. Sensitive Traffic
- ii. Best-effort Traffic
- iii. Undesired Traffic

i. Sensitive Traffic

Sensitive traffic always delivers traffic on time and it includes VoIP, video conferencing and web browsing. In this kind of traffic, quality of service is guaranteed.

ii. Best-effort Traffic

Best effort traffic is also known as non-detrimental traffic and used in peer to peer and email applications. In this left over traffic is generated after sensitive traffic.

PREPROCESSING AND FILTERING OF NETWORK TRAFFIC CLASSIFICATION USING ADABOOST ALGORITHM

iii. Undesired Traffic

Undesired traffic is limited to delivery of spam messages as Skype.

1.7.2 Traffic types

Generally traffic is of four parts which includes

- i. Bursty Traffic
- ii. Interactive Traffic
- iii. Latency Sensitive Traffic
- iv. Non-real time Traffic

i. Bursty Traffic

As in video streaming the buffering occurs unevenly, the same pattern is applied in bursty traffic for transmitting data.

ii. Interactive Traffic

This kind of traffic takes place in web browsing and online purchasing means which take shorter time to request/response.

iii. Latency sensitive Traffic

This traffic is opposite to Bursty, means the delivery of data is at regular time intervals.

iv. Non-real time Traffic

As in email applications the request/response time does not matter, in this type of traffic also delivery of data transmission on time does not matter.

Machine learning [4] deals with data mining also whether it is an artificial intelligence tool and we do studies related to learning so that we can learn from known applications of data. It also deals with generalization and representation of data in data mining. Machine learning and data mining in most aspects are the same because machine learning predicts data based upon known properties whereas data mining is discovery of unknown properties.

This means how [1] computer can learn or improve their performance based on data. In machine the basic concept is to deal with computer so that they can automatically learn to capture pattern and from those achieved patterns can make intelligent decisions which is

PREPROCESSING AND FILTERING OF NETWORK TRAFFIC CLASSIFICATION USING ADABOOST ALGORITHM

based on data related to the pattern. Machine learning is a fast growing discipline. It is of further three types:

- Supervised machine learning
- Unsupervised machine learning
- Semi-supervised machine learning

In supervised learning, it deals with classification of data. It deals with classification of data. It always comes from training dataset which is of labeled. In unsupervised learning, deal with clustering and not labeled set. We use cluster to form classes in data. Semi-supervised learning falls under both labeled and unlabeled.

Machine learning also gave some of the algorithms as in data mining such as:

a) Supervised learning(Classification)

- Decision tree
- K-Nearest Neighbor
- Linear regression
- Naive Bayes
- Neural networks

b) Unsupervised learning (Clustering)

- K-Mean
- Expectancy Mean
- DBSCAN
- OPTICS

CHAPTER 2 LITERATURE REVIEW

To understand the concept of network traffic classification using machine learning techniques many research papers were consulted to grasp the problem domain and to understand the basic idea behind it. Some of the research papers are discussed here:

Shital S. et al(2014) Clustering of network traffic by online streaming focuses on implementing the new technique to cluster network data, traffic which eliminates the limitations in existing online clustering algorithm and in this paper we have to identify robustness and accuracy over large streamed data. [5] The work is divided into three parts: first part is to do packet sniffing and second part is to apply online clustering algorithm and third part shows the comparison with existing approaches. The packets are used with the help of RAH clustering algorithm. In this paper the whole work is done on packet sniffer. This packet sniffer shows all port no, length, IP address. This approach is less complex than existing approach. The limitations of new approach are that the clustering does not require predefined number of cluster and threshold values. If these values are unidentified then it will show wrong results. Due to comparison of new approaches with the existing approaches the streaming shows better results than old approach.

Jun Z. et al(2013) An untreated network traffic classification method with unknown flow detection. The new method is generated to detect the unknown applications for small supervised training set. [6] The proposed method uses correlation information for capturing the unknown flow. In this Erman gave a new method Erman semi-supervised method and use of nearest cluster based classifier is there for capturing the unknown flow of applications. The two traffic datasets are used: wide traffic dataset and ISP dataset and on the basis of three factors precision, accuracy and recall. For detection of unknown flow they had calculated true detection rate and false detection rate. The researcher had compared their proposed approach with the existing approaches. The calculated result is 0.65 for training clusters and 0.8 testing purities. These all applications uses TCP flow for recognizing the flow but for future UDP is being left. The proposed method uses two

PREPROCESSING AND FILTERING OF NETWORK TRAFFIC CLASSIFICATION USING ADABOOST ALGORITHM

techniques NCC (Nearest Cluster based Classifier) and Compound Classification to judge the bag of flows (BoFs).

Traffic classification technique is an essential tool for network and system security in the complex environments such as cloud computing based environment. The state-of-the art traffic classification methods aim to take the advantages of flow statistical features and machine learning techniques, however the classification performance is severely affected by limited supervised information and unknown applications. To achieve effective network traffic classification, we propose a new method to tackle the problem of unknown applications in the crucial situation of a small supervised training set. The proposed method possesses the superior capability of detecting unknown flows generated by unknown applications and utilizing the correlation information among real-world network traffic to boost the classification performance. A theoretical analysis is provided to confirm performance benefit of the proposed method. Moreover, the comprehensive performance evaluation conducted on two real-world network traffic datasets shows that the proposed scheme outperforms the existing methods in the critical network environment.

Armin D. et al(2013) Semi-Supervised Outlier Detection with only positive and unlabeled data based on Fuzzy clustering. The task of this paper is to find instances with the use of labeled examples. The outlier detection is issued in various applications like fraud and intrusion detection [7]. This method is based on extracting negative instances by k neural network and then using fuzzy clustering with both negative and positive examples. To address the problem of detection outliers using just few positive examples and unlabeled data, this paper proposes an accurate method based on fuzzy clustering. At first, enough reliable negative instances are extracted by means of kNN method and then fuzzy rough C-means clustering [7] will be applied to identify the other positive examples as outliers. The proposed approach SSODPU (semi-supervised outlier detection with positive and unlabeled data) is used for solving the problem. Yet, considering the fact that there is no other method dealing with this problem, we decided to compare the performance of our method against some state-of-the-art unsupervised algorithms. The methods are as follow: ABOD algorithm, Gaussian method, INFLO algorithm, kNN (K-Nearest Neighbor) algorithm, LDOF algorithm, LOF algorithm and LOOP algorithm. In future the use of only positive labeled approach for detecting outliers should be performed.

PREPROCESSING AND FILTERING OF NETWORK TRAFFIC CLASSIFICATION USING ADABOOST ALGORITHM

Yibo L. et al(2013) Traffic Classification: Issues and Challenges. In the past years, the techniques which are being evolved with internet protocol and applications, many new algorithms have been proposed and identified [3]. The issue and challenges are the ever emerging protocols and applications and ever increasing bandwidth. Traffic classification has been extensively examined in recent years, as it is widely used in network management, design, security, advertising and research. In the past few years, the traffic classification techniques have been evolved along with the development of Internet protocols and applications, and many approaches have been investigated, proposed and developed. Nowadays, the ever increasing network bandwidth, the constantly sophisticated applications and the growth incentives to confuse classification systems to avoid filtering or blocking are among the reasons that traffic classification remains one of the hot areas in network research. In this paper, we first attempt to present an analysis of the existing traffic classification techniques, and dwell on their issues and challenges, then address some recommendations that can improve the performance of traffic classification systems. The four are following techniques for network traffic:

- a) Port-based technique
- b) Payload-based technique
- c) Flow-based technique
- d) Host-based technique

All four techniques have their own issues and challenges. Many of new more techniques have been emerged to solve the problem. The new technique for identifying feature set is flow statistical feature selection and this is best suited with supervised and unsupervised for capturing known and unknown applications whose feature set is not known. Now a day, the flow based techniques are in progress and they have overcome the limitations of statistical methods by using algorithm.

Chien-Liang L. et al(2013) Semi-supervised Linear Discriminant Clustering (Semi-LDC). The algorithm proposed in this paper is k-means clustering and Linear Discriminant Analysis. The goal is to find a feature space where the k-mean can perform well in the space. To exploit the information brought by unlabeled examples, this paper proposes to use soft labels to denote the labels of unlabeled examples. The Semi-LDC uses the proposed algorithm, called constrained-PLSA, to estimate the soft labels of unlabeled examples. We use soft LDA with hard labels of labeled examples and soft labels of unlabeled examples to find a projection matrix. The clustering is then performed

PREPROCESSING AND FILTERING OF NETWORK TRAFFIC CLASSIFICATION USING ADABOOST ALGORITHM

in the new feature space. We conduct experiments on three data sets. The experimental results indicate that the proposed method can generally outperform other semi-supervised methods. We further discuss and analyze the influence of soft labels on classification performance by conducting experiments with different percentages of labeled examples. The finding shows that using soft labels can improve performance particularly when the number of available labeled examples is insufficient to train a robust and accurate model. Additionally, the proposed method can be viewed as a framework, since different soft label estimation methods can be used in the proposed method according to application requirements. In this both labeled and unlabeled examples are used to find projection matrix [8].

To compare hard labels with soft labels for generating an effective semi-supervised flow. Three data sets used in this paper for reducing the dimensionality. As k-mean is unsupervised machine learning and linear discriminant analysis is supervised machine learning, so by comparing both we proposed a new method name soft linear discriminant analysis. This paper also suggests that the proposed method can benefit from soft label representation particularly when only a few labeled examples are available.

Tamilkili *et al*(2013) A survey on recent traffic classification techniques using machine learning methods need to classify statistical feature. To calculate flow five tuples are needed. The novel unsupervised approach collects dataset or capture packet as input and then based on flow type, based on payload do training and testing, on training the pre-processing technique is used to know the flow type. And finally generate [9] the output. To use support vector machine based traffic classifier the data set is passed through testing flow as a single class then add or check for boundaries: 0(unknown), one (assigned) and less than one (multiclass).

Jamuna A. *et al*(2013) Efficient flow based network traffic classification using machine [10] learning compares the traditional approach with C4.5, naive bayes, nearest neighbor and decision tree. Then two methods are used full feature selection and reduced feature set for classification. But from two only one give best results i.e. given by reduced feature set. In this paper we classify that real-time traffic classifiers will work under constraints, which limit the number and type of features that can be calculated. In comparing the classification speed C4.5 is best.

PREPROCESSING AND FILTERING OF NETWORK TRAFFIC CLASSIFICATION USING ADABOOST ALGORITHM

Yu *et al*(2012) In internet traffic clustering with constraints paper, the limitations of traditional-based and payload-based traffic classification has been removed by suggesting clustering method. In past, it is [12] shown that unsupervised learning is best to capture unknown applications. So now we focus on correlation and to remove the constraints we develop a k-mean algorithm. In particular, previous studies have shown that the unsupervised clustering approach is both accurate and capable of discovering previously unknown application classes. In this paper, we explore the utility of side information in the process of traffic clustering. Specifically, we focus on the flow correlation information that can be efficiently extracted from packet headers and expressed as instance-level constraints, which indicate that particular sets of flows are using the same application and thus should be put into the same cluster. To incorporate the constraints, we propose a modified constrained K-Means algorithm. A variety of real-world traffic traces are used to show that the constraints are widely available. The experimental results indicate that the constrained approach not only improves the quality of the resulted clusters, but also speeds up the convergence of the clustering process. The two real world datasets are used like Keio, Wide traces. In the problem domain of network traffic classification, we observe that there exists some side information telling partial correlations across flows in addition to the flow descriptions themselves. For example, concurrent flows connecting to the same destination IP address at the same TCP port are typically using the same network application. Extra information of this kind could be derived from domain knowledge of networking without knowing the actual class labels of flows. In general, given an unlabeled traffic data set, we could extract not only the statistical attributes describing each flow, but also a large amount of constraints representing the correlation information, which could be useful in clustering the flows. In the result not only performance of traffic clustering gets improved but also speed up the convergence of clustering process.

