A DISSERTATION

ON

COMPARATIVE ANALYSIS OF CLASSIFICATION ALGORITHMS FOR DEFECT PREDICTION

Submitted in partial fulfillment of the requirements

For the award of the degree of

MASTER OF TECHNOLOGY

IN

SOFTWARE TECHNOLOGY

by

Yesha Grover

Roll No: 2K11/SWT/19

Under the Guidance of

Dr. Kapil Sharma

Associate Professor

Department of Software Engineering , DTU

Department of Software Engineering

Delhi Technological University

New Delhi

2014

DECLARATION

I hereby declare that the thesis entitled "**COMPARATIVE ANALYSIS OF CLASSIFICATION ALGORITHMS FOR DEFECT PREDICTION"** which is being submitted to the **Delhi Technological University**, in partial fulfillment of the requirements for the award of degree of **Master of Technology in Software Technology** is an authentic work carried out by me. The material contained in this thesis has not been submitted to any university or institution for the award of any degree.

Yesha Grover

Department of Software Engineering

Delhi Technological University,

Delhi.

CERTIFICATE

DELHI TECHNOLOGICAL UNIVERSITY

Date: _____

This is to certify that the thesis entitled **"COMPARATIVE ANALYSIS OF CLASSIFICATION ALGORITHMS FOR DEFECT PREDICTION "** submitted by **Yesha Grover (Roll Number: 2K11/SWT/19),** in partial fulfillment of the requirements for the award of degree of Master of Technology in Software Engineering, is an authentic work carried out by her under my guidance. The content embodied in this thesis has not been submitted by her earlier to any institution or organization for any degree or diploma to the best of my knowledge and belief.

Project Guide

Dr. Kapil Sharma

Associate Professor

Department of Software Engineering

Delhi Technological University, Delhi-110042

ACKNOWLEDGEMENT

I take this opportunity to express my deepest gratitude and appreciation to all those who have helped me directly or indirectly towards the successful completion of this thesis.

Foremost, I would like to express my sincere gratitude to my guide **Dr. Kapil Sharma**, **Associate Professor, Department of Software Engineering**, **Delhi Technological University, Delhi** whose benevolent guidance, constant support, encouragement and valuable suggestions throughout the course of my work helped me successfully complete this thesis. Without his continuous support and interest, this thesis would not have been the same as presented here.

Besides my guide, I would like to thank Dr. Ruchika Malhotra and entire teaching and non-teaching staff in the Department of Software Engineering, DTU for all their help during my course of work.

Yesha Grover

2K11/SWT/19

Master of Technology (Software Technology)

Delhi Technological University

Bawana road, Delhi - 110042

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Introduction

The objective of this thesis is to evaluate the performance of the classification algorithms on binary classification problems using a variety of performance metrics: classification accuracy, precision, recall (sensitivity), F-Measure and ROC area. The evaluation is performed with an intention to identify which algorithm suits best for prediction of defect prone classes in software based on software quality metrics.

Motivation

As independent testing team, it is important to plan and manage the test execution activities in order to meet the tight deadline for releasing the software to end-users. Since the aim of test execution is to discover as many defects as possible, testing team is usually put into burden to ensure all defects are found and fixed by the developers within the system testing phase. Additional number of days has to be added to the timeline to accommodate testing team in completing their test with the hope that all defects have been found and fixed. On the other hand, the stakeholders would also ask the testing team on the forecasted defects in the software so that they could decide whether the software is feasible and fit for release. This is due to the nature that system testing is the last gate before the software is made visible to end-users, thus as the custodian of executing system testing, the independent testing team has to take responsibility to ensure software to be released is of high quality.

Therefore, the ability to predict how many defects that can be found and what are the defect prone areas; at the start of system testing shall be a good way to tackle this issue. This becomes the reason for conducting this study.

Besides serving as a target on how many defects to capture in system testing, defect prediction can also become an early quality indicator for any software entering the testing phase. Testing team can use the predicted defects to plan, manage and control test execution activities. This could be in the form aligning the test execution time and number of test engineers assigned to particular testing project.

Having defect prediction as part of the testing process allows testing team to strengthen their test strategies by adding more exploratory testing and user experience testing to ensure known defects are not escaped and re-introduced to end-users. Test engineers would be able to have better root cause analysis of the defects found. In the long run, testing can achieve what is called as zero-known post release defects for that particular software.

Work

In this thesis, Random Forest , LogitBoost , Bagging , KStar , RBFNetwork , Logistic , Naïve Bayes , Bayes Net algorithms have been implemented on a real-world dataset. The goal of the research was to evaluate the performance of the classification algorithms on binary classification problems using a variety of performance metrics: classification accuracy, precision, recall (sensitivity), F-Measure and ROC area.

Conclusion

As obtained in the experimental results, true positive rate for LogitBoot algorithm is the best. However, for Naïve bayes algorithm, true positive rate is the lowest. Random Forest Algorithm and Bagging Algorithm also fared well for the true positive rate. It was also

observed that false positive rate for Bagging and Random Forest algorithm are the highest. However, for RBF Network algorithm, false positive rate is the lowest. Recall was found to be highest for LogitBoost algorithm. However, for naïve bayes algorithm, recall is the lowest. It can be observed that false precision for Naïve Bayes algorithm are the highest. However, for Bagging and Random Forest algorithms, precision is the lowest. F-Measure for LogitBoost algorithm is the highest. However, for Naïve Bayes algorithm, F-Measure is the lowest. ROC area is the highest for KStart algorithm. However, for Bagging algorithm, ROC Area is the lowest.

We found that which algorithm performed best depended on the type of problem being considered, dataset characteristics and the performance matrix used.

Chapter 1

INTRODUCTION

In the recent past, there has been an exponential increase in the amount of stored data. Managers and decision makers are faced with the problem of information overload. For example, in 1992, Frawley, Piatetsky-Shapiro and Matheus reported that the amount of data in the world doubles every twenty months. Cios, Pedrycz, and Swiniarski in 1998 reported that, Wal-Mart alone uploads twenty million point of sale (POS) transactions every day. Today we have far more information stored than we can handle. But as data volume increases, making meaningful decisions becomes increasingly difficult. To address these issues, researchers turned to a new research called Data Mining and Knowledge Discovery in Databases. In the past decades data mining methods have been widely used for the purpose of extracting knowledge from large data. Classification, a supervised method used to partition variables into several classes, represents the most widely used data mining method.

There have been several studies on comparing classification algorithms. However, most of these studies have been limited to only a very few classification algorithms. The theme of my thesis is to compare and better understand the prevalent classification algorithms, by evaluating the performance of nine different classification algorithms on several real world datasets.

Statement of the Problem

Machine learning and data mining researchers have developed an abundance of classification algorithms to solve classification problems. A number of commercial tools are also available today to provide decision makers a range of classification techniques. However, no single classification algorithm has been demonstrated to be superior to the others in all scenarios. Neither is it totally clear as to which algorithm should be preferred over the others under specific circumstances. Decision makers are therefore faced with the important question: "What is the best choice of a classification algorithm for my particular application?" This problem is termed *classification algorithm selection.* The primary focus of my research will be to evaluate the application of several classification algorithms using both statistical and machine learning methods on a dataset. An important aspect of my thesis is to use a variety of performance criteria to evaluate the learning method. The performance criteria we have chosen to evaluate the algorithms are model accuracy, precision, recall, specificity, F-Measure and ROC area.

Contribution of the Thesis

Although there have been much research comparing classification algorithms, unfortunately, no one has been able to determine which algorithm is superior to the others in all scenarios, nor is it totally clear as to which algorithm should be preferred given specific circumstances. There have been inconsistencies in the results as shown in the literature review. For example in 1998,

Bhattacharyya and Pendharkar argued that the interpretation of results from these studies is difficult considering the variations in data and algorithms used, data pre-processing steps and the optimization of technique-related parameters in some studies, etc. Considering the vast

amounts of data collected everyday in various domains such as health care, financial services, point of sale transactions and many others, there is a pressing need to convert this information into knowledge. Machine learning and data mining are both concerned with achieving this goal in a scalable fashion. More important, since classification algorithms are the most widely used data mining method, there is the need to better understand the classification techniques and paradigms which form an integral part of machine learning and data mining research, in order to eliminate the gap between the results predicted by theory and the behavior observed in practice.

Organization of the Thesis

The thesis is organized into six chapters. In Chapter 2 a review of the relevant related background of the literature is discussed which provides the groundwork for this thesis. In Chapter 3, research background and dataset description is provided. Chapter 4 describes the methods used for the thesis, the classification task in general, the algorithms, the datasets used, and the feature selection method. Chapter 5 includes the analysis and discussion of results and performance of the algorithms. Conclusion and Summary is in Chapter 6.

