A Comparative Study of Classification Algorithms for Cancer Datasets

Dissertation

Submitted To:

Central Department of Computer Science & Information Technology

Tribhuvan University

Kirtipur, Kathmandu

Nepal

In partial Fulfillment of the requirements for the Degree of Master of Science in Computer Science & Information Technology

Submitted by:

Ishwari Prasad Kandel

October, 2015

Supervisor

Prof. Shashidhar Ram Joshi (Ph.D)

Tribhuvan University

Institute of Science and Technology

Central Department of Computer Science and Information Technology

Student's Declaration

I hereby declare that I am the only author of this work and that no sources other than the listed here have been used in this work.

… … … … … … …
Ishwari Prasad Kandel
Date: 6th October, 2015

Tribhuvan University

Institute of Science and Technology

Central Department of Computer Science and Information Technology

Supervisor's Recommendation

I hereby recommend that the dissertation prepared under my supervision by **Mr. Ishwari Prasad Kandel** entitled "**A Comparative Study of Classification Algorithms for Cancer Datasets**" be accepted as in fulfilling partial requirement for the completion of Masters Degree of Science in Computer Science & Information Technology.

Prof. Dr. Shashidhar Ram Joshi

Department of Electronics & Computer Engineering,

Institute of Engineering,

Pulchowk, Nepal

Date: 6th October, 2015

Tribhuvan University

Institute of Science and Technology

Central Department of Computer Science and Information Technology

LETTER OF APPROVAL

We certify that we have read this dissertation work and in our opinion it is appreciable for the scope and quality as a dissertation in the partial fulfillment of the requirements of Masters Degree of Science in Computer Science & Information Technology.

Evaluation Committee

Asst. Prof. Nawaraj Paudel
Head of Department
Central Department of Computer Science
& Information Technology
Tribhuvan University
Kirtipur

Prof. Dr. Shashidhar Ram Joshi
(Supervisor)
Department of Electronics & Computer
Engineering,
Institute of Engineering,
Pulchowk, Nepal

(External Examiner)

(Internal Examiner)

Date: 6th October 2015

Acknowledgement

I would never have been able to finish my dissertation without the guidance, support and encouragement of numerous people including my supervisor, my friends, colleagues and support from my family. At the end of my thesis I would like to thank all those people who made this thesis possible and an unforgettable experience for me.

First, I would like to express my gratitude to my supervisor **Professor Dr. Shashidhar Ram Joshi,** Institute of Engineering, Pulchowk Campus for his continuous support without which the thesis wouldn't have been possible to complete. His suggestions, guidance, thorough knowledge and expertise helped me immensely in understanding and developing this thesis. I thank him immensely for his patience and generous time spent to guide me through the entire process.

Most importantly I would like to thank to respected Head of Department of Central Department of Computer Science and Information Technology, Asst. Prof. Nawaraj Paudel for his kind support, help and constructive suggestions. I am very much grateful and thankful to all the respected teachers Prof. Dr. Subarna Sakya, Mr. Dheeraj Kedar Pandey, Mr. Sarbin Sayami, Mrs. Lalita Sthapit, Mr. Arjun Singh Saud, Mr. Bikash Balami and Mr. Jagdish Bhatt for providing me such a broad knowledge and inspirations. I am so much thankful to Mr. Roshan Silwal for his continuous support throughout the thesis work and also like to thank Kumar, Shyam, Dinesh, Barun, Kamal, Nabin, Rajendra, Bhim, Lalit for their cooperation.

Special thanks to my family for their endless motivation, constant mental support and love which have been influential in whatever I have achieved so far. All my class fellows are worthy of my gratefulness for their direct or indirect support in completion of my dissertation. Finally, I would like to thank Mr Chiranjivi Sitaula for his guidance and kind co-operation during my whole work.

I have done my best to complete this research work. Suggestions from the readers are always welcomed, which will improve this work.

Abstract

Classification algorithms of data mining have been successfully applied in the recent years to predict cancer based on Micro-array Gene Expression Data. Various classification algorithms can be applied on such Micro-array Gene Expression Data to devise methods that can predict the occurrence of cancer.

In this study, Comparison of five different algorithms i.e. Voted Perceptron, LWL, DECORATE, Random Forest and RIDOR is presented. The main aim of this study is to evaluate the performance of those five algorithms for different cancer datasets with different dimensions. The datasets used for the study are chosen such a way that they differ in size, mainly in the terms of number of instances and number of attributes. When comparing the performance of all five algorithms, Random Forest is found to be the better algorithm in most of the cases.