Tomasz *et al*(2012) In this paper a method for classification of network traffic based on c5.0 machine learning algorithm has been introduced by Tomas. [13] To improve the performance of network is a challenging task. QoS and Multi-hop networks are improving the quality of current network traffic. To remove the limitations of previous methods we proposed a c5.0 machine learning algorithm on the basis of statistical technique. In this we classify the dataset for applications like Skype, FTP, Torrent, Web browser etc. To evaluate the result we use decision tree to classify the test cases. In this

PREPROCESSING AND FILTERING OF NETWORK TRAFFIC CLASSIFICATION USING ADABOOST ALGORITHM

research both training and test datasets were disjoint, but collected from the same users. As the next step we consider involving numerous users to assess the accuracy using datasets obtained from different network.

Thuy *et al*(2012) In timely and continuous machine learning based classification for interactive IP traffic, for classifying the traffic machine learning techniques used only for analysis of first few packets. QoS monitor well and quick the flow and the statistical feature derived from sub-flows and then search for sub-flows selection. [14] In this training and testing dataset derive from the flow. The flow is searched from client-to-server and server-to-client directions. We propose to achieve this by using statistics derived from sub-flows—a small number of most recent packets taken at any point in a flow's lifetime. Then, the ML classifier must be trained on a set of sub-flows, and we investigate different sub-flow selection strategies. We also propose to augment training datasets so that classification accuracy is maintained even when a classifier mixes up client-to-server and server-to-client directions for applications exhibiting asymmetric traffic characteristics. Two best approaches Naive Bayes and C4.5 Decision Tree machine learning algorithms has been used. The three factors show a great progress within less than 1s. Applications used to detect flow are FPS game traffic and VOIP traffic. First we compare full flow and then illustrate the benefit of SSP-ACT.

The results showed that using a sub-flow size of $N=25$ packets, the Naive Bayes classifier achieved 98.9% Recall and 87% Precision when classifying ET traffic and 99.6% Recall and 95.4% Precision when classifying VOIP traffic. The C4.5 Decision Tree classifier achieved 99.3% Recall and 97% Precision when classifying ET traffic and 95.7% Recall and 99.2% Precision when classifying VOIP traffic. Both classifiers maintained their performance and missed packets using unsupervised approach.

T. Karasgianni *et al*(2012) In 2012, machine learning based network traffic classification gave applications regarding QoS, accounting and intrusion detection. Previously, we had traditional methods like port match method and after having enhancements in internet we shifted to payload analysis. This analysis is not so popular because payload is unable to encrypt traffic so researchers moved to statistical feature based approach. In this paper, [15] we focused on two techniques of machine learning: Supervised and Unsupervised. We focused on to identify the flow classification using statistical approach for calculating the better performance of flow. Paxson V. Gave

PREPROCESSING AND FILTERING OF NETWORK TRAFFIC CLASSIFICATION USING ADABOOST ALGORITHM

relationship between flow category and flow statistical for the first time and analyze the connection between them. The concept of machine learning is going to be useful for reducing the workload of traffic.

The machine learning techniques are divided into four categories:

- a) Classification
- b) Clustering
- c) Association mining
- d) Numerical prediction

Bin *et al*(2012) We can perform classification on various applications like QoS, intrusion detection and accounting. In this paper, port-match is proposed to be a best method for classification of network traffic on machine learning. The classification of network applications is essential to numerous network activities, including quality of service, accounting, and intrusion detection. Originally, port-match was regarded as the most popular and effective method. After that, with the booming of new Internet businesses, researchers shifted to the use of payload analysis. However, the payload analysis is unable to process encrypted traffic, which motivated scientists to develop more general and effective solutions. To do so, traffic classification has been of great concern to academia and industry and gradually formed a relatively independent area of research. Machine Learning based classification has stood out among multitude research findings. This paper aims to provide an overview of recent advances in such study area. We focus on how to divide machine learning based classification into two categories: supervised and clustering. We also present the algorithms of creating specific feature sets and classification models such as Genetic Algorithm and Bayesian algorithm. Finally, we compare the efficiency of these algorithms and discuss the future direction of machine learning based classification. Due to growth in internet, researchers started using payload analysis for classifying the network traffic.

To overcome the limitations of both we gave a machine learning technique [16] namely, Supervised and Unsupervised algorithm such as Genetic and Bayesian. The researcher proposed a method on his name called Erman's method semi-supervised learning to limiting the clustering techniques. Finally compare new methods with old methods. Training and Testing of datasets are also classified so that it can guarantee high identification accuracy.

PREPROCESSING AND FILTERING OF NETWORK TRAFFIC CLASSIFICATION USING ADABOOST ALGORITHM

Guowu X. *et al*(2012) The Subflow: Towards practical flow-level traffic classification demonstrates that the statistical methods have not shown greater progress on machine learning so new method of flow-level is proposed by many of the researchers. [17] In flow-level we constitute packet sizes and Interarrival times. The key point of this paper is we can easily identification of the traffic for each application. In this paper the use of subspace clustering is proposed by the author and subspace clustering is used for profiling of various applications by eliminating redundant features. This application shows great impact on accuracy. The advantages of this application over previous methods are that our proposed approach is flexible and having adaptive nature to change; by applying these methods bootstrapping becomes easier and practical. The algorithm used in this paper is Bayesian Network Classifier for improving the accuracy of proposed methods. In the result it was shown that the accuracy increases by an average higher than 95% for detection for new applications are successful.

Jian M. *et al*(2010) Study on the process of network traffic classification using supervised learning. Classification of network traffic is the essential step for many researches. Machine learning approaches are best approaches in the field of network traffic classification. Many statistical techniques have been proposed for addressing the problem of traffic classification. In this paper a new approach is determined which is based on logistic regression.

In this approach, we can automatically select best feature from the rest of traffic to separate the flow of network. [18] It overcomes many of the weaknesses of five states of art techniques. It uses the applications like HTTP streaming and P2P. Many more applications have also observed by researcher like e-Donkey and Gnutella. They achieve great progress in terms of performance in flow-based statistical feature set. The workflow has been taken of SunYatSen University. The workflow of this paper is logically correct.

Finamore *et al*(2010) Network traffic classification is rising day by day from past few years so to overcome the traffic flow we analyzed two techniques, supervised classification algorithms and unsupervised clustering algorithms. Recently the work on methods used for statistical has [19] not been solved by the researchers and the work is going on by improving the performance of flow. In unsupervised traffic classification, it is difficult to build a best classifier with the help of clustering and moreover without knowing the real traffic classes. The supervised classifier can be divided into two

PREPROCESSING AND FILTERING OF NETWORK TRAFFIC CLASSIFICATION USING ADABOOST ALGORITHM

categories: parameterized and nonparametrized classifiers. Parametric classifiers use algorithms such as C4.5 decision trees, Bayesian networks, neural networks and nonparametric classifiers such as K-Nearest Neighbor. In this paper, the nonparametric approach had given best solution regarding the performance of correlation information by opting Nearest Neighbor algorithm. This solution arises on the basis of both empirical and theoretical perspectives. In this new real-world methods and datasets have been introduced to show the performance under few training samples. The same proposed work we can opt for semi-supervised also as a future work so that an accurate result can be produced by the information.

Wang R. *et al*(2010) In this paper, a new re-sampling method for network traffic classification using SML is being introduced by the author. By resolving the limitations of old techniques in machine learning we had gone through flow based feature technique which comes under statistical technique. In this paper we explore the information about traffic clustering with constraints using correlation information, by using K-mean algorithm. When we gone through this paper we came to know that unsupervised is the best technique for traffic classification. This paper showed us that not only the convergence speed got improved but also the quality of clusters. The new re-sampling method is proposed tuning sampling to solve the problem of [20] data skew in network traffic. And compared our methods with uniform and stratified sampling which has been proposed earlier. The result came in between the both that accuracy varies.

To evaluate the best result we use five real time traffic data sets and by using unsupervised we found 10% improvement in our data set. So this will improve the performance of clustering by feature discretization. If we want more accurate data flow then apply C4.5 classifiers for improving the performance of clustering with constraints.

Shrivastav *et al*(2010) Network traffic classification using semi-supervised approach is analyzed and implemented. This takes only few labeled flow and many labeled flow. The approach is divided into two parts:

- a) Clustering
- b) Classification

Clustering is [21] used to partition training dataset. After making clusters, classification is performed in which labeled data are used for assigning class labels to the clusters. A KDD cup 1999 data set is being taken for testing this approach. The tested results are

PREPROCESSING AND FILTERING OF NETWORK TRAFFIC CLASSIFICATION USING ADABOOST ALGORITHM

compared with support vector machine based classifier for results. The compared result will show the final result and then use this result to classify semi-supervised approach.

Bregni *et al*(2010) Due to Rottondi, flow of packets is generated to classify the traffic by the use of many newer techniques like Shallow Packet Inspection. It examines only outermost header of packets flow and it uses both synthetic data and real time traffic data. It deploys UDP and TCP port numbers for deep packet inspection for overcome the issues generated in this paper by using packet Interarrival time for [22] classification.

It also compares many of the algorithms to calculate the accuracy and composition of training data. We used the statistical feature called index of variability in order to compare or integrate between different traffic classes. A captured wired traffic approach is used to map patterns with network traffic so that the useful information can be used to predict future traffic analyses. This paper will capture live traffic and analyze the statistical values. In this paper for capturing network traffic wireshark tool is used. The traffic will constitute on very small packet size. The whole scenario is based on confidence interval. The bandwidth fluctuates and the average number of bytes per packet for TCP was 652 bytes and UDP was 594 bytes.

Dong *et al*(2009) The study of network Traffic Identification based on Machine Learning Algorithm. In this paper the machine learning algorithm is now compared with traffic classification for better performance and accurate results. Network traffic identification is one of the hot research fields for network management and network security; machine learning is an important method during the network traffic identification research. This paper describes the current situation and common methods of network traffic identification, at the same time this paper also states the currently popular Machine learning methods. We compared and evaluated the supervised and unsupervised classification and clustering algorithms, the experiment results show that feature selection algorithm has great effect on supervised machine learning and DBSCAN algorithm which belongs to unsupervised clustering algorithm has great potential in precision. Both machine learning techniques Supervised and Unsupervised had great [23] effect on feature selection and DBSCAN algorithm for precision. The current network traffic methods such as:

- a) Port-based method
- b) Deep packet inspection (DPI)

PREPROCESSING AND FILTERING OF NETWORK TRAFFIC CLASSIFICATION USING ADABOOST ALGORITHM

- c) Host behavior method
- d) Flow-based method on machine learning.

The algorithm used in this paper is Naive Bayes, SVM (Support Vector Machine). The two datasets are used in this paper are MOORE-SET and AUCKLAND-SET. The two real world data sets show that the accuracy and precision factor used in paper shows great progress by comparing the previous methods. The results show that the supervised machine learning is influenced by feature selection algorithm. This algorithm improves accuracy and time efficiency.

This paper has studied and analyzed the machine learning algorithm for network traffic identification and mainly studied unsupervised and supervised machine learning. Through experiment on the classification algorithm of two different datasets, comparing the classical unsupervised and supervised algorithm; the experiment result show that the supervised machine learning is greatly influenced by feature selection algorithm. The suitable feature selection algorithm can improve the accuracy and time efficiency of classification algorithm. By comparing several unsupervised machine learning algorithm (cluster algorithm), results show that DBSCAN algorithm has great potential and has more advantage than other two kinds of algorithms in precision, besides the modeling time is between the K-Means method and DBSCAN method.

Tavallaee *et al*(2009) Online classification of network flows is very challenging and stills an issue to be solved due to increase in new applications and traffic encryption. In this we apply signature based method to online flow. [24] Due to limitations in previous methods we proposed a hybrid approach. As traditional method is used to assign port number and now not used for modeling. We can not only obtain signature from unencrypted traffic, but can also be extracted from the encrypted traffic. To overcome the shortcomings of traditional method, researchers have proposed new approaches based on either the content of the payload or the statistical feature such as connection duration. And also apply a learning approach to classify the unknown portion based on the network statistical features. All the selected features have two important characteristic:

- They have shown to be effective in distinguishing different applications.
- They can be calculated in real time and impose no delay to the classifier.

In order to process the packets, classify them into flows and extract the features, we used a commercial network security management tool. The methods are implemented by

PREPROCESSING AND FILTERING OF NETWORK TRAFFIC CLASSIFICATION USING ADABOOST ALGORITHM

WEKA on the basis of accuracy, learning time and classification time. The main drawback of this approach is that it fails to find the novel unknown applications because of applying a supervised classification method.