Chapter 2

LITERATURE SURVEY

Several previous studies have been done to compare classification algorithms. Perhaps the most extensive is the STATLOG project, by King, Feng, and Sutherland (1995). Their work was carried out on several datasets to determine to what extent the various techniques met the data analytic needs of industry. They argued that the compiled interpretation of results from various studies is however, difficult, considering the variations in the data and the algorithm used. They said data pre-processing steps employed and the optimization of technique-related parameters could possibly favor one algorithm over the other.

Several studies have been focused on traditional statistics and artificial intelligence; however, very few have focused on machine learning. Lim, Loh, and Shih (2000) compared twenty-two decision tree algorithms, nine statistical techniques, and two neural network algorithms in terms of classification accuracy, training time and number of trees. The study found that C4.5, IND-CART, and QUEST have the best combinations of error rate and speed. QUEST and logistic regression were substantially faster. They also reported that comprehensibility of tree structures decreases with increase in tree size and complexity. They claim that if you use the same kind of test on two trees with the same prediction accuracy, the one with fewer leaves is usually preferred. King, Feng, and Sutherland (1995) did a comparison of algorithms for symbolic learning, statistics, and neural networks. They used twelve datasets and found out that the performance was greatly affected by the dataset used instead of the algorithm used.

Pesonen (1997) compared discriminant analysis, logistic regression analysis, cluster analysis and a backpropagation network in the diagnosis of acute appendicitis. The results of the four classification methods were evaluated using the receiver operating characteristic (ROC) curve. Discriminant analysis and backpropagation showed slightly better results than the other methods. Pesonen concluded that a backpropagation neural network offers a good choice for statistical classification methods. Chen (1991) compared three types of neural networks (backpropagation, radial basis functions, and probabilistic neural networks) with the statistical method of nearest neighbor decision rule. The classification target was simulated active sonar waveforms. He found out that all three neural networks outperformed the nearest neighbor algorithm.

Ripley (1994) compared discriminant analysis, nearest neighbor, backpropagation neural networks, MARS, and a classification tree on a few classification problems. The evaluation measure used was the percentage correctly classified. It turned out that the various tools were approximately equally matched. Ripley concluded that: "Neural networks emerge as one of a class of flexible non-linear regression methods which can be used to classify via regression". Curran and Mingers (1994) compared discriminant analysis, decision trees, and neural networks across seven datasets. Four datasets contained real data and three were artificially created. Discriminant analysis performed well when the dataset proved to be linearly separable. Neural Networks performed well on the sphere data and fairly well on across all datasets. It did better than discriminant analysis when there were non-linear relationships between predictors and classes

but slightly worse when the data were linearly separable. Recent studies have shown that artificial intelligence (AI) methods achieve better performance than traditional statistics. In an attempt to provide a model with better explanatory power, Huang, Chen, Hsu, and Chen (2004) did an analysis of credit rating using Support Vector Machines (SVM) and Neural

Networks. They used backpropagation neural network (BNN) as a benchmark. However, they noticed only slight improvement in the SVM. They used recent research results in neural network models to interpret the AI models. Mushtaq et al. (2006) applied different classification algorithms on preprocessed financial data in an attempt to evaluate the classification algorithm that would have the best predictive accuracy given different parameters. They concluded that the C4.5 model had better predictive accuracy for the transactional and frequently occurring data than either ID3 or ZeroR.

Brown, Corruble, and Pittard (1993) compared back propagation neural networks and decision trees for multimodal classification problems. Decision trees performed better on datasets which contained irrelevant attributes which they were able to ignore. On two other datasets in which most variables were useful in discriminating the classes, neural networks outperformed decision trees. The bottom line seems to be that neural networks do not have the capability to deal well with irrelevant attributes in the dataset, but decision trees do. Several comparative studies have been done ranging from single algorithm with variations of the algorithm to multiple algorithms. There have been a number of studies comparing the performance of machine learning and statistical methods.

Bhattacharyya and Pendharkar (1998) argue that the interpretation of results from these studies is difficult considering the variations in data and algorithms used, data pre-processing steps employed that could give the edge to one algorithm over the other, optimization of technique-related parameters in some studies and so on.

Hand and Henley (1997) conducted a review of different statistical classification methods used for credit rating. They analyzed particular problems arising in the credit scoring system and reviewed the statistical methods which have been used. They found that discriminant analysis and logistic regression analysis are the most widely used techniques for building credit scoring rating.

In a separate study in 1996, Henley and Hand used k-nearest neighbor (KNN) method to access credit worthiness of consumers applying for a loan. Hand and Henly proposed this technique as an improvement over the traditional credit scoring technique. They showed that KNN method with adjusted Euclidean distance metrics can give a slightly improved prediction of consumer credit risk than the traditional technique.

Finch and Schneider (2006) compared the classification accuracy of linear discriminant analysis (LDA), quadratic discriminant analysis (QDA) logistic regression, and classification and regression trees (CART). They based their comparison on the condition of the data collected. They used Monte Carlo Simulation to access the cross-validated predictive accuracy of the methods, and concluded that QDA performed as well as or better than the other alternatives in virtually all conditions. Similar studies for comparing cross-validated classification accuracies of predictive discriminant analysis, and logistic regression classification models under various data conditions for two-group classification problems have been conducted by Meshbane and Morris (1996). The most commonly used methods for solving two-group classification problems are the two most popular methods of Yarnold, Hart, and Soltysik (1994).

There have been several but inconsistent results in the study to compare the classification accuracy of logistic regression and predictive discriminant analysis (PDA). Some of these studies suggested that logistic regression is more accurate than PDA for non-normal data. However, several other studies show otherwise. They found no difference in the accuracy of the two techniques using non-normal data, Morris and Lieberman (2007).

Berardi, Patuwo and Hu (2004) presented a principled approach for building and evaluating neural network classification models for decision support system implementation and e-commerce application in their study. The study was aimed at understanding how to utilize e-commerce data for Bayesian classification within a neural network framework to yield more

accurate and reliable classification decisions and showed that neural networks are ideally suited for noisy data like e-commerce data.

Kiang (2003) did a comparative assessment of classification methods. Kiang considered classification techniques used in data mining i.e. NN, decision trees models and statistical methods, i.e. LDA, logistic regression analysis and KNN models. In the study, Kiang used synthetic data to perform a controlled experiment in which the data characteristics were systematically altered to introduce imperfections such as nonlinearity, multicollinearity, unequal

Covariance, etc. The study was performed to investigate how these different classification methods performed when certain assumptions about the data characteristics were violated. Kiang showed that data characteristics considerably impacted the classification performance of the methods.

Pohar, Blas, and Turk (2004) conducted comparative studies between linear discriminant analysis and logistic regression. They learned that linear discriminant analysis is the more appropriate method to use when the exploratory variables are normally distributed. However, linear discriminant analysis fails when the number of categories is really small, for example, two or three. Despite this they concluded that linear discriminant analysis still remains preferable to logistic regression when the assumptions are met. When the assumptions of linear discriminant analysis are not met, the logistic regression gives good results regardless of the distribution.

Willett (2002) created predictive models using both classical statistics and data mining algorithms to develop the most accurate model for predicting non-persistence, and developed a targeted experimental intervention to increase student persistence. The study found that C5.0 boosted five-fold was the most accurate method. Suresh and Balasaheb (2008) conducted a study on the problem of classification of sonar signals. The purpose of their

study was to make a comparison of decision tree induction and neural network classifiers in a classification of sonar signal databases. They wanted to discriminate between sonar signals bounced off of a metal cylinder and those bounced off of a roughly cylindrical rock using discriminant functions such as multilayer perception neural

networks (MLP NN) and a classification and regression trees (CART). Their conclusion was that the MLP NN classifier worked as an optimal classifier for the given task with average classification accuracy at about 90%.

Lin, Huang, and Chang (2004) used statistical techniques to predict the correct placement of a student in his appropriate group. They considered five science-educational indicators for each student who was intended to be placed in one of three reference groups: advanced, regular or remedial. They compared several discriminant techniques, including: Fisher's discriminant analysis and kernel-based non-parametric discriminant analysis, using five school datasets. The study shows that a kernel-based nonparametric procedure performs better than Fisher's discriminant rule. Several other predictive models such as multiple regression, logistic regression, discriminant analysis, path analysis, factor analysis and Bayesian models have been created out of student data (Willett, 2002). Chiang, Yhang and Zhou (2006) conducted a comparative study of neural network and logistic regression to predict customer patronage behavior towards the web and traditional stores. They found in their study that in most of the selected products, neural networks significantly outperformed logistic regression in terms of its predictive power. Studies have also shown that in situations where neural networks have been used to model business problems, mostly in finance and marketing decision making, neural networks have outperformed traditional models such as discriminant and regression analysis. For example, Tuluca and Stalinski (2006) compared neural networks, discriminant analysis and logistic regression techniques to predict a firm's decision to list shares on a foreign stock exchange. They used a sample of 95 US manufacturing firms, including the

firms that are listed abroad and those that are listed only on the U.S. stock exchanges. The study showed that neural networks outperformed both discriminant analysis and logistic regression techniques. Similar studies were also done by Fadlalla and Lin (2001), Hung, Liang and Liu (1996) and West, Brockett and Golden (1997).