Keywords- Micro-array, Gene Expression Data, Breast cancer, Lymphoma, Leukemia, bioinformatics, Voted Perceptron, DECORATE, RODOR, LWL

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviations	Full Form
RNA	Ribo Nucleic Acid.
mRNA	Messenger Ribo Nucleic Acid
DNA	Desoxy-ribo Nucleic Acid
LWL	Locally Weighted Learning
DECORATE	Diverse Ensemble Creation by Oppositional Relabeling of Artificial Training Examples
RIDOR	Ripple Down Rule learner
WEKA	Waikato Environment for Knowledge Analysis
SDK	Software development kit
IDE	Integrated development Environment
SWT	Standard Widget toolkit

Chapter 1

1. Introduction

1.1 Introduction

Classification is the process of sorting and categorizing data into various types, forms or any distinct class. Data classification enables the separation and classification of data according to dataset requirement for various objectives.

Classification is mainly focused on predicting group membership for data instances. It predicts categorical class level (discrete /nominal) and classifies data based on the training set and the values in a classifying attribute and uses it in classifying new data. It finally categorize and assigns class levels to a pattern set.

The goal of classification is to predict the value of a designated discrete class variable, given a vector of predictors or attributes [1]. In the age of bioinformatics, cancer data sets have been used for the cancer diagnosis and treatment than can improve human aging [2].

Cancer is a disease characterized by out-of-control cell growth, spread of abnormal cells and the capability of invade other tissues used by external or internal factors. There are more than 100 different types of cancer, and each is classified by the types of cell that is initially affected. Cancer harms the body when altered cells divide uncontrollably to form lumps or masses of tissue called tumors (except in case of Leukemia where cancer prohibits normal blood function by abnormal blood function by abnormal cell division in the blood stream). Tumors that stay in one spot and demonstrated limited growth are generally considered to be benign [3].

Gene expression profiling is a technique used in molecular biology to query the expression of thousands of genes simultaneously [4]. Gene expression analysis of cancer is used to regulatory gene defects and other devastating diseases, cellular responses to the environment, cell cycle variation, etc. When genes are expressed, the genetic information (base sequence) on DNA is first transcribed (copied) to a molecule named messenger RNA (mRNA). The mRNA molecules further participate in protein synthesis by specifying the particular amino acids that make up individual proteins. Gene Expression Analysis is one of the major applications of the Micro-array**.** Microarray is a hybridization of a nucleic acid sample (target) to a very large set of oligo-nucleotide probes, which are attached to a solid support (chip), to

determine sequence or to detect variations in a gene sequence or expression levels or for gene mapping [5].

In the recent years, several data mining classification algorithms are applied to classify cancer datasets to predict the presence of cancer. But, there is always a confusion to select the right algorithm. The comparative analysis of different algorithms over different cancer datasets will be helpful to overcome such confusions.

1.2 Problem definition

With the enormous amount of data stored in database, in is increasingly important to discover a robust, efficient and versatile data exploration technique. Data mining is the process inferring knowledge for such purpose which has major components like Pattern Recognition, Clustering, Association Rule and Classification. Classification has been identified as an important problem in the emerging field of data mining [5] as they try to find meaningful ways to interpret data sets. Classification of data is very typical task in data mining.

There are so many classification algorithms for data mining had been developed in past some years and also they have successful implementation. However, choosing the best algorithm for different datasets is always a challenging task as different algorithm can have different performance and accuracies with different type of datasets.

1.3 Objective

The objectives of this research are:

- To analyze classification algorithms.
- To trace algorithms.
- To compare performance of algorithms on different cancer datasets.

1.4 Motivation

The explosive growth of data leads to tremendous amount of data stored in database. However, those data are starving for knowledge. The discovery of knowledge form such data is very important job as in practical we have to make decision rapidly from data analysis with maximum knowledge.

In the recent decades, there is a dramatic changes in biometric research which contributes explosive growth of biomedical data like *micro array gene expression data.* Discovered

knowledge from such data can be applied for further research purpose. In case of cancer, the micro-array gene expression data of cancer contained cell can be used for future cancer prediction. Such data can be subjected to classification as a training set and that proper classified knowledge is very successful implementation to predict cancer in future.

On the other hand, recent progress in data mining research has led to the development of so many classification algorithms. Moreover, implementation of such classification algorithms on bio-medical data is one of the most emerging research area as more reliable prediction methodology to diagnose human disease is always in high demand for medical professionals. That is why, comparative analysis of such algorithms on medical data is very useful to figure out the proper selection of algorithms to specific datasets.