Shijun H. et al(2009) A statistical feature based approach to internet traffic classification using Machine Learning. This classification technique is focusing on modelling attributes and features of data flow for the identification of flow in traffic. We do compare this new statistical technique with traditional based techniques. In this also supervised algorithm K Nearest Neighbor is adopted for the sake of efficiency and it is easy to calculate the operational background of the flow and we had [25] got the better achievement in flow based approach. We are having two types of phases: training and testing phase. The traffic data for known applications had been flow from files to the feature set of traffic for training phase. In this we do mark known applications for each and every flow. After combining these flow based feature set we got a perfect trained set to modify the machine learning classifier model. In this paper, I had gone through K-Nearest Neighbor algorithm over statistical flow with the improvement of 90% and this achieves a great access to traditional flow. At previous papers only 50-70% achievement in known applications of flow.

In this paper we have demonstrated a statistical-feature based approach to classify Internet traffic using supervised Machine Learning (ML). No more information than headers in IP and transport layers is need for classification. Taking no account of extra-added Instant Messaging (IM) flows (mainly for testing), the classifier model performs well in traffic classification with above 90% flow accuracy, which shows no inferiority compared with that of similar research work like Roughan et al. 's work [25]. Moreover, the simplified statistical features and the easy-to-use k-Nearest Neighbor (KNN) estimator result in lower space and time complexity, which is worth mentioning.

Xu T. et al(2008) In this paper, Traffic classification is an emerging network management which is becoming popular for managing the network and measurement tasks. Following are the two main contributions: First approach is novel integrated dynamic online traffic classification rule [26] or schema called data stream based traffic classification. Second approach is mining algorithm called very fast decision tree. And this decision tree is implemented in data stream based traffic classification. This paper mark several advantages:

PREPROCESSING AND FILTERING OF NETWORK TRAFFIC CLASSIFICATION USING ADABOOST ALGORITHM

- This is build to handle multiple, rapid network traffic.
- It is real time memory efficient method.
- The training phase simultaneously works with testing phase.

The results show that first approach is better than second due to small cost with high accuracy of above 98%.

Arthur et al(2007) An survey on internet traffic identification and classification. The author suggest that due to hilarious enhancement in the growth of internet, users, increase in speed and many new applications had affect the work of Internet Service Provider (ISP) and network administrators. In this paper we will face two problems that firstly separate packet based and flow based from network traffic classification. Secondly, the technique to be used for solving these problems is [27] signature-matching, inference. On the basis of two factors accuracy and completeness the results have been conducted and final result have been conducted from following applications like HTTP, HTTPS, SMTP are approx. 99%. This survey explains the main problems in the field of IP traffic analysis and focuses on application detection. First, it separates traffic analysis into packet-based and flow-based categories and details the advantages and problems of each approach. Second, this work cites the techniques for traffic analysis available in the literature, along with the analyses performed by the authors. These techniques include signature-matching, sampling and inference. Third, this work shows the trends in application behavior analyses and cites important and recent references in the subject. Lastly, this survey enlists the open topics of research by explaining the questions that are still open in traffic analysis and application detection and makes some final remarks.

In conclusion we came to know that the signature-matching technique for payload analysis is having some legal problems and inference method only identify some applications correctly. Some claim that these techniques were time-consuming and some claim the loss of information but some claim for better achievement in high efficiency and precision.

Li et al(2007) Accurate classification of the internet traffic based on the SVM model recognizes the features by pattern evaluation for classifying the unknown flows. We gather packet from only one direction according to 4 tuples [28] (source IP, destination IP, source port, destination port). Then merge these tuples into bidirectional flow for

PREPROCESSING AND FILTERING OF NETWORK TRAFFIC CLASSIFICATION USING ADABOOST ALGORITHM

checking the payload. The SVM method is developed in this paper to classify the internet flows. Optimization is carried out to obtain the best combination of parameters. For regular traffic samples with biased prior probability, we achieve a accuracy of 99.4% and for unbiased 96.9%.

CHAPTER 3 PRESENT WORK

3.1 Problem Formulation

Due to increase in internet growth as well as emerging protocols and applications, the internet bandwidth has to be promoted. The more increase in internet, more bandwidth is required and more the use of security threats. Traffic classification deals with traditional method and statistical methods. In traditional methods port-based and payload-based methods come which now a day is not used. To overcome their problems we came across flow-based statistical method and the issues or challenges which is generating are of selection of flow statistical features and uniqueness proof of flow statistical features. So from these issues we went through many research papers and recognize the problem of preprocessing and filtering of network traffic classification using adaboost algorithm.

We proposed from previous approaches that the use of Naive Bayes with consistency feature selection gives best efficiency of 95% but without feature set gets 94%. So to get more efficient result we compare with bagging and boosting and classify the dataset using Consistency Feature Selection. In previous approach algorithm used are Naive Bayes, C 4.5, KNN, Bayesian network. We also compared our new results with previous results of C4.5 and we get the accuracy of 91%. So to get better result we perform mining as hybrid classifier by comparing the algorithm like Bagging and Boosting and by comparing these two algorithms we achieved the accuracy of 97% which is more efficient than previous approaches and this efficient result got with the help of parameters like Precision, Recall, TP rate, FP rate etc.

3.2 Objectives

After considering the whole scenario, our aim is to solve the problem with the help of different objectives such as:

- To do collection of raw data from KDD dataset.
- To perform preprocessing and filtering of data by extracting best feature set.
- To use Consistency Feature Selection (CFS) method, feature selection is done.
- To do classification of filtered data with the help of Naïve Bayes.

PREPROCESSING AND FILTERING OF NETWORK TRAFFIC CLASSIFICATION USING ADABOOST ALGORITHM

- To perform mining using Bagging and Boosting of hybrid to get best instance.
- To classify the data, compare the results and analyze performance of the proposed algorithm on the dataset.
- To use various parameters like precision, recall and accuracy etc.

3.3 Research Methodology

Broadly there are two types of research methodologies one is Quantitative approach and other one is Qualitative approach. Quantitative approach is like result, throughput and Qualitative approach is used for gathering and analyzing data. Research methodology is a way to solve the problem in a step by step process. We adopt many steps before starting of research and it is necessary to know the methods and techniques. So we will be using quantitative approach for solving our problem so that we can get best result and throughput. We don't only talk of research but we had to consider the logic behind those methods for solving our problem. The Qualitative approach is done in literature review and from there we have come across many problems and from that finally one problem originates. Now we have taken that problem and the methods behind that have also been implemented.

Now the trend of internet is changing day by day so new applications have been proposed and more security threats also been recognized. Machine learning [4] deals with data mining also whether it is an artificial intelligence tool and we do studies related to learning so that we can learn from known applications of data. It also deals with generalization and representation of data in data mining. Machine learning and data mining in most aspects are the same because machine learning predicts data based upon known properties whereas data mining is discovery of unknown properties. In our report we had gone through various statistical feature based classification methods and came to know that it comes under flow-based technique. For solving these types of problem we need best tool and methods. My work can be implemented by using JAVA tool because it is best suited for machine learning algorithms. In base paper NN classifier is used. Machine learning algorithms are supervised (classification) for labelled set of data, unsupervised (Clustering) and semi-supervised algorithms. In the proposed work Network traffic is captured by known applications and unknown applications using various unsupervised and semi supervised machine learning techniques but we solve our problem by capturing unknown applications using supervised learning. This is done with

PREPROCESSING AND FILTERING OF NETWORK TRAFFIC CLASSIFICATION USING ADABOOST ALGORITHM

the help of machine learning because it performs better result in network traffic. The traffic log will capture unknown flow at the start and end of unknown session.

Traffic classification is an automatic process for generating traffic according to different parameters into a number of traffic classes. In network traffic data is wrapped in packets and each packet contains control information and user data.

Traffic classification has been emerging day by day from past few years. It is widely used in networks, including intrusion detection, security and research. Many of the protocols and proposed applications have been investigated and developed by using machine learning algorithm. The growth of today's internet is giving popularity to the needs of research. Many emerging tools have been proposed by many of the researchers. This is the hot area of research in emerging field of networks. With the use of machine learning algorithms which have been used in data mining fields taken in convergence. Data mining is knowledge discovery of data. It includes many features like clustering, classification, outlier detection, noise detection etc. But in our field we had combined the research area with networking so that an better or optimal solution could be found. In past year research fields, traditional method such as port-based method and payload based method had been taken place but unfortunately, these methods gained no success due to some problems.

Port based method assign port numbers to network traffic given by IANA. This technique was successful because many of the applications used fixed port numbers but some applications like P2P assign random port numbers to traffic so due to this problem port based method is not effective. The limitations of port based method has overcome by payload based but it also gains no popularity in network traffic classification because it can't opt encrypted traffic. So the limitations of both overcome by statistical method for generating flow based method. Now a days, this technique is achieving great success in the filed of network traffic. In our work we are also using statistical technique with the help of machine learning algorithm

Traffic classification [3] has extensively researched in recent years and many techniques have been proposed including Flow-Based technique, Host-Based technique and Graph-Based technique. Some of them were under research but many of them had achieved great success in the area of research. Now due to enhancement in today's internet new applications came and they become sophisticated so with respect to these emerging technologies many of the issue and challenges have been raised and faced by the researcher.

PREPROCESSING AND FILTERING OF NETWORK TRAFFIC CLASSIFICATION USING ADABOOST ALGORITHM

To start up with programming part we had use Netbeans Java. To add any new project on java we click on File menu and from there we select new Project tab and then add new project. If any error occur during adding the project then it means we don't have sql connector file and for adding sql connector file we do right click on project name and select its properties. After selecting properties the new window appear on the screen, then go to categories and select Libraries. After selecting Libraries remove the old files and then click on Add JAR/Folder then browse the connector file and open it and then add it. That is how we add connector files to Java.

To solve our problem we had taken a KDD dataset which is the collection of raw data namely as Final Dataset.arff in which includes two attributes as src_bytes and dest_host_srv_count and had two classes as normal for known applications and anomaly for unknown applications. After getting the dataset we applied preprocessing and filtering of data by selecting the best features from dataset using consistency feature selection method and then the classification of filtered data is done. Then perform mining using Naïve Bayes with feature selection as in figure 3.1 and correctly classified instances is 95%, incorrectly classified instances is 4%. As in previous approach the use of Naïve Bayes without feature selection, correctly classified instances is 94% and incorrectly classified instances is 5%. In C4.5 mining performed correctly classified instances as 91% and incorrectly classified instances is 8%. Now perform mining with the help of Bagging and Boosting to get more efficient result than naïve Bayes. In Bagging algorithm the classifier used is RP Tree which is the form of Decision Tree. But instead of RP Tree we are using AdaBoost classifier to make our result more accurate. In the result we can clearly see the difference in correctly classified traffic upto 97% and incorrectly classified traffic upto 2% which is the major difference in results and this makes our work more accurate and efficient. Then at the end classify the data and evaluate the results and analyze various parameters like Accuracy, Precision, Recall etc.

- Accuracy: Proportion of total number of predictions that are correctly classified in class C.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

where TP is True Positive

TN is True Negative

FP is False Positive

FN is False Negative

PREPROCESSING AND FILTERING OF NETWORK TRAFFIC CLASSIFICATION USING ADABOOST ALGORITHM

- Precision: Percentage of selected documents that are correctly classified in class C out of all documents in class C.

$$\text{Precision} = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)} + \text{False Positive (FP)}}$$

- Recall: Percentage of correct documents that are selected in class C from the entire document actually belonging to class C.

$$\text{Recall} = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)} + \text{False Negative (FN)}}$$

- Confusion Matrix: Also known as contingency table or error matrix in supervised learning and in unsupervised learning is called matching matrix. In confusion matrix, ROC is used to plot graph.