Classification algorithms have also been used in the medical field. One example is in gene classification where genes need to be classified based on their functionality. Several scientists and researchers have compared the outcome of predictive models using neural network and multivariable logistic regression analysis in many areas of health care. Li, Liu, Yang, and Chiu (1997) compared the performance of three mathematical models for surgical decisions on head injury patients. The study was performed using logistic regression, and two neural networks (a multi-layer perception MLP and radial-basis-function RBF). In the study they concluded that neural networks may be a better solution for complex, non-linear medical decision support systems than conventional statistical techniques such as logistic regression. Eftekhar et al. (2005) compared the performance of artificial neural network and multivariable logistic regression models in prediction of outcomes of head trauma. They concluded that artificial neural network (ANN) significantly outperformed the logistic regression in both discrimination and calibration. They found that the ANN lagged behind in terms of accuracy. Song et al. (2005), in their article entitled "Comparative Analysis of Logistic Regression and Artificial Neural Network for Computer-Aided Diagnosis of Breast Masses" found no difference in the performance between logistic regression and the artificial neural network as measured by the area under the ROC curve. However, they concluded that at 95% fixed sensitivity, the artificial neural network had a higher specificity compared with the logistic regression. Eng (2002) compared neural networks with one hidden layer and multivariate logistic regression on 1,064 patients who received an angiograghically based diagnosis of pulmonary embolism. The models were compared for accuracy in predicting the

presence or absence of pulmonary embolism on subsequent pulmonary arteriography. The objective of the study was to determine if neural networks would have better performance over conventional logistic regression. However, the results of the study showed no significant difference between the methods.

Chapter 3

RESEARCH BACKGROUND

The datasets chosen for the thesis is obtained by determining various software metrics for software - jajuk. The classification problem here is to classify the java classes in the software into defect prone and non-defect prone classes.

Software description

Jajuk is software that organizes and plays music. It is a full-featured application geared towards advanced users with large or scattered music collections. Using multiple perspectives, the software is designed to be intuitive and provide different visions of your collection.

There have been criticisms of the datasets used in conducting comparative studies. Some of the criticisms include: consideration of too few datasets, small dataset sizes, popularity of datasets, and outdated datasets (King, Feng, & Sutherland, 1995). In this thesis, the datasets were selected using the following criteria: number of records to class ratio, number of records in dataset, and class size.

Dataset description

The datasets chosen for the thesis is obtained by determining various software metrics for software - jajuk. The classification problem here is to classify the java classes in the software into defect prone and non-defect prone classes.

Dataset Statistics

Total Classes	328
No of defective classes	267
% of defective classes	81.40%

Table 1: Dataset Statistics

Attributes Description

S.No	Attribute	Description
1	Kind	Type of class [Public / Private]
2	Name	Name of the class
3	LOC	Lines of Code
4	CBO	Coupling between Object Classes CBO = number of classes to which a class is coupled Two classes are coupled when methods declared in one class use methods or instance variables defined by the other class. The uses relationship can go either way: both uses and used-by relationships are taken into account, but only once. Multiple accesses to the same class are counted as one access. Only method calls and variable references are counted. Other types of reference, such as use of constants, calls to API declares, handling of events, use of user-defined types, and object instantiations are ignored. If a method call is polymorphic (either because of Overrides or Overloads), all the classes to which the call can go are included in the coupled count. High CBO is undesirable. Excessive coupling between object classes is detrimental to modular design and prevents reuse.
5	WMC	Weighted Methods Per Class WMC = number of methods defined in class Keep WMC down. A high WMC has been found to lead to more faults. Classes with many methods are likely to be more more application specific, limiting the possibility of reuse. WMC is a predictor of how much time and effort is required to develop and maintain the class. A large number of methods also means a greater potential impact on derived classes, since the derived classes inherit (some of) the methods of the base class. Search for high WMC values to spot classes that could be restructured into several smaller classes.

6	NOC	Number of Children NOC = number of immediate sub-classes of a class NOC equals the number of immediate child classes derived from a base class. In Visual Basic .NET one uses the Inherits statement to derive sub-classes. In classic Visual Basic inheritance is not available and thus NOC is always zero. NOC measures the breadth of a class hierarchy, where maximum DIT measures the depth. Depth is generally better than breadth, since it promotes reuse of methods through inheritance. NOC and DIT are closely related. Inheritance levels can be added to increase the depth and reduce the breadth. A high NOC, a large number of child classes, can indicate several things: High reuse of base class. Inheritance is a form of reuse. Base class may require more testing. Improper abstraction of the parent class. Misuse of sub-classing. In such a case, it may be necessary to group related classes and introduce another level of inheritance. High NOC has been found to indicate fewer faults. This may be due to high reuse, which is desirable.
7	RFC	Response for a Class The response set of a class is a set of methods that can potentially be executed in response to a message received by an object of that class. RFC is simply the number of methods in the set. RFC = M + R (First-step measure) RFC′ = M + R′ (Full measure) M = number of methods in the class R = number of remote methods directly called by methods of the class R′ = number of remote methods called, recursively through the entire call tree A given method is counted only once in R (and R′) even if it is executed by several methods M. Since RFC specifically includes methods called from outside the class, it is also a measure of the potential communication between the class and other classes. A large RFC has been found to indicate more faults. Classes with a high RFC are more complex and harder to understand. Testing and debugging is complicated. A worst case value for possible responses will assist in appropriate allocation of testing time.
8	DIT	Depth of inheritance Tree DIT = maximum inheritance path from the class to the root class The deeper a class is in the hierarchy, the more methods and variables it is likely to inherit, making it more complex. Deep trees as such indicate greater design complexity. Inheritance is a tool to manage complexity, really, not to not increase it. As a positive factor, deep trees promote reuse because of method inheritance. A high DIT has been found to increase faults. However, it's not necessarily the classes deepest in the class hierarchy that have the most faults.
9	LCOM	Lack of cohesion of methods This metric measures the correlation between the methods and the local instance variables of a class. High cohesion indicates good class subdivision. Lack of cohesion or low cohesion increases complexity. Classes with low cohesion could probably be subdivided into two or more subclasses with increased cohesion. It is calculated as the ratio of methods in a class that do not access a specific data field, averaged over all data fields in the class.
10	ALTER	This indicates whether the java class is faulty or not

Table 2: Attributes Description of dataset

Chapter 4

RESEARCH METHODOLOGY

The following classification algorithms have been selected within the scope of this thesis and would be described briefly. They are:

- LogitBoost
- Bagging
- KStar
- RBFNetwork
- Logistic
- Naïve Bayes
- Bayes Net

There are two phases to this project: building the models (learning/training phase), and validating the models (using hold out data).

Weka has been used as the tool to build all the models. The models were evaluated using the following evaluation methods:

- Model accuracy,
- Precision,
- Recall/Sensitivity, and
- Specificity
- ROC Area

- F-Measure

The results of the implementations have been tabulated. A descriptive analysis was conducted to answer the research question.

Description for Algorithms

Next follow descriptions of the eight classification algorithms

LogitBoost

LogitBoost is a boosting algorithm formulated by Jerome Friedman, Trevor Hastie, and Robert Tibshirani. The original paper casts the AdaBoost algorithm into a statistical framework. Specifically, if one considers AdaBoost as a generalized additive model and then applies the cost functional of logistic regression, one can derive the LogitBoost algorithm.

Bagging

Bootstrap aggregating, also called bagging, is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. It also reduces variance and helps to avoid overfitting. Although it is usually applied to decision tree methods, it can be used with any type of method. Bagging is a special case of the model averaging approach. Given a standard training set D of size n, bagging generates m new training sets D_i, each of size n′, by sampling from D uniformly and with replacement. By sampling with replacement, some observations may be repeated in each D_i. If n′=n, then for large n the set D_i is expected to have the fraction (1 - 1/e) (≈63.2%) of the unique examples of D, the rest being duplicates.This kind of sample is known as a bootstrap sample. The m models are fitted using the above m bootstrap samples and combined by averaging the output (for regression) or voting (for classification). Bagging leads to "improvements for unstable procedures" (Breiman, 1996), which include, for

example, neural nets, classification and regression trees, and subset selection in linear regression (Breiman, 1994). An interesting application of bagging showing improvement in preimage learning is provided here. On the other hand, it can mildly degrade the performance of stable methods such as K-nearest neighbors (Breiman, 1996).