1.5 Thesis Organization

- Chapter 1 of this dissertation work is introduction part, which is organized into subsequent four chapters.
    - First chapter is focused on overview of Classification, Cancer disease and Micro-array Gene Expression data (cancer datasets).
    - Second chapter is about problem analysis of existing or previous works which demands further study to get better solutions.
    - Third chapter describes main objective of this dissertation work.
    - Fourth chapter is about motivation of this dissertation work.
- Chapter 2 contains explanation of all the previous studies related to this topic in detail under literature review.
- Chapter 3 includes details of all algorithms to be studied.
- Chapter 4 describes the implementation details.
- Chapter 5 contains all the details of data which is applied for analysis purpose and comparative performance measure of all five algorithms over collected cancer datasets. The result of the study is shown in tabular form as well as in graphs.
- Chapter 6 provides final conclusion and future works of the study.

Chapter 2

2. Background and Literature Review

2.1 Literature Review

The problem and scope of classification is one of the most widely studied topic in the field of data mining and machine learning communities. Classifiers are one of the technique of data mining which is applied to various domains to discover knowledge and improving decision making. Healthcare is the one of the biggest domain in which many researchers are involving to discover more efficient methods.

There are so many classifier algorithms developed since early 1960's to this date. Performance evaluation of such algorithms over cancer data domain is analyzed in different timeline. Certain research and comparative studies conducted earlier over cancer datasets before.

D.S.V.G.K. Kaladhar and B. Chandana [6] studied comparative analysis of CART, LMT, Random Forest, ADT, and Naive Bayesian algorithm for Breast Cancer. They concluded that Random Forest method has predicted better results than the other algorithms used for comparison. Absolute relative error of Random Forest is also found to be less than the absolute relative error of other algorithms.

Mohd Fauzi bin Othman,Thomas Moh Shan Yau [7] presented the comparison of different classification techniques using Waikato Environment for Knowledge Analysis or in short, WEKA. The aim of their paper is to investigate the performance of different classification or clustering methods for a set of large data. The algorithm or methods tested are Bayes Network, Radial Basis Function, Pruned Tree, Single Conjunctive Rule Learner and Nearest Neighbors Algorithm. In their study, they found the best algorithm based on the breast cancer data is Bayes Network Classifier with an accuracy of 89.71% and with execution time 0.19 second. Bayes Network Classifier has found average error of 0.2140 compared to others. They suggested that among all algorithms used for comparison, Bayes Network Classifier has the potential to significantly improve the conventional classification methods for use in bio-informatics field.

Rohit Arora and Suman [8] evaluated the performance in term of classification accuracy of J48 and multilayer Perceptron algorithms using various accuracy measures like TP rate, FP rate, Precision recall, F-measure and ROC Area. They measured accuracy of each datasets. In their evaluation, they found multilayer perceptron slightly better algorithm in most of the cases.

They conclude that algorithms based on neural network has better learning capability hence suited for classification problem if learned properly. They have also suggested some future work related to this topic. For the future work, more algorithms from classification can be incorporated and much more datasets should be taken or try to get the real datasets from different domain to have actual impact of the performance of algorithms taken into consideration.

Dursun Delen, Glenn Walker, Amit Kadam [9] researched several prediction models for breast cancer survivality. They used three popular data mining methods: two from machine learning (ANN, decision trees) and one from statistics (logistic regression). They have analyzed datasets along with a 10-fold cross validation with different data mining methods. Their study result indicates that the decision tree is the best prediction algorithm with 93.6% accuracy on the dataset applied, artificial neural network is come out to be the second with 91.2% accuracy and logistic regression model come out with 89.2% accuracy.

Gopala Krishna Murthy Nookala, Bharath Kumar Pottumuthu, Nagaraju Orsu Suresh B. Mudunuri [5] studied comprehensive comparative analysis of 14 different classification algorithms: A good mix of algorithms have been chosen from these groups that include Bayes Net & Naive Bayes (from Bayes), Multilayer Perceptron, Simple Logistics & SMO (from functions), IBk &KStar (from Lazy), NNge, PART & ZeroR (from Rules) and ADTree, J48, Random Forest and Simple Cart (from Trees). Their performance has been evaluated by using 3 different cancer data sets namely: Breast Cancer, Lymphoma and Leukemia. They found none of the classifiers outperformed all others in terms of accuracy when applied to all three datasets. Their result indicates that the performance of the classifier depends upon the datasets, especially on the number of attributes used in the dataset.