		Predicted	
		Negative	Positive
Actual	Negative	a	b
	Positive	c	d

Example:

A	b
108 (TP)	9 (FP)
1 (FN)	95 (TN)

$$a = 108 + 95 = 203$$

$$b = 9 + 1 = 10$$

where a is classified as normal and b is classified as anomaly

PREPROCESSING AND FILTERING OF NETWORK TRAFFIC CLASSIFICATION USING ADABOOST ALGORITHM

- True Positive (TP) and False Positive (FP) Rate: For multiple comparisons TP and FP is used and it is a type of error. TP is also called Sensitivity as if a person has a disease how often will the test be positive is referred to as true positive rate. FP is an error in a test result indicates presence of a condition.

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

- F-measure: A measure that combines precision and recall.

$$\text{F-measure} = \frac{2 * (\text{Precision} * \text{recall})}{\text{Precision} + \text{Recall}}$$

3.3.1 Flowchart of previous and new approach

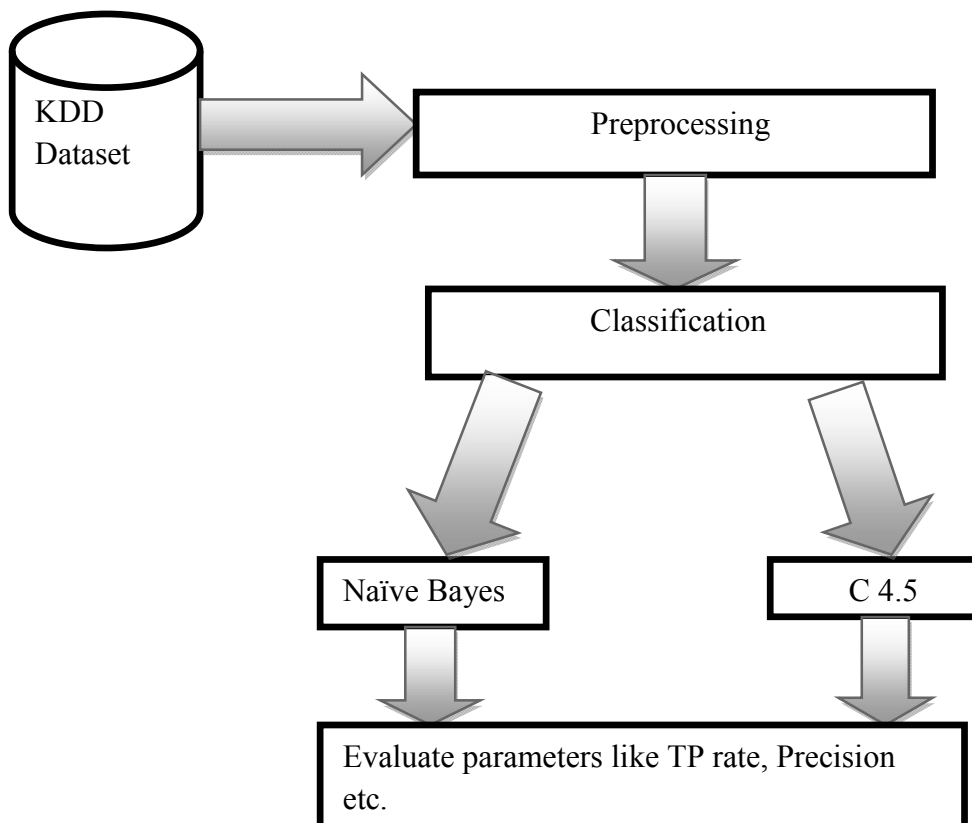


Figure 3.1: Flowchart for Previous approach

PREPROCESSING AND FILTERING OF NETWORK TRAFFIC CLASSIFICATION USING ADABOOST ALGORITHM

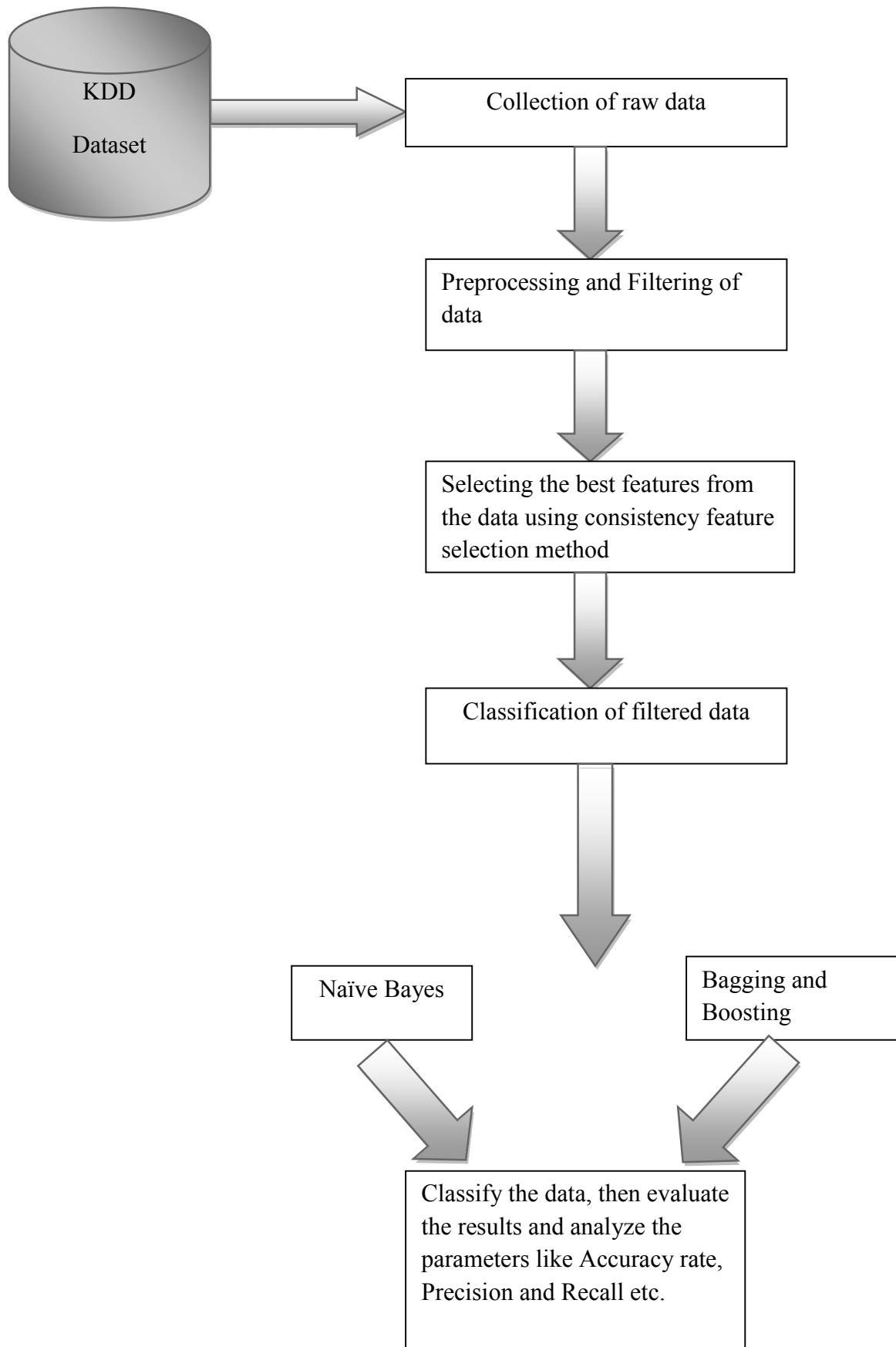


Figure 3.2: Flowchart for New Approach

3.3.2 Algorithm Used

Algorithms used in our approach are:

- Bagging and Boosting

Bagging also referred to as bootstrap aggregation and it is a meta-learning technique. In bagging, the classifiers are divided into different parts to training class and use the majority vote to get final result. The final result is used to reduce the variance associated with prediction and thereby improve the prediction process. Unlike in boosting we convert weak learners into strong learners. In this weak learners are those who are slightly correlated to classification and strong learners are those who are fully correlated to classification. We proposed these two algorithms to get more accurate result than Naive Bayes. In bagging classifier used is RP Tree which is the form of decision tree and due to or proposed work we are using Adaboost as a classifier instead of RP tree to combine with boosting algorithm.

- Adaboost

Is known as Adaptive Boosting, these are weak learners and makes the use of decision stump. Adaboost is always limited to noisy data and outliers. To each instance assigned an equal weight. If the previous classifier is not updated then no weight is assigned to next classifier. If there is error below 0 and above 0.5 then model is terminated. If the error is between 0.1 to 0.4 then model is successfully computed.

Algorithm:

- i. Assign equal weight to each row of instance used in dataset.
- ii. Then perform iterations unless
If error $e = 0$ or $e \geq 0.5$
then
Model is terminated
Else
If error $e = 0.1$ to 0.4
then
Model is successfully computed
- iii. Multiply assigned weight to each instance by $e(1-e)$
- iv. Classification
if weight = 0 then

PREPROCESSING AND FILTERING OF NETWORK TRAFFIC CLASSIFICATION USING ADABOOST ALGORITHM

```
add -log e/ (1-e)
return instance with highest weight
```

- Naive Bayes

Naive bayes are statistical classifiers. They can predict class labels such as the probability to which tuple belongs to a particular class and also called class-conditional independence. It is an independent feature model and straight forward classifier. Naive bayes can work without receiving Bayesian probability. Naive bayes also work well with machine learning and is a statistical classification and is used by the researcher now a days.

$$P(h/c) = \frac{P(c/h) P(h)}{P(c)}$$

Where P (h): independent probability of h as prior probability

P (c): independent probability of c

P (c/h): conditional probability of c given h as likelihood

P (h/c): conditional probability of h given c as posterior probability [2]

3.3.3 Tool Used

The tools used are:

3.3.3.1 Java Netbeans IDE

To start up with programming part we had use Netbeans Java. To add any new project on java we click on File menu and from there we select new Project tab and then add new project. If any error occur during adding the project then it means we don't have sql connector file and for adding sql connector file we do right click on project name and select its properties. After selecting properties the new window appear on the screen, then go to categories and select Libraries. After selecting Libraries remove the old files and then click on Add JAR/Folder then browse the connector file and open it and then add it. That is how we add connector files to Java.

Java is a programming [29] language and computing platform, first released by Sun Microsystems in 1995. There are lot of applications and websites that will work unless you have java installed, and more are created every day. Java is fast, secure and reliable. Form laptop to datacenter, game console to scientific supercomputer, cellphone to internet, we can say java is everywhere. [30] A high levelled programming language is spread by Sun Microsystem. Java was originally called OAK, and was developed for

PREPROCESSING AND FILTERING OF NETWORK TRAFFIC CLASSIFICATION USING ADABOOST ALGORITHM

handheld devices. OAK was failed so in 1995 Sun renamed to java and improves the language to take benefit of the World Wide Web.

Java is an [30] object oriented language which resembles like C++ but simplified to remove language characteristics that cause common programming mistakes. Java source code files are compiled into a format called byte code, which can then be implemented by a java interpreter. Compiled java code can run on most computers because java interpreter and runtime environment known as java virtual machines. Byte code can then converted into machine language by a just-in-time compiler.

Java is a universal programming language with a number of characteristics that make the language well acceptable for use on the World Wide Web. Small java applications are called java applets and can be downloaded from a web server and run on your computer by a java compatible code.

3.3.3.2 MySQL

It is an open source relational SQL database management system (RDBMS) and also uses triggers, cursors, view, schema etc. Unlike SQL database, it does not uses full SQL functions for implementation. It is used to create,update and delete rows in table or to delete full table from dataset. To startup with MySQL command client some commands are used

```
mysql> show tables;
```

```
database changed
```

```
mysql> show tables;
```

```
tables will appear
```

```
mysql> select * from table name;
```

```
mysql> delete from table name where filename = ' ';
```

CHAPTER 4

RESULTS AND DISCUSSIONS

Java is an [30] object oriented language which resembles like C++ but simplified to remove language characteristics that cause common programming mistakes. Java source code files are compiled into a format called byte code, which can then be implemented by a java interpreter. Compiled java code can run on most computers because java interpreter and runtime environment known as java virtual machines. Byte code can then converted into machine language by a just-in-time compiler.

Java is a universal programming language with a number of characteristics that make the language well acceptable for use on the World Wide Web. Small java applications are called java applets and can be downloaded from a web server and run on your computer by a java compatible code.