KStar

K* is a simple, instance based classifier, similar to K-Nearest Neighbor (K-NN). New data instances, x, are assigned to the class that occurs most frequently amongst the k-nearest data points, yj, where j = 1,2…k (Hart, 1968). Entropic distance is then used to retrieve the most similar instances from the data set. Using entropic distance as a metric has a number of benefits including handling of real valued attributes and missing values (Cleary and Trigg, 1995).

RBFNetwork

In the field of mathematical modeling, a radial basis function network is an artificial neural network that uses radial basis functions as activation functions. The output of the network is a linear combination of radial basis functions of the inputs and neuron parameters. Radial basis function networks have many uses, including function approximation, time series prediction, classification, and system control. They were first formulated in a 1988 paper by Broomhead and Lowe, both researchers at the Royal Signals and Radar Establishment.

Logistic

In statistics, logistic regression, or logit regression, is a type of probabilistic statistical classification model.[1] It is also used to predict a binary response from a binary predictor, used for predicting the outcome of a categorical dependent variable (i.e., a class label) based on one or more predictor variables (features). That is, it is used in estimating the parameters of a qualitative response model. The probabilities describing the possible outcomes of a

single trial are modeled, as a function of the explanatory (predictor) variables, using a logistic function. Frequently (and subsequently in this article) "logistic regression" is used to refer specifically to the problem in which the dependent variable is binary—that is, the number of available categories is two—while problems with more than two categories are referred to as multinomial logistic regression or, if the multiple categories are ordered, as ordered logistic regression.

Naïve Bayes

In machine learning, naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features. Naive Bayes is a popular (baseline) method for text categorization, the problem of judging documents as belonging to one category or the other (such as spam or legitimate, sports or politics, etc.) with word frequencies as the features. With appropriate preprocessing, it is competitive in this domain with more advanced methods including support vector machines.

Bayes Net

A Bayesian network, Bayes network, belief network, Bayes(ian) model or probabilistic directed acyclic graphical model is a probabilistic graphical model (a type of statistical model) that represents a set of random variables and their conditional dependencies via a directed acyclic graph (DAG). For example, a Bayesian network could represent the probabilistic relationships between diseases and s ymptoms. Given symptoms, the network can be used to compute the probabilities of the presence of various diseases. Formally, Bayesian networks are DAGs whose nodes represent random variables in the Bayesian sense: they may be observable quantities, latent variables, unknown parameters or hypotheses. Edges represent

conditional dependencies; nodes that are not connected represent variables that are conditionally independent of each other. Each node is associated with a probability function that takes as input a particular set of values for the node's parent variables and gives the probability of the variable represented by the node. For example, if the parents are m Boolean variables then the probability function could be represented by a table of 2^m entries, one entry for each of the 2^m possible combinations of its parents being true or false. Similar ideas may be applied to undirected, and possibly cyclic, graphs; such are called Markov networks.

Random Forest algorithm

Random forests are an ensemble learning method for classification (and regression) that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes output by individual trees. The algorithm for inducing a random forest was developed by Leo Breiman and Adele Cutler,and "Random Forests" is their trademark. The term came from random decision forests that was first proposed by Tin Kam Ho of Bell Labs in 1995. The method combines Breiman's "bagging" idea and the random selection of features, introduced independently by Ho and Amit and Geman in order to construct a collection of decision trees with controlled variance.

Application of algorithms

The algorithms are applied using WEKA. Feature Selection is applied as preprocessing technique and below attributes were found to be significant:

- LOC
- CBO

- LCOM
- ALTER

The models are generated based on the above attributes. After this, the confusion matrix was obtained and the evaluation methods were applied.

Model Evaluation techniques for Classification Algorithms

Once a classification model has been built, the next task is to know how well your model will perform. In this section will discuss the following evaluation methods and tools used for evaluating different classification algorithms: *classification accuracy, lift charts and gain charts, precision, recall/sensitivity and specificity*. All of these measures are based on the definition of a confusion matrix. An example of classification matrix for a binary classification is shown in Figure 6. Misclassification cost is not discussed within the scope of this thesis.

Classification Accuracy and Confusion Matrix

A confusion matrix (Kohavi and Provost, 1998) contains information about actual and predicted classifications done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix. The following table shows the confusion matrix for a two class classifier.

The entries in the confusion matrix have the following meaning in the context of our study:

- *a* is the number of **correct** predictions that an instance is **negative**,
- *b* is the number of **incorrect** predictions that an instance is **positive**,

- *c* is the number of **incorrect** of predictions that an instance **negative**, and
- *d* is the number of **correct** predictions that an instance is **positive**.

		Predicted	
		Negative	Positive
Actual	Negative	a	b
	Positive	c	d

Table 3: Confusion Matrix

Several standard terms have been defined for the 2 class matrix:

- The *accuracy* (*AC*) is the proportion of the total number of predictions that were correct. It is determined using the equation:

  AC = (a+d) / (a+b+c+d)

- The *recall* or *true positive rate* (*TP*) is the proportion of positive cases that were correctly identified, as calculated using the equation:

  TP = d/(c+d)

- The *false positive rate* (*FP*) is the proportion of negatives cases that were incorrectly classified as positive, as calculated usingthe equation:

  FP = b/(a+b)

- The *true negative rate* (*TN*) is defined as the proportion of negatives cases that were classified correctly, as calculated using the equation:

  TN = a/(a+b)

- The *false negative rate* (*FN*) is the proportion of positives cases that were incorrectly classified as negative, as calculated using the equation:

  FN = c/(c+d)

- Finally, *precision* (*P*) is the proportion of the predicted positive cases that were correct, as calculated using the equation:

  P = d/(b+d)

- F-Measure is a measure that combines precision and recall is the harmonic mean of precision and recall, the traditional F-measure or balanced F-score:

  F = 2 x (precision. recall) / (precision + recall)

- A receiver operating characteristic (ROC), or simply ROC curve, is a graphical plot which illustrates the performance of a binary classifier system as its discrimination threshold is varied. It is created by plotting the fraction of true positives out of the total actual positives (TPR = true positive rate) vs. the fraction of false positives out of the total actual negatives (FPR = false positive rate), at various threshold settings. TPR is also known as sensitivity or recall in machine learning. The FPR is also known as the fall-out and can be calculated as one minus the more well known specificity. The ROC curve is then the sensitivity as a function of fall-out. In general, if both of the probability distributions for detection and false alarm are known, the ROC curve can be generated by plotting the Cumulative Distribution Function (area under the probability distribution from -inf to +inf) of the detection probability in the y-axis versus the Cumulative Distribution Function of the false alarm probability in x-axis.

ROC analysis provides tools to select possibly optimal models and to discard suboptimal ones independently from (and prior to specifying) the cost context or the class distribution. ROC analysis is related in a direct and natural way to cost/benefit analysis of diagnostic decision making.

Chapter 5

RESULTS AND DISCUSSION

In this section the performance results of each algorithm on each dataset will be discussed and the research question will be addressed.

Experimental Results

Below are values of the evaluation parameters obtained by applying the discussed algorithms over *final_jajuk* dataset.

Evaluation Parameters for LogitBoost algorithm

Table 4 gives the values of evaluation parameters for LogitBoost algorithm. It can be observed that True pass rate and recall are highest for the algorithm and FP rate is lowest.

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
LogitBoost	82.9	58.3	80.4	82.9	80.6	0.822

Table 4 | Evaluation parameters for LogitBoost algorithm

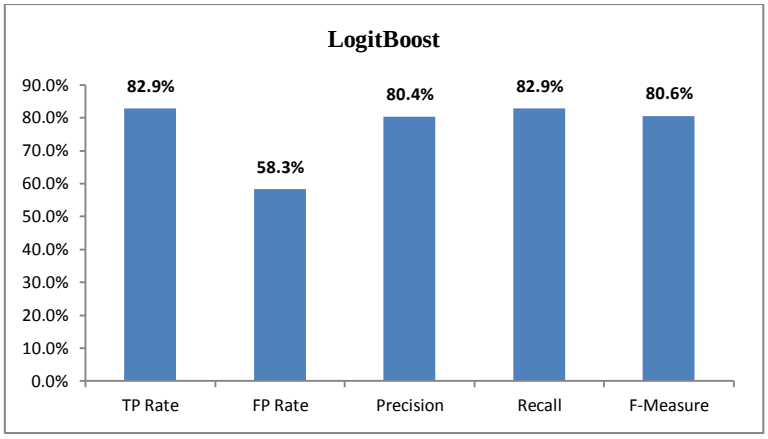

Fig1 | Evaluation parameters for LogitBoost algorithm

Evaluation Parameters for Bagging algorithm

Table 5 gives the values of evaluation parameters for Bagging algorithm. It can be observed that True pass rate and recall are highest for the algorithm and precision is lowest.