To start up with programming part we had use Netbeans IDE 8.0 as in figure 4.1. To add any new project on java we click on File menu and from there we select new Project tab and then add new project. If any error occur during adding the project then it means we don't have sql connector file and for adding sql connector file we do right click on project name and select its properties. After selecting properties the new window appear on the screen, then go to categories and select Libraries. After selecting Libraries remove the old files and then click on Add JAR/Folder then browse the connector file and open it and then add it or we can go to our project library and right click on it then some options will appear on the screen, from their also we can choose Add JAR/Folder.

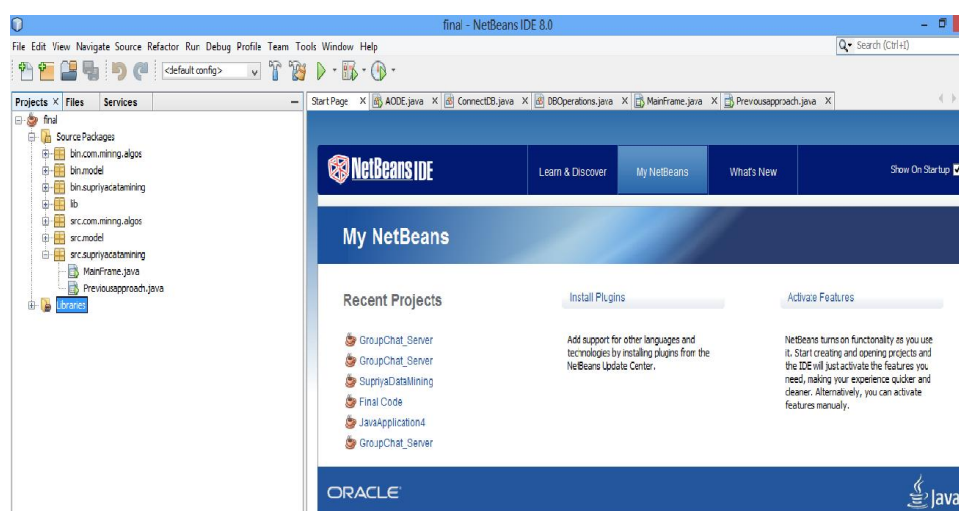


Figure 4.1: Netbeans IDE 8.0

PREPROCESSING AND FILTERING OF NETWORK TRAFFIC CLASSIFICATION USING ADABOOST ALGORITHM

First of all we go to Final (project name) then by clicking on to Mainframe.java, right clicking on it run file and an dialog box appears on the screen as in Figure 4.2. Then we had to choose dataset by clicking on Browse button and by specifying our dataset location click on upload then the name of dataset will appear in the output window.

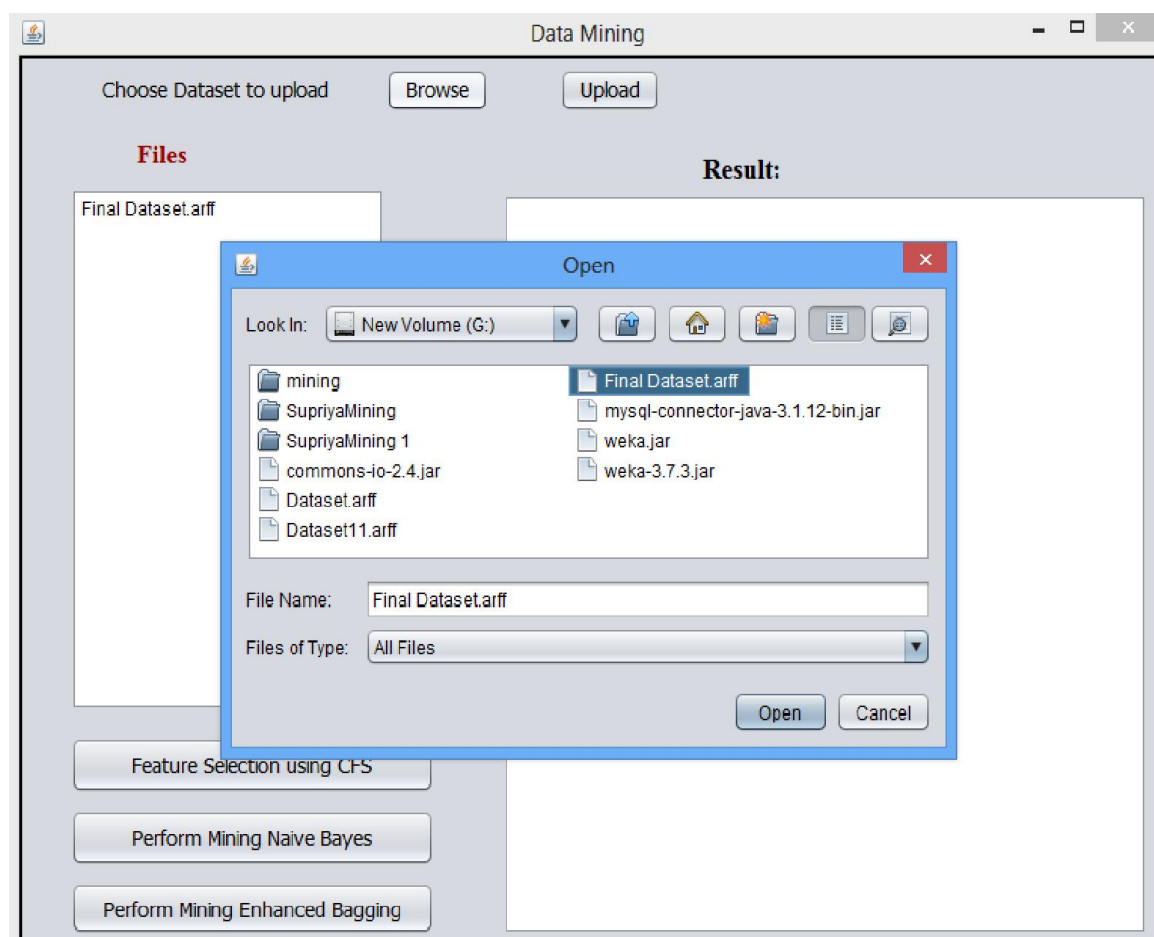


Figure 4.2: Choose Dataset to Upload

After uploading dataset, select the respective dataset and click on feature selection using CFS as in figure 4.3. Then at init panels and output window dataset will run and the CFS will perform an search method using backward search, forward search and bi-directional search and then we had choose forward search so that the first come first serve will perform to get best first feature and then perform filtering of data on the basis of two attributes src_bytes and dst_host_srv_count and then preprocessing of data will perform for extracting best feature. After preprocessing and filtering, classification of data will performed on selected set to get an accurate result. Then the next step is to mine proposed

PREPROCESSING AND FILTERING OF NETWORK TRAFFIC CLASSIFICATION USING ADABOOST ALGORITHM

algorithms with previous algorithms to compare their results on the basis of correctly and incorrectly classified instances.

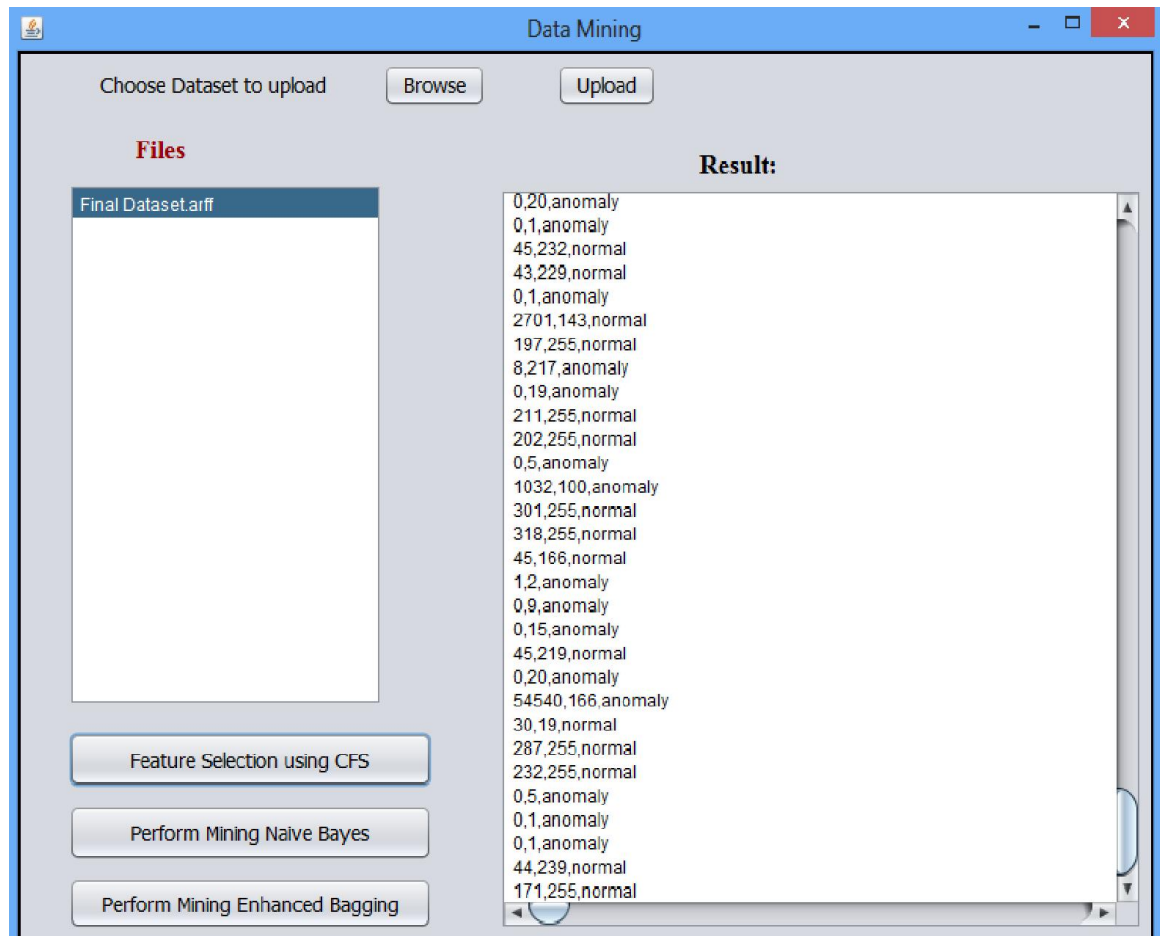


Figure 4.3: Feature Selection

As in figure 4.4, shown the CFS selecting best features from dataset to perform Naive Bayes.

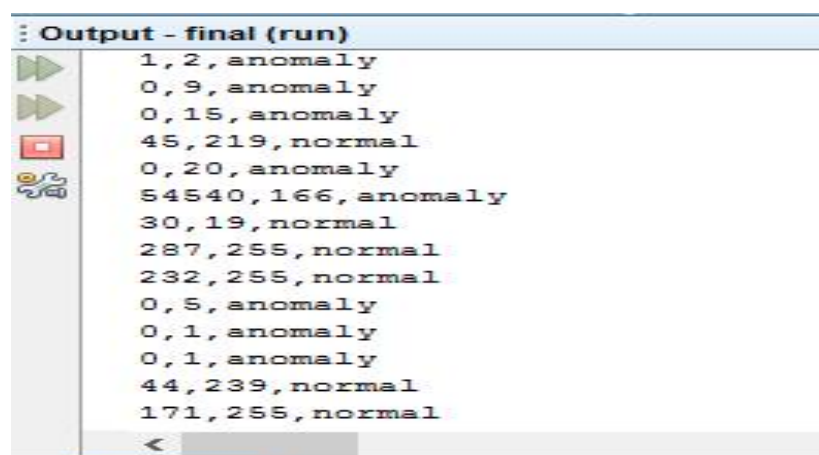


Figure 4.4: Dataset extracting best Features

PREPROCESSING AND FILTERING OF NETWORK TRAFFIC CLASSIFICATION USING ADABOOST ALGORITHM

In figure 4.5, after selecting the feature set using Consistency Feature Selection we can perform mining using Naive Bayes and then the result window will appear on the screen and the accuracy of our new approach is shown in correctly classified instances and incorrectly classified instances with the help of various parameters like TP rate, FP rate, Precision, Recall etc.

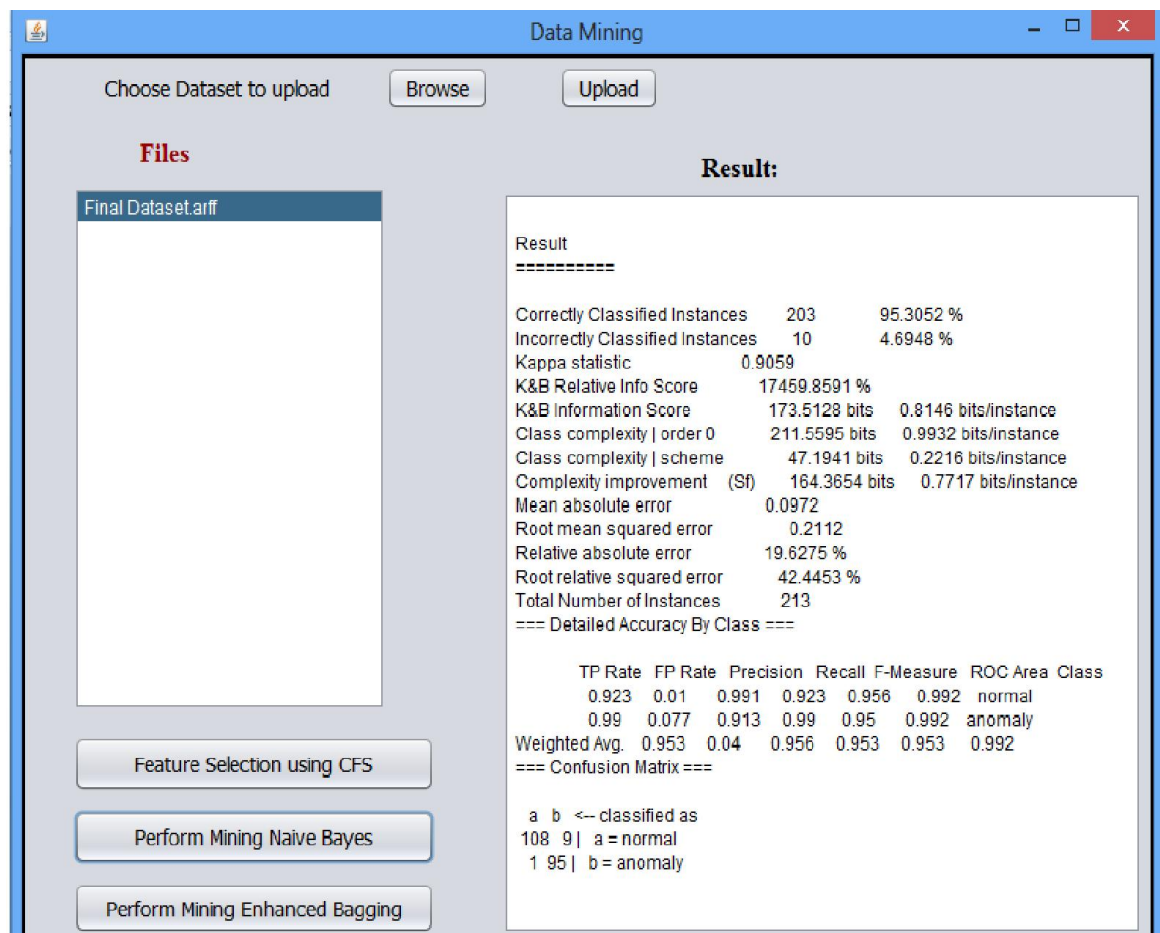


Figure 4.5: Naive Bayes using Feature Selection

In our old approach the use of feature selection is not given so the accuracy of Naive Bayes without using the CFS is 94% which is below than our new approach as in figure 4.6 and the old approach is also using C4.5 algorithm to improve the accuracy of Naive Bayes but it is not successful and the new accuracy is below 91% as shown in figure 4.7. As we can see in our new approach the correctly classified instances is 95% and the in correctly classified instances is 4% and with this we can say the data classified is more efficient than old approach. So total number of instances get 213 by this we can calculate

PREPROCESSING AND FILTERING OF NETWORK TRAFFIC CLASSIFICATION USING ADABOOST ALGORITHM

confusion matrix and get the weighted average by calculating the F-measure and ROC area and specify to which class they belong.

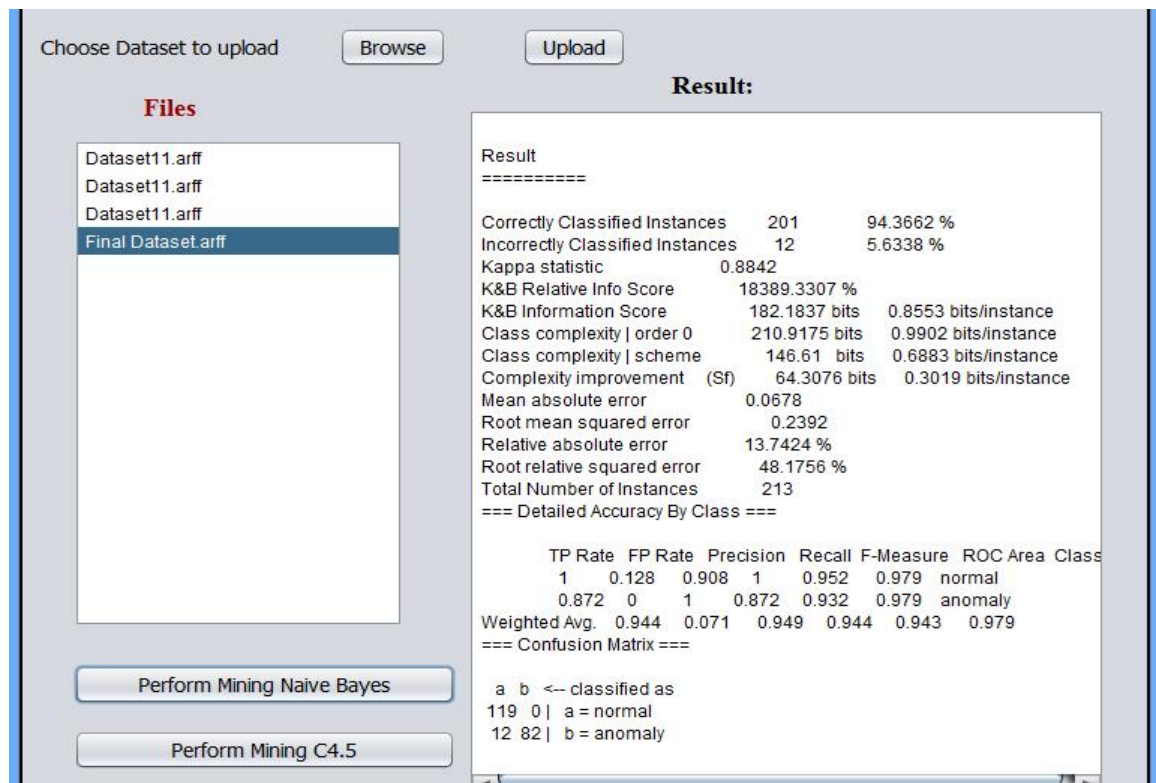


Figure 4.6: Naive Bayes without feature selection

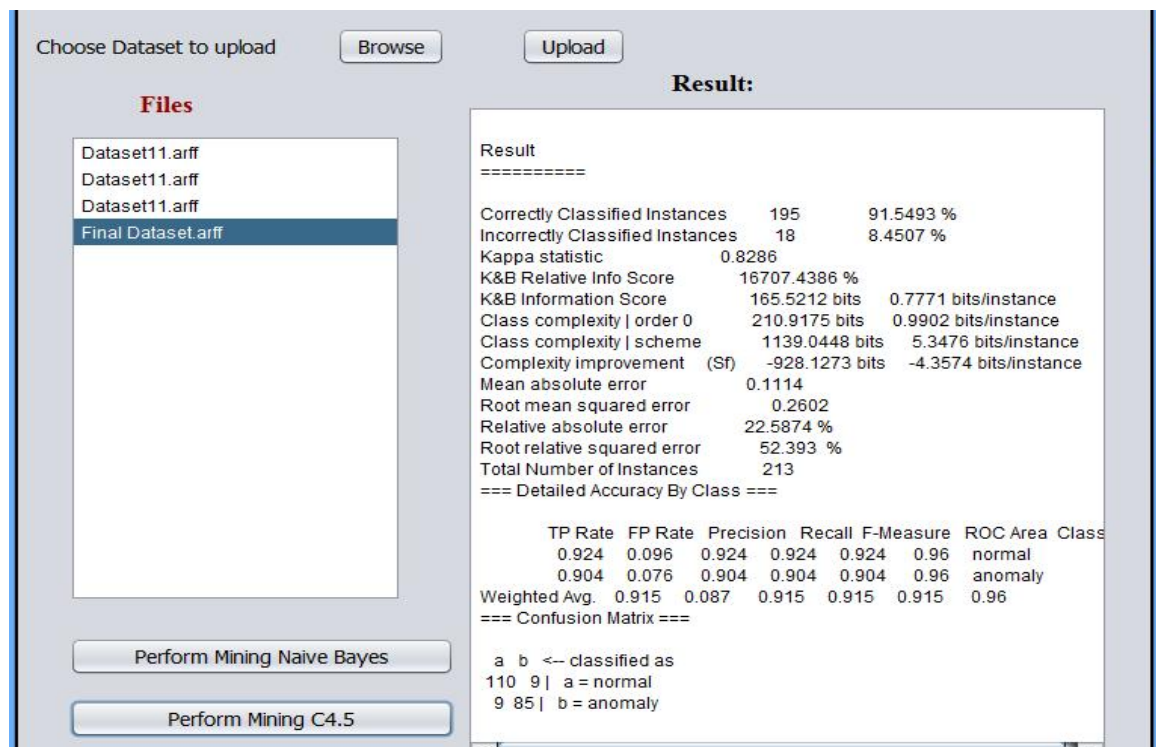


Figure 4.7: Perform Mining using C4.5

PREPROCESSING AND FILTERING OF NETWORK TRAFFIC CLASSIFICATION USING ADABOOST ALGORITHM

In figure 4.8, we performed the mining using bagging and bagging using the classifier named RP tree but in our work there is no need of decision tree so instead of RP Tree we used the classifier named AdaBoost which is made by combining Boosting and Bagging. By combining them we get the accuracy of 97%. Bagging algorithms are weak learners and boosting algorithm convert weak learners to strong learners. They use meta-learning technique and perform decision stump. The result is shown on the basis of main three parameters used for classification are Accuracy, Precision and Recall. In the result panel the confusion matrix is classified as normal and anomaly class which means known and unknown traffic.