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Baggigng	81.4	81.4	66.3	81.4	73.1	0.498

Table 5| Evaluation parameters for Bagging algorithm

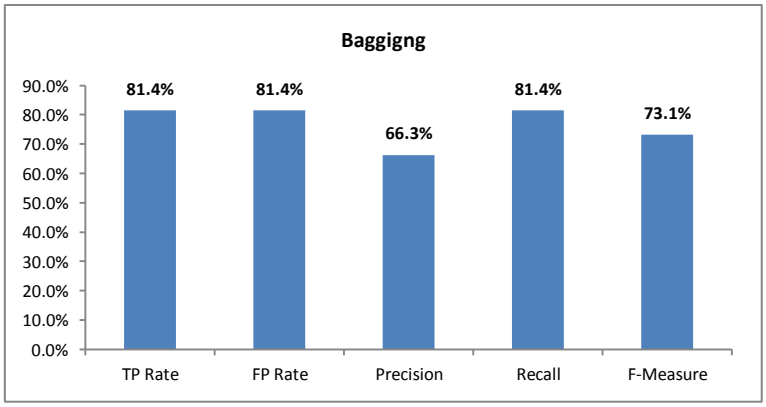

Fig 2 | Evaluation parameters for Bagging algorithm

Evaluation Parameters for KStar algorithm

Table 6 gives the values of evaluation parameters for KStar algorithm. It can be observed that True pass rate and recall are highest for the algorithm and FP rate is lowest.

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
KStar	79.9	56.5	77.9	79.9	78.7	0.828

Table 6| Evaluation parameters for KStar algorithm

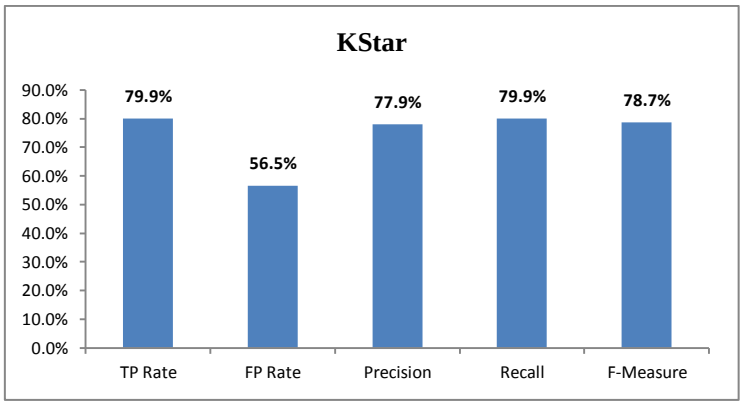

Fig 3| Evaluation parameters for KStar algorithm

Evaluation Parameters for RBFNetwork algorithm

Table 7 gives the values of evaluation parameters for RBFNetwork algorithm. It can be observed that precision is highest for the algorithm and FP rate is lowest.

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
RBFNetwork	78	45.5	79.2	78	78.6	0.745

Table 7| Evaluation parameters for RBFNetwork algorithm

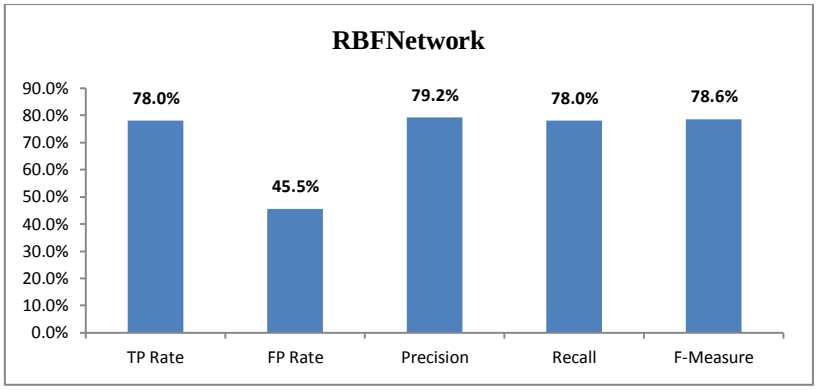

Fig 4 | Evaluation parameters for RBFNetwork algorithm

Evaluation Parameters for Logistic algorithm

Table 8 gives the values of evaluation parameters for Logistic algorithm. It can be observed that True pass rate is highest for the algorithm and F-Measure is lowest.

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Logistic	80.5	77.8	72.8	80.5	74.1	77.3

Table 8 | Evaluation parameters for Logistic algorithm

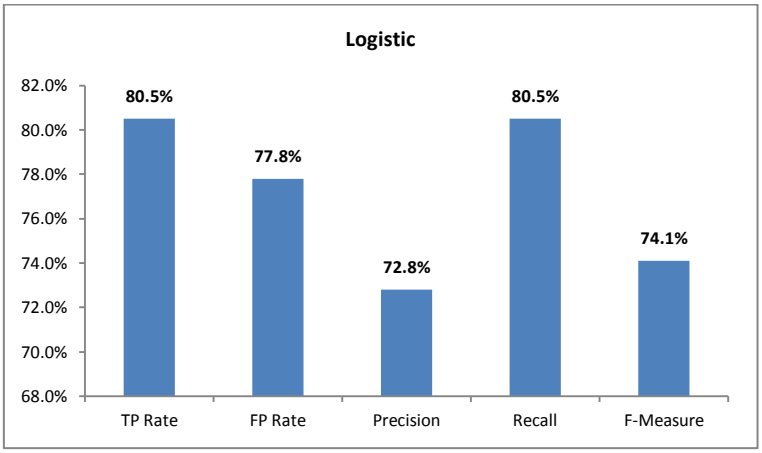

Fig 5 | Evaluation parameters for Logistic algorithm

Evaluation Parameters for Naïve Bayes algorithm

Table 9 gives the values of evaluation parameters for Naïve Bayes algorithm. It can be observed that precision is highest for the algorithm and FP rate is lowest.

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Naïve Bayes	54.3	21.8	81	54.3	58.6	0.751

Table 9| Evaluation parameters for Naïve Bayes algorithm

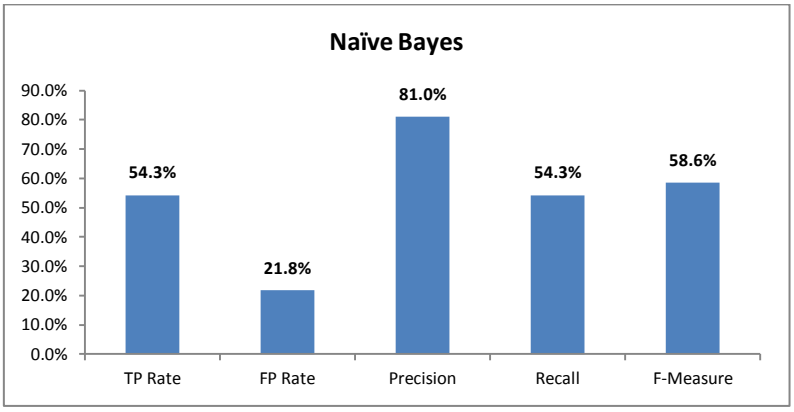

Fig 6 | Evaluation parameters for Naïve Bayes algorithm

Evaluation Parameters for Bayes Net algorithm

Table 10 gives the values of evaluation parameters for Bayes Net algorithm. It can be observed that F-Measure is highest for the algorithm and FP rate is lowest.

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Bayes Net	76.5	43.3	79	76.5	77.6	0.763

Table 10 | Evaluation parameters for Bayes net algorithm

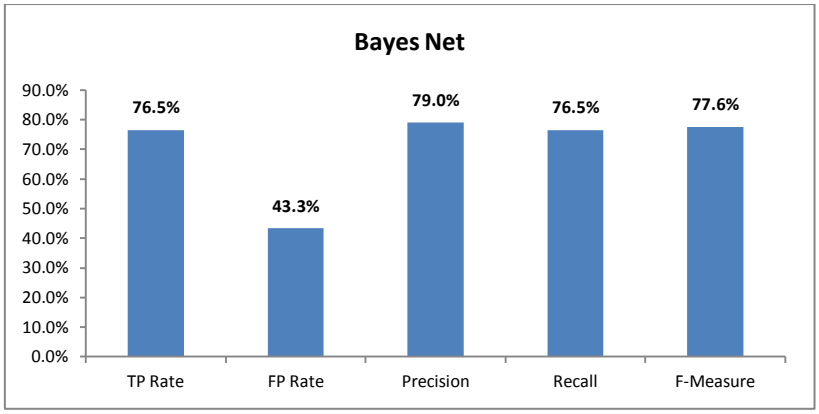

Fig 7 | Evaluation parameters for Bayes Net algorithm

Evaluation Parameters for Random Forest algorithm

Table 11 gives the values of evaluation parameters for Random Forest algorithm. It can be observed that True pass rate and recall are highest for the algorithm and precision is lowest.