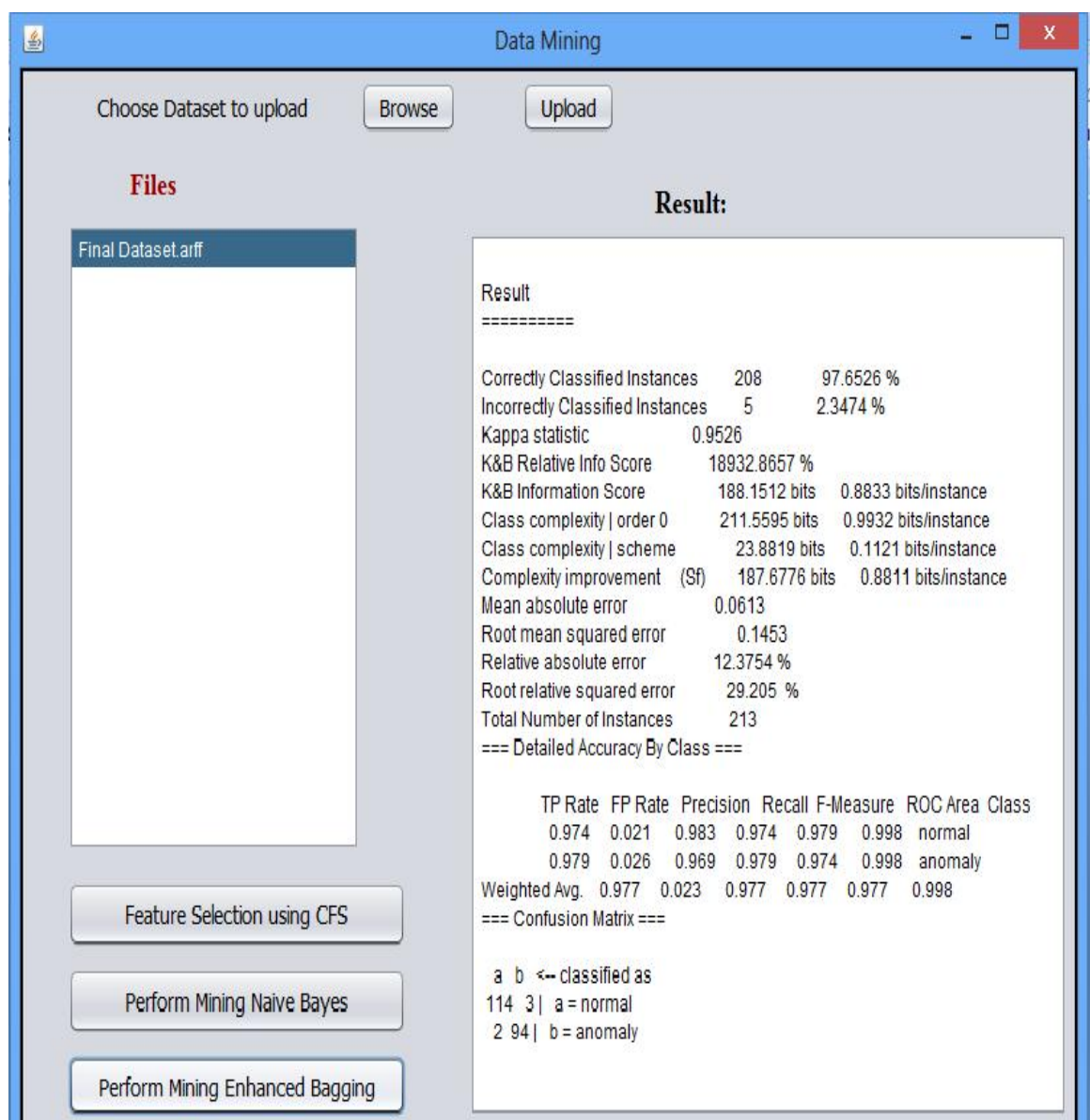
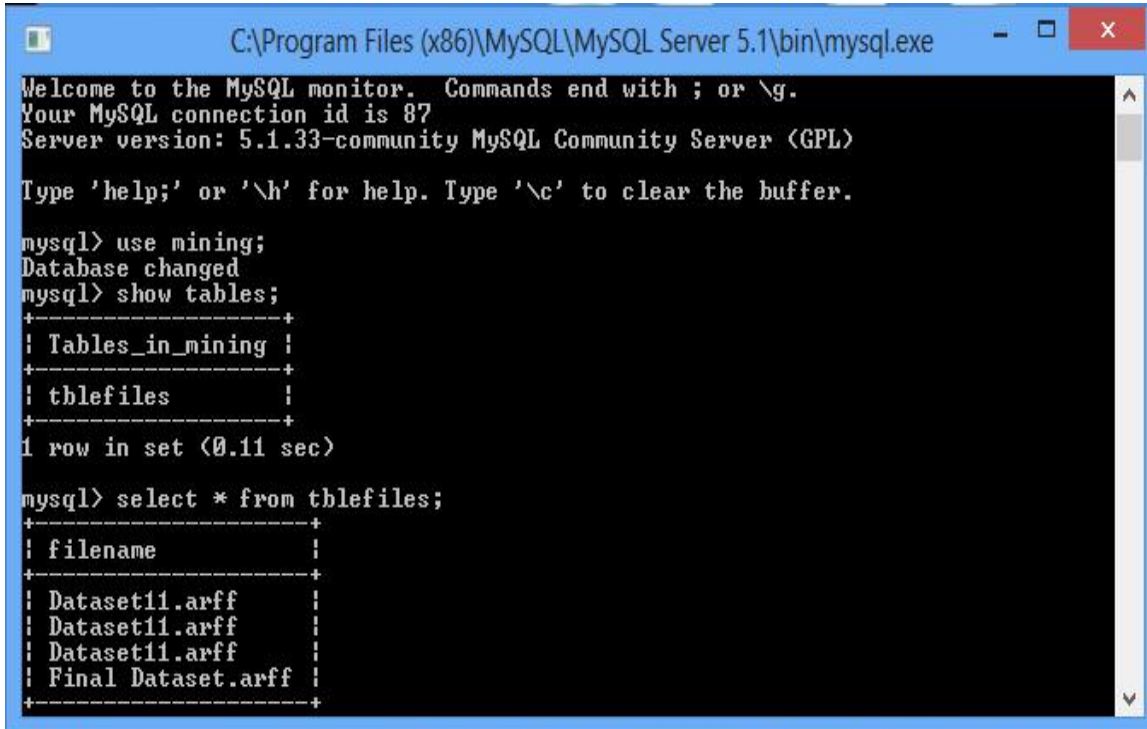


Figure 4.8: Perform Mining using Bagging

PREPROCESSING AND FILTERING OF NETWORK TRAFFIC CLASSIFICATION USING ADABOOST ALGORITHM

In figure 4.9, the MySQL server 5.1 is performing big role as the backend. It is the best relational SQL database management system which includes cursor, view, schema etc. for implementation of queries in SQL. In this database can be created by making tables and data get stored and if we want to operate any insert, delete and update operation is also done in this MySQL command line prompt. Our project is running under port 3307 using the username “root” with no password. The program files are automatically get stored in the C:\ drive.

A screenshot of a Windows command prompt window titled "C:\Program Files (x86)\MySQL\MySQL Server 5.1\bin\mysql.exe". The window displays the MySQL command-line interface. The text shown is: "Welcome to the MySQL monitor. Commands end with ; or \g. Your MySQL connection id is 87 Server version: 5.1.33-community MySQL Community Server (GPL) Type 'help;' or '\h' for help. Type '\c' to clear the buffer. mysql> use mining; Database changed mysql> show tables; +-----+ | Tables_in_mining | +-----+ | tblefiles | +-----+ 1 row in set (0.11 sec) mysql> select * from tblefiles; +-----+ | filename | +-----+ | Dataset11.arff | | Dataset11.arff | | Dataset11.arff | | Final Dataset.arff | +-----+" data-bbox="151 281 878 608"/>

```
C:\Program Files (x86)\MySQL\MySQL Server 5.1\bin\mysql.exe
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 87
Server version: 5.1.33-community MySQL Community Server (GPL)

Type 'help;' or '\h' for help. Type '\c' to clear the buffer.

mysql> use mining;
Database changed
mysql> show tables;
+-----+
| Tables_in_mining |
+-----+
| tblefiles        |
+-----+
1 row in set (0.11 sec)

mysql> select * from tblefiles;
+-----+
| filename |
+-----+
| Dataset11.arff |
| Dataset11.arff |
| Dataset11.arff |
| Final Dataset.arff |
+-----+
```

Figure 4.9: MySQL Server 5.1

In this figure 4.10, accuracy of Naive Bayes and C4.5 of previous approach without feature selection is shown. The correctly and incorrectly classified instances calculated during implementation and shown with two different perspectives that correctly classified instance should increase from other algorithm and incorrectly classified instance get decreased from first algorithm. They work opposite to each other. As in figure 4.11, the graph shows the performance of Naive Bayes with feature selection and Bagging using AdaBoost algorithm. The performance is shown on the basis of accuracy due to correctly and incorrectly classified instances. From these two graphs we can recognize the difference in accuracy very clearly and the flow of traffic is generated in the form of normal and anomaly classes.

PREPROCESSING AND FILTERING OF NETWORK TRAFFIC CLASSIFICATION USING ADABOOST ALGORITHM

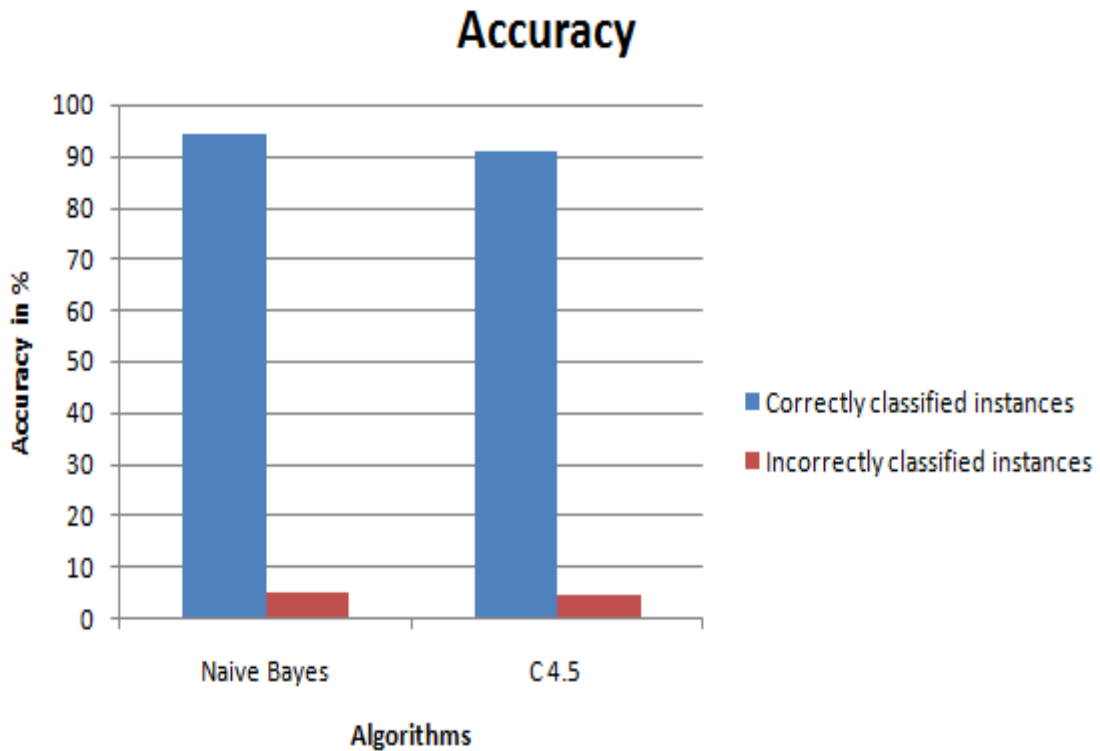


Figure 4.10: Naive Bayes accuracy without Feature Selection

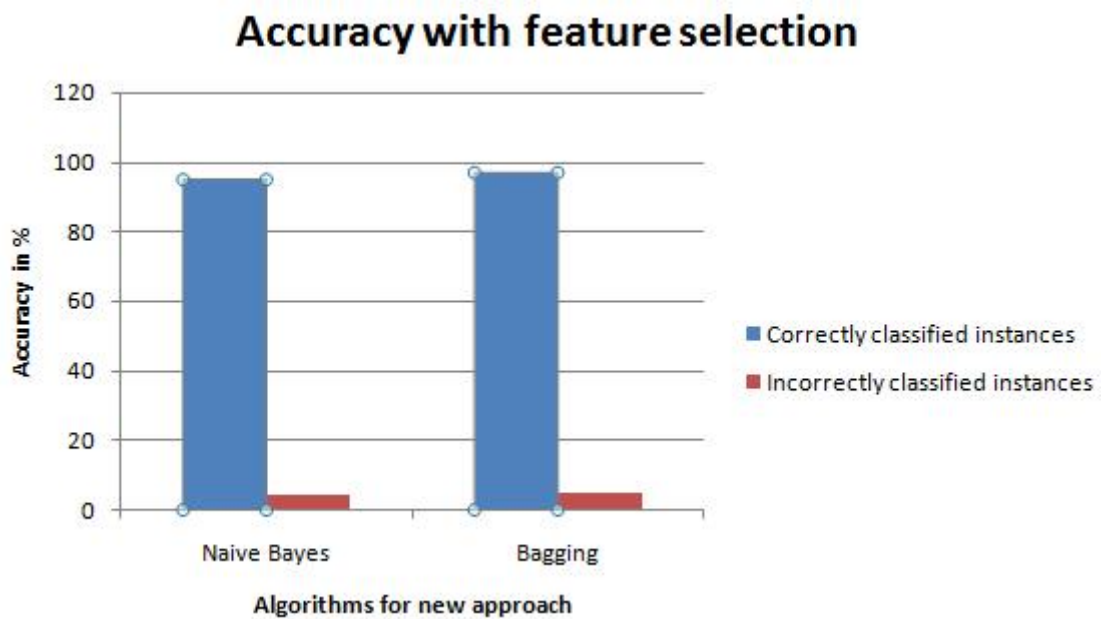


Figure 4.11: Naive Bayes accuracy with Feature Selection

This graph shows the flow of traffic is generated in the form of known and unknown traffic in each case as shown in figure 4.12. the normal and anomaly traffic is calculated by using confusion matrix and shows the result by multiplying diagonal columns with one another. The mining performed using Bagging and Boosting shows the correct traffic instances.