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Random Forest	81.4	81.4	66.3	81.4	73.1	0.718

Table 11 | Evaluation parameters for Random Forest algorithm

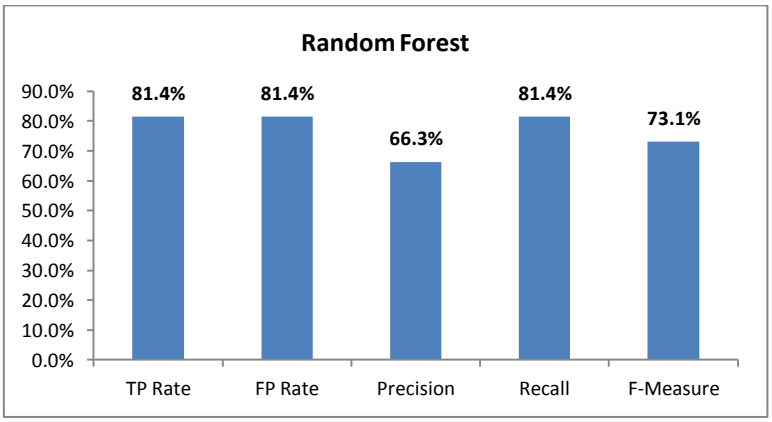

Fig 8 | Evaluation parameters for Random Forest algorithm

4.2.1 True Positive Rate

It can be observed that true positive rate for LogitBoot algorithm is the best. However, for Naïve bayes algorithm, true positive rate is the lowest. Random Forest Algorithm and Bagging Algorithm also fared well for the true positive rate.

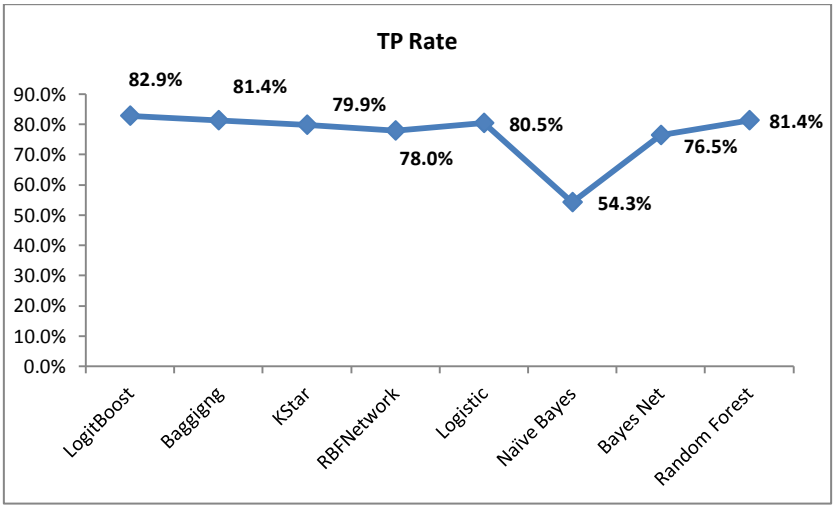

Fig 9 | TP rate over all algorithms

4.2.2 False Positive Rate

It can be observed that false positive rate for Bagging and Random Forest algorithm are the highest. However, for RBF Network algorithm, false positive rate is the lowest.

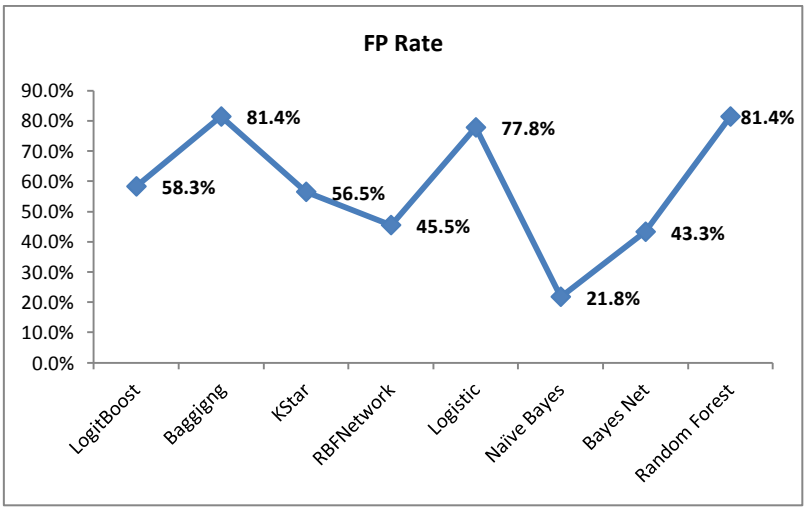

Fig 10 | FP rate over all algorithms

4.2.3 Recall

It can be observed that recall for LogitBoost algorithm the highest. However, for naïve bayes algorithm, recall is the lowest.

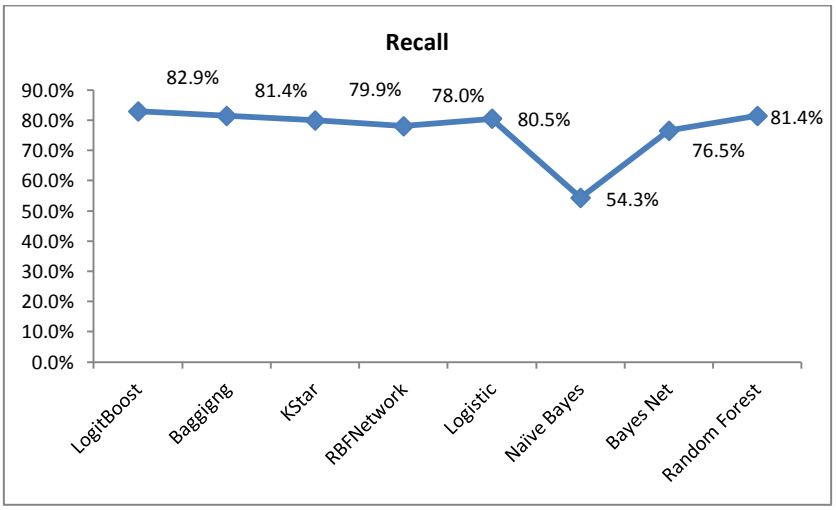

Fig 11 | Recall over all algorithms

4.2.4 Precision

It can be observed that false precision for Naïve Bayes algorithm are the highest. However, for Bagging and Random Forest algorithms, precision is the lowest.

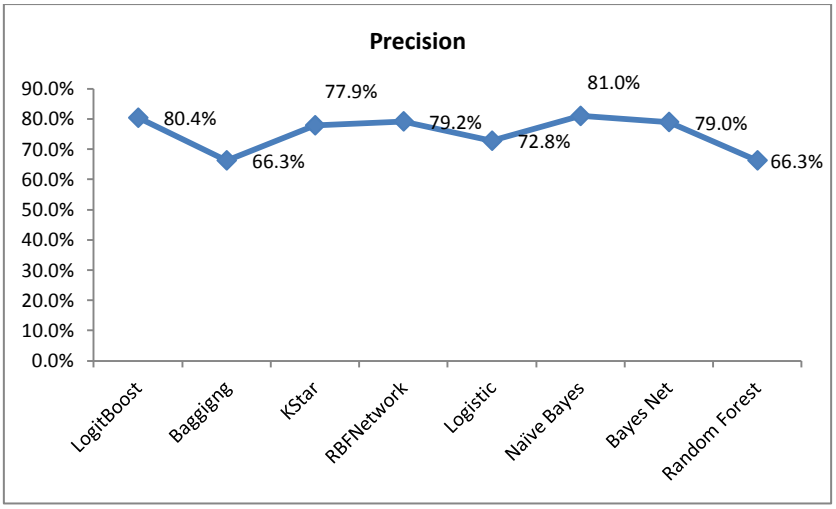

Fig 12 | Precision over all algorithms

4.2.5 F-Measure

It can be observed that F-Measure for LogitBoost algorithm is the highest. However, for Naïve Bayes algorithm, F-Measure is the lowest.

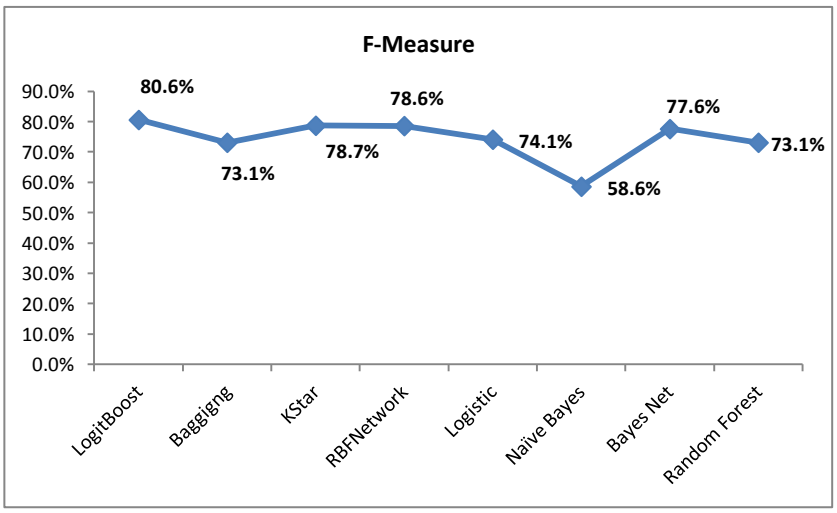

Fig 13 | F-measure over all algorithms

4.2.6 ROC Area

It can be observed that ROC area for KStart algorithm is the highest. However, for Bagging algorithm, ROC Area is the lowest.

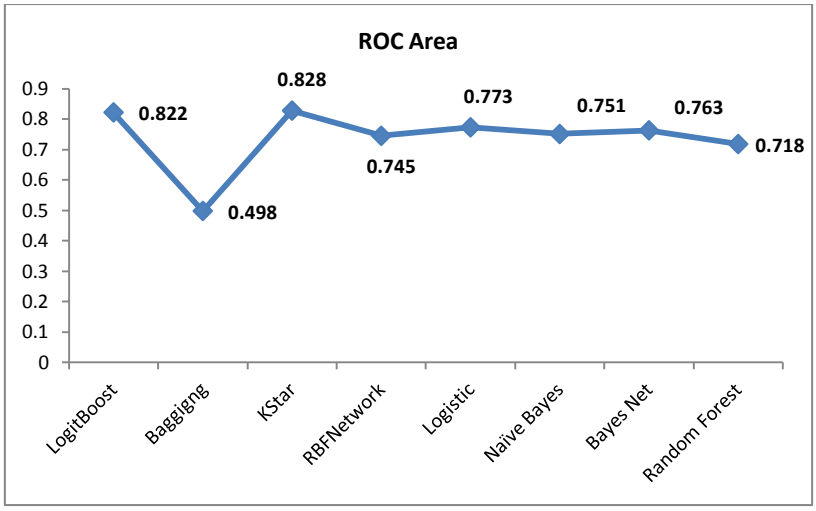

Fig 14 | ROC area over all algorithms

Chapter 6

CONCLUSION

As obtained in the experimental results, true positive rate for LogitBoot algorithm is the best. However, for Naïve bayes algorithm, true positive rate is the lowest. Random Forest Algorithm and Bagging Algorithm also fared well for the true positive rate. It was also observed that false positive rate for Bagging and Random Forest algorithm are the highest. However, for RBF Network algorithm, false positive rate is the lowest. Recall was found to be highest for LogitBoost algorithm. However, for naïve bayes algorithm, recall is the lowest. It can be observed that false precision for Naïve Bayes algorithm are the highest. However, for Bagging and Random Forest algorithms, precision is the lowest. F-Measure for LogitBoost algorithm is the highest. However, for Naïve Bayes algorithm, F-Measure is the lowest. ROC area is the highest for KStart algorithm. However, for Bagging algorithm, ROC Area is the lowest.

According to the experimental results, the Bayes Net algorithm proved to have the best performance. It performed better for all the performance parameters. RBF network and KStar also performed well. However, there is no universally best learning algorithm. From the analysis none of the algorithms outperformed the others in every problem. However, the results showed that the performance of each of the classification algorithm depends on what type of problem is being considered. The performance of classification algorithm also depends on the performance matrix and the characteristics dataset.

5.1 Limitations of the Thesis

The relationships between dataset characteristics and model accuracy were not discussed in this thesis. It is known that dataset characteristics influence the accuracy of classification algorithm and therefore this may influence the conclusion of the findings. Another limiting factor is the size of the datasets. Four out of the eight datasets have less than 2,000 instances. Research conducted by Shavlik, Mooney, and Towell (1991) showed that neural networks performs better when the training dataset is small.

Chapter 7

REFERENCES

Atlas, L., Connor, J., Park, D., El-Sharkawi, M., Marks, R., Lippman, A., Muthasamy, Y. (1991). A Performance Comparison of Trained Multi-layer Perceptions and Trained Classification Trees. *Systems, man, and cybernetics: proceedings of the IEEE international conference, ,* 915-920.

Berardi, V. L., Patuwo, B. E., & Hu, M. Y. (2004). A principled Approach for Building and Evaluating Neural Network Classification Models. *Decision Support Systems,* 233-246.

Bhattacharyya, S., & Pendharkar, P. C. (1998). Inductive, Evolutionary and Neural Computing Techniques for Discrimination: A Comparative Study. *Decision Sciences.*

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and Regression Trees. Wadsworth, Belmont.

Brown, D., Corruble, V., & Pittard, L. (1993). A Comparison of Decision Tree Classifiers with Backpropagation Neural Networks for Multimodal Classification Problems. *Pattern Recognition, 26,* 953-961.

Burges, C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery. Kluwer Academic Publishers. Boston 1998, 2.*

Caruana, R., & Niculescu-Mizil, A. (2006). An Empirical Comparison of Supervised Learning Algorithms. *Proceedings of the 23rd International Conference on Machine Learning, , .*

Chapelle, O. H. P, & Vapnik, V. N. (1999). Support Vector Machines for Histogram-Based Image Classification. *IEEE Trans Neural Networks, 10,* 1055-1064.

Chen, C. H. (1991). On the Relationship between Statistical Pattern Recognition and Artificial Neural Networks. *Neural Networks in Pattern Recognition and their Applications. New York, World Scientific*.

Chiang, W. K., Zhang, D., & Zhou, L. (2006). Predicting and Explaining Patronage Behavior towards Web and Traditional Stored using Neural Networks. *Decision Support Systems, 41,* 514-531.

Chung, H. M., & Silver, M. S. (1992). Rule-based Expert Systems and Linear Models: An Empirical Comparison of Learning-by-Examples Methods. *Decision Sciences.*

Cios, K., Pedrycz, W., & Swiniarski, R. (1998). *Data Mining Methods for Knowledge Discovery*. Norwell, MA: Kluwer Academic Publishers.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling Wine Preferences by Data Mining from Physicochemical Properties. *Decision Support Systems, Elsevier, 47*(4), 547-553.

Cristianini, N. (2001). *Support Vector and Kernel Machines* [PDF]. Retrieved from http://www.svms.org/tutorials/Cristianini2001.pdf

Curran, S. P., & Mingers, J. (1994). Neural Networks, Decision Tree Induction and Discriminant Analysis: An Empirical Comparison. *Journal of the Operational Research Society, 45*, 440-450.

Defries, R. S., & Chan, J. C-W (2000). Multiple Criteria for Evaluating Machine Learning Algorithms for Land Cover Classification from Satellite Data. *Remote Sensing of Environment, 74,* 503-515.

Dogan, N., & Tanrikulu, Z. (2010). A Comparative Framework for Evaluating Classification Algorithms. *Proceedings of the World Congress of Engineering.*

Eftekhar, B., Mohammad, K., Ardebili, H. E., Ghodsi, M., & Ketabchi, E. (2005). Comparison of Artificial Neural Network and Logistic Regression Models for Prediction of

Mortality in Head Trauma based on Initial Clinical Data. *BMC Medical Informatics and Decision Making, 5*(3) Doi: 10.1186/1472-6947-5-3

Eng, J. (2002). Predicting the Presence of Acute Pulmonary Embolism: A Comparative Analysis of the Artificial Neural, Logistic Regression, and Threshold Models. *AJR American Journal of Roentgenology, 179*(4).

Fadlalla, A., & Lin, C-H (2001). An Analysis of the Applications of Neural Networks in Finance. *Interfaces, 32*(4), 112-122.

Finch, W. H., & Schneider, M. K. (2006). Misclassification Rates for Four Methods of Group Classification. *Educational and Psychological Measurement, 66*(2), 240-257.

Fisher, D., & McKusick, K. (1989). An Empirical Comparison of ID3 and Backpropagation. *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence,* 788-793.