PREPROCESSING AND FILTERING OF NETWORK TRAFFIC CLASSIFICATION USING ADABOOST ALGORITHM

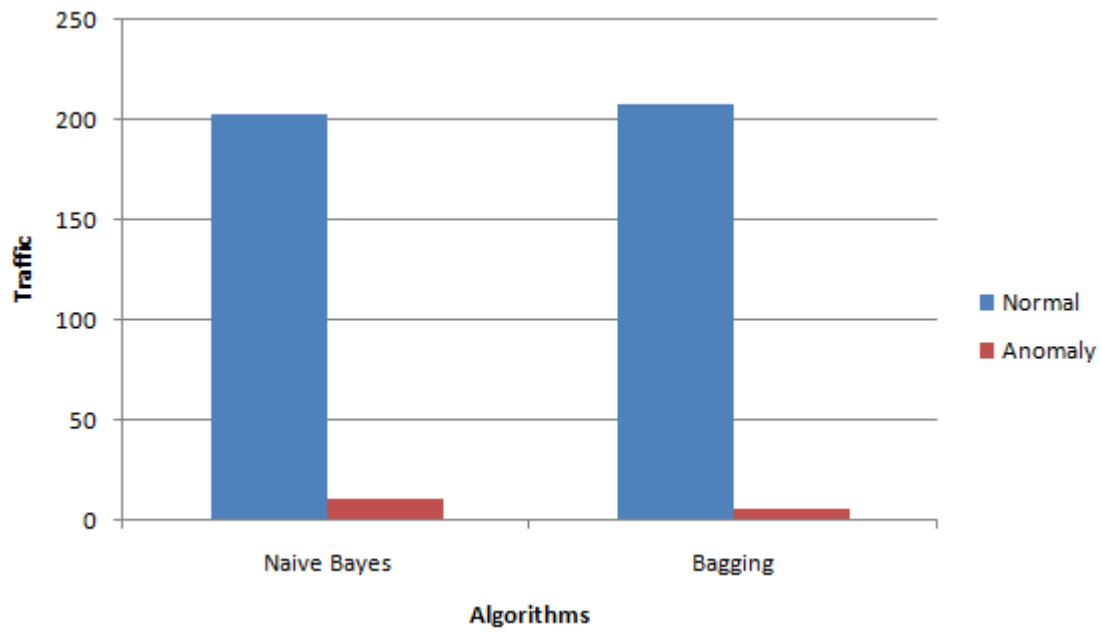


Figure 4.12: Graph for Known and Unknown Traffic

CHAPTER 5

CONCLUSION AND FUTURE SCOPE

5.1 Conclusion

The proposed solution and implementation justify that from previous approaches our results are more accurate. This is by combining Bagging and Boosting. As from previous approach the performance of Naive Bayes and C4.5 are not much efficient without the use of feature selection. The known and unknown traffic is also calculated with confusion matrix as for Naive Bayes, known traffic is 203 and unknown traffic is 10 but in the case of bagging known traffic is mor and unknown decreased from Naive Bayes which means the use of Bagging and Boosting improved our results more accurately. In means of accuracy the correctly classified instances in case of Naive Bayes is lower than Bagging and total instances used in both are equal. The accuracy using Naive Bayes is 95% and by using Bagging is 97%. In previous approach the accuracy using Naive Bayes without feature selection is 94%. But now with feature selection the results are more accurate and the proposed methods have improved the parameters like TP rate, FP rate, Precision etc.

5.2 Future Scope

In future we can propose more efficient algorithms so that they can perform more accurate results. In future we can perform majority voting on the Naive Bayes by using any clustering and classification algorithms.

-
- [1] Micheline Kamber Jiawei Han, *Data Mining concepts and Techniques*. Waltham, USA: Morgan Kaufman Publishers, 2012.
- [2] Kamber Han J, "*Data Mining: Concepts and Techniques*". India, pp. 6-10: Morgan Kaufman, 2006.
- [3] Luoshi Yibo, ""Traffic Classification: Issues and Challenges", "*Journal of Communications*, pp. China, April, vol. 8,no.4., 2013.
- [4] Afshin Rostmizadeh. (2012, October) Wikipedia. [Online].
en.wikipedia.org/wiki/Supervised_learning
- [5] Sanchika Bajpai,Sonali Khairnar Shital Salve, ""Clustering of Network Traffic by Online Streaming", "*International journal of advanced research in computer science and software engineering*, pp. pp. 149-155, vol. 4(3), March, Pune, 2014.
- [6] Yang, Chao Jun Zhang, ""An effective Network Traffic Classification method with unknown flow detection", "*IEEE Transactions on network and Service Management*, June, pp. vol. 10, no. 2., 2013.
- [7] Ashkan Sami Armin Daneshpazhoun, "Semi-supervised outlier detection with only positive and unlabeled data based on Fuzzy Clustering," in *Information and Knowledge Technology(IKT)*, Iran, 2013.
- [8] Wen-Hoar Hsaio Chien-Liang Liu, "Semi-Supervised Linear Discriminant Clustering," *IEEE transactions on Cybernetics*, 2013.
- [9] M. Tamilkili, "A survey on recent Traffic Classification techniques using Machine learning methods," *International Journal of advanced Research in Computer science and Software engineering*, pp. vol. 3(12), pp. 368-373, 2013.
- [10] Vinodh Edwards S.E Jamuna A, ""Efficient flow based network traffic classification using machine learning", "*International journal of Engineering reserach and Applications (IJERA)*, pp. vol. 3(2), pp. 1324-1328, 2013.

PREPROCESSING AND FILTERING OF NETWORK TRAFFIC CLASSIFICATION USING ADABOOST ALGORITHM

- [11] Robert Kapo Karl B. Dyer, "COMPOSE: A Semi-supervised learning framework for initially labeled nonstationary streaming data," *IEEE Transactions on Neural networks and Learning Systems*, 2013.
- [12] Yang, Jun Yu, ""Internet Traffic Clustering with Constraints"," *IEEE*, p. Australia., 2012.
- [13] Tahir Tomasz, ""A method for classification of network traffic based on C5.0 Machine Learning Algorithm"," *IEEE*, p. Denmark., 2012.
- [14] Grenville Thuy, ""Timely and Continuous Machine-Learning-Based Classification for Interactive IP Traffic"," *IEEE/ACM Transactions on Networking*, pp. Melbourne, December, vol. 20, no. 6., 2012.
- [15] K. P. T. Karasgianni, ""A survey of techniques for internet traffic classification using machine learning"," *IEEE*, 2012.
- [16] Yi Shen Bin, ""An Survey on Machine Learning based Network Traffic Classification"," *Journal of Information and Computational Science*, pp. China, October., 2012.
- [17] Marios Iliofotu Guowu Xie, "Subflow: Towards practical Flow-level Traffic Classification," in *IEEE International Conference*, California, 2012.
- [18] C. H. C. Jian Min Wang, ""Study on process of network traffic classification using machine learning"," in *IEEE China Grid Conference*, 2010, pp. China, pp. 262-266.
- [19] M. Mellia, M, Meo & D. Rossi A. Finamore, ""Stochastic Packet inspection classifier for UDP traffic"," pp. October, pp. 1505-1575, vol. 18, no. 5, 2010.
- [20] Z. L. Wang Ruoyu, ""A new resampling method for network traffic classification using Supervised Machine Learning"," *IEEE*, p. China., 2010.
- [21] Shrivastav, "Network Traffic classification using Semi-Supervised approach," in *Second International Conference on Machine learning and Computing*, Bangalore, 2010, pp. pp. 345-349.

PREPROCESSING AND FILTERING OF NETWORK TRAFFIC CLASSIFICATION USING ADABOOST ALGORITHM

- [22] D. Lucerna, C. Rottondi & G. Verticale S. Bregni, "Using per-source measurements to improve performance of Internet Traffic Classification", 2010.
- [23] Zhou Dong, "The Study of Network Traffic Identification based on Machine Learning Algorithm", *IEEE, China.*, 2009.
- [24] Mahbod Tavallaee, "Online Classification of Network flows," in *Seventh annual communication network and services research conference* , Canada, 2009, pp. pp. 78-84.
- [25] Kai Chen Shijun Huang, "A Statistical-feature-based approach to Internet traffic Classification using Machine Learning", *IEEE*, pp. China, pp. 856-863, 2009.
- [26] Qiong Sun Xu Tian, "Dynamic Online Traffic Classification using data stream mining," in *State key lab. of networking and switching technology*, Beijing, 2008, pp. pp. 104-107.
- [27] Carlos Arthur, "A Survey on Internet Traffic identification and classification", *IEEE*, pp. Brazil, September., 2007.
- [28] Zhu Li, "Accurate classification of the Internet Traffic based on the SVM model", *IEEE communication society*, pp. pp. 1373-1378, 2007.
- [29] [Online]. <http://www.java.com/en/downloads/faq/>
- [30] webopedia. [Online]. webopedia.com/term/j/java.html
- [31] (2014, May) wikipedia. [Online]. www.wikipedia.com/
- [32] (2014, May) Wampserver. [Online]. www.wampserver.com/en/
- [33] (2014, May) Yahoo Answers. [Online]. <https://in.answers.yahoo.com/question/>
- [34] S. Zander & G. Armitage, "Distributed firewall and flows-shaper using statistical evidence," 2010.
- [35] Bo Yang, Lei Zhang & Shan Jhang Runyuan Sun, "Traffic classification using

PREPROCESSING AND FILTERING OF NETWORK TRAFFIC CLASSIFICATION USING ADABOOST ALGORITHM

Probabilistic Neural Network", 2010.

[36] Jiahai Donghong, "IP Traffic Classification based on Machine Learning", *IEEE*, p. China., 2011.

[37] Luoshi and Dawei Yibo, "Computer science and technology, Traffic classification: issues", *Journal of Communication*, pp. April, Vol. 8, No. 4., 2013.

[38] Sasan Adibi, "A captured wired traffic approach," *International journal of advanced Science and Technology*, pp. vol. 21, no. 2, 2010.

CHAPTER 7 APPENDIX

7.1 Glossary of terms

PREPROCESSING AND FILTERING OF NETWORK TRAFFIC CLASSIFICATION USING ADABOOST ALGORITHM

1. Data Mining: Data mining is a knowledge discovery of data and to mine huge amount of data.
2. JAVA: is a programming language and resembles like C++.
3. MySQL: MySQL is a database system from which a website pick data through query. Query is executed through execution of code written in server side programming language such as Java. It stands for Structured Query Language.
4. KDD Dataset: Knowledge Discovery of Data also refers as mined data and KDD dataset is used for networking to calculate different traffic classes.
5. Accuracy: is correctly classified instances in class.
6. Precision: is correctly selected instances out of total instances.
7. Naive Bayes: are statistical classifiers which can predict class labels and also called class-conditional independence.
8. Bagging: is meta-learning technique and divide training data into different parts to perform majority vote to get final result.

7.2 Publications

Supriya Katal, Asstt. Prof. Hardeep Singh, "A survey of Machine Learning Algorithm in Network Traffic Classification", *International Journal of Computer Trends and Technology*, vol. 9, no. 6, March 2014.