Frawley, W. J., Pietetsky-Shapiro, G., & Zimmerman, H. G. (1992). Knowledge Discovery in Databases: An Overview. *AI Magazine, 13,* 57-70.

Ge, E., Nayak, R., Xu, Y., & Li, Y. (2006). Data Mining for Lifetime Prediction of Metallic Components. *In Proc. Fifth Australian Data Mining Conference.*

Guszcza, J. (2005). *The Basics of Model Validation* [PowerPoint slides].

Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society, Series A, 160,* 523-541.

Henley, W. E., & Hand, D. J. (1996). A k-nearest-neighbor classifier for assessing consumer. *Statistician, 45,* 77-95.

Huang, Z., Chen, H., Hsu, C. J., & Chen, W. H., S. (2004). Credit Rating Analysis with Support Vector Machines and Neural Networks: A Market Analysis Comparative Study. *Decision Support System, 37*(4), 543-558.

Hung, S-Y, Liang, T-P, & Liu, V. W-C (1996). Integrating arbitrage Pricing Theory and Artificial Neural Networks to Support Portfolio Management. *Decision Support Systems, 18,* 301-316.

Ivanciuc, O. (2007). Applications of Support Vector Machines in Chemistry. *In: Reviews in Computational Chemistry, 23,* 291-400. Retrieved from http://www.ivanciuc.org/Files/Reprint/Ivanciuc_SVM_CCR_2007_23_291.pdf

Jakkula, V. Tutorial on Support Vector Machine (SVM). Retrieved August 15, 2011, from http://eecs.wsu.edu/~vjakkula/SVMTutorial.doc

Kass, G. V. (1980). Applied Statistics. , *29,* 119-127.

Kiang, M. Y. (2003). A Comparative Assessment of Classification Methods. *Decision Support System, 35,* 441-454.

King, R. D., Feng, C., & Sutherland, A. (1995). StatLog: Comparison of Classification Algorithms on a Large Real-World Problems. , , .

Klecka, W. R. (1980). *Discriminant Analysis. Quantitative Applications in the Social Sciences Series, No. 19*. Thousand Oaks, CA: Sage Publication.

Koh, H. C., Tan, W. C., & Goh, C. P. (2006). A Two-step Method to Construct Credit Scoring Models with Data Mining Techniques. *International Journal of Business and Information, 1*(1), 96-118.

Larose, D. T. (Ed.). (2006). *Data Mining Methods and Models*. Hoboken, NJ: Wiley.

Larose, D. T. (Ed.). (2005). *Discovering Knowledge in Data: An Introduction to Data Mining.* Hoboken, NJ: Wiley.

Last, M., & Maimon, O. (2002). A Compact and Accurate Model for Classification. *IEEE.*

Li, Y-C, Liu, L., Yang, T-F, & Chiu, W-T (1997). Comparing the Performance of Mathematical Models for Surgical Decisions on Head Injury Patients.

Li, Z., Liu, Y., Hayward, R., & Walker, R. (2010). Empirical Comparison of Machine Learning Algorithms for Image Texture Classification with Application t Vegetation Management in Power Line Corridors. *In Wagner W., Szekely, B (eds): ISPRS TC VII Symposium - 100 Years ISPRS, Vienna, Austria, XXXVIII* (Part 7A) Lim, T., Loh, W., & Shih, Y. (2000). A Comparison of Predictive Accuracy Complexity, and Training Time of Thirty-Three Old and New Classification Algorithms. *Machine Learning, 40*(3), 203-228. Kluwer Academic Publishers, Boston.

Lin, M., Huang, S., & Chang, Y. (2004). Kennel-Based Discriminant Technique for educational Placement. *Journal of Educational and Behavioral Statistics, 29,* 219-240.

Lu, Z., Szafron, D., Dreiner, R., Lu, P., Wishart, D. S., Poulin, B.,...Eisner (2003). Predicting Sub-cellular Localization of Proteins using Machine-Learned Classifiers. *Department of Computing Science University of Alberta, Edmonton, AB, Canada, T6G 2E8.*

Masuyama, T., & Nakagawa, H. (2002). Applying Cascade Feature Selection to Support Vector Text Categorization. *In TJOA, A.M and Wagner, R.R (eds.) Proceedings of the 13th International Workshop on Database and Expert Systems Applications, ,* 241-245.

McCallum, A., & Nigam, K. (1998). A Comparison of Event Models for Naive Bayes Text Classification. *Proceedings of Workshop on Learning for Text Categorization, American Association for Artificial Intelligence, , .*

Meshbane, A., & Morris, J. D. (1996). Predictive Discriminant Analysis vs. Logistic Regression in Two-Group Classification Problems. Paper presented at the meeting of the American Educational Research Association, New York.

Michie, D., Spiegelhalter, D. J., & Taylor, C. C (Eds.). (1994). *Machine Learning, Neural and Statistical Classification.*

Morris, J. D., & Lieberman, M. G. (2007). Achieving Accurate Prediction Models: Less is Almost Always More. *Multiple Linear Regression Viewpoints, 33*(2).

Mushtaq, S., Hameed, S., Asad, A., & Sheikh, L. M. (2006). Artificial Intelligence, Knowledge Engineering and Data bases. *5th WSEAS International Conference,* 354-359.

Neville, P. G. (1999). Decision Trees for Predictive Modeling. In (p. 20).

Nisbet, R., Elder, J., & Miner, G. (Eds.). (2009). *Handbook of Statistical Analysis and Data Mining Applications*.

Pesonen, E. (1997). Is Neural Network better than Statistical methods in diagnosis of Acute Appendicitis? *Medical Informatics Europe.*

Pohar, M., Blas, M., & Turk, S. (2004). Comparison of Logistic Regression and Linear Discriminant Analysis: A simulation Study. *Metodoloski Zvezki, 1,* 143-161.

Powers David, M. W. (2007). Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. *Flinders InfoEng Tech Rept, SIE-07-001*. Retrieved from http://www.infoeng.flinders.edu.au/research/techreps/SIE07001.pdf

Provost, F., Fawcett, T., & Kohavi, R. (1998). The Case against Accuracy Estimation for Comparing Induction Algorithms. *Proceedings of the Fifteenth International Conference on Machine Learning* Quinlan, J. R. (1994). Comparing Connectionist and Symbolic Learning Methods. *Computational Learning Theory and National Learning Systems: Constraints and Prospects, 1,*

Quinlan, R. (1998). *C5.0: An Informal Tutorial*. *Rulequest*. Retrieved from http://www.rulequest.com/see5-win.html

Ripley, B. D. (1994). Neural Networks and Related Methods for Classification. *Journal of the Royal Statistical Society, B, 56*(409-459).

Salzberg, L. S. (1997). On Comparing Classifiers: Pitfalls to Avoid and a Recommended Approach. *Data Mining and Discovery,* (1), 317-328.

Shavlik, J. W., Mooney, R. J., & Towell, G. G. (1991). Symbolic and Neural Learning Algorithms: An Experimental Comparison. *Machine Learning, 6,* 111-143.

Song, J. H., Venkatesh, S. S., Conant, E. A., Arger, P. H., & Sehgal, C. M. (2005). Comparative Analysis of Logistic Regression and Artificial Neural Network for Computer-Aided Diagnosis of Breast Masses. *Academic Radiology, 12*(4),.

Stalinski, P., & Tuluca, S. A. (2006). Determinants of Foreign Listing Decision: Neural Networks versus Traditional Approaches. *International Research Journal of Finance and Economics,* (4), . Retrieved from http://www.eurojournals.com/finance.htm

Suresh, S. S., & Balasaheb, M. P. (2008). Neural Network and Decision Tree Induction: A Comparison in the Domain of Classification of Sonar Signal. *First International Conference on Emerging Trends in Engineering and Technology,,* 595-600.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., Mclachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z-h., Steinbach, M., Hand, D. J., Steinberg, D (2008). Top 10 Algorithms in Data Mining. *Knowledge and Information System, 14,* 1-37. DOI: 10.1007/s10115-007-0114-2

West, P. M., Brockett, P. L., & Golden, L. L. (1997). A Comparative Analysis of Neural Networks and Statistical Methods for Predicting Consumer Choice. *Marketing Science, 16*(4),

Wikipedia online. Retrieved August 15, 2011, from http://en.wikipedia.org/wiki/Support_vector_machine

Willett, T. (2002). *Increasing persistence with an experimental intervention directed by data mining and statistical predictive models*. Pasadena CAIR conference.

Wolpert, D. H., & Macready, W. G. (1996). No Free Lunch Theorems for Optimization. http://axon.cs.byu.edu/~martinez/classes/678/Papers/Wolpert_NLFoptimization.pdf