R.S. Michalski, I. Mozetic, J. Hong, and N. Lavrac [10] studied the incremental learning system AQ15 on some practical problem. For this they evaluated this incremental learning system on different three medical domains including breast cancer. They argue that the most important one is the classification accuracy of the induced rules on new objects. In this study, they had presented an experimental evaluation of the AQ15 program for learning from examples in therr medical domains: Lymphography, prognosis of breast cancer recurrence and location of primary tumor.

They had characterized those three domains consecutively larger amount of overlapping and sparse learning events. They took 70% of example randomly for rule learning and the rest for rule testing. They dis the experiments repeatedly to confirm the evaluation. The major contribution of this paper is to show that a relatively simple attribute based inductive learning method is able to produce decision rules of sufficiently high quality to be applicable to practical problems with noisy, overlapping and incompletely specified learning events. They gave some further research suggestion to find any given domain a rule reduction criterion that leads to the best trade-off between accuracy and complexity of a rule base.

Ash A. Alizadeh, Michael B. Eisen, R. Eric Davis, Chi Ma, Izidore S. Lossos, Andreas Rosenwald, Jennifer C. Boldrick, Hajeer Sabet, Truc Tran, Xin Yu, John I. Powell, Liming Yang, Gerald E. Marti, Troy Moore, James Hudson Jr, Lisheng Lu, David B. Lewis, Robert Tibshirani, Gavin Sherlock, Wing C. Chan, Timothy C. Greiner, Dennis D. Weisenburger, James O. Armitage, Roger Warnke, Ronald Levy, Wyndham Wilson, Michael R. Grever, John C. Byrd, David Botstein, Patrick O. Brown & Louis M. Staudt [11] conducted a systematic characterization of gene expression in B-cell malignancies using DNA microarrays. They showed that there is diversity in gene expression among the tumors of DLBCL patients, apparently reflecting the variation in tumor proliferation rate, host response and differentiation state of the tumor. They identified two molecularly distinct forms of DLBCL which had gene expression patterns indicative of different stages of B-cell differentiation. One type expressed genes characteristic of germinal centre B cells (germinal centre B-like DLBCL); the second type expressed genes normally induced during in vitro activation of peripheral blood B cells (activated B-like DLBCL). Patients with germinal centre B-like DLBCL had a significantly better overall survival than those with activated B-like DLBCL. The molecular classification of tumors on the basis of gene expression can thus identify previously undetected and clinically significant subtypes of cancer.

Their study shows that a genomic view of gene expression in cancer can bring clarity to previously muddy diagnostic categories. The classification scheme highlighted in their study divided DLBCCL on the basic of genes that are differentially expressed within the B-cell lineage.

T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, E. S. Lander [12] in their study

described generic approach to cancer classification based on gene expression monitoring by DNA microarrays and applied to human acute Leukemia as a test case. A class discovery procedure automatically discovered the distinction between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) without previous knowledge of these classes. An automatically derived class predictor was able to determine the class of new leukemia cases. The results demonstrate the feasibility of cancer classification based solely on gene expression monitoring and suggest a general strategy for discovering and predicting cancer classes for other types of cancer, independent of previous biological knowledge. They divided cancer classification into two challenges: class discovery and class prediction. Class discovery refers to defining previously unrecognized tumor subtypes. Class prediction refers to the assignment of particular tumor samples to already-defined classes, which could reflect current states or future outcomes. They focused their study mainly in three issues: The first one was the issue to explore whether there were genes whose expression pattern was strongly correlated with the class distinction to be predicted, the second one was how to use a collection of known samples to create a "class predictor" capable of assigning a new sample to one of two classes and the third one was how to test the validity of class predictors.

Yoav Freund and Robert E Schapire [13] introduced and analyzed a new algorithm for linear classification which combines Rosenblatt's perceptron with Helmbolt and Warmuth's leave-one-out method called Voted Perceptron. Their algorithm has advantage of data that are linearly separable with large margins. They compared their algorithms with Vapnik's algorithm and found to be much simpler to implement and much more efficient in term of computation time. They also showed that their algorithm their algorithm can be efficiently used in very high dimensional spaces using kernel functions. In addition, the theoretical analysis of the expected error of the perceptron algorithm yields very similar bounds to those of support vector machine. They also concluded that voting and averaging work better than just using the final hypothesis.

Peter Englert in his paper [14] presented the Locally Weighted Learning algorithm in details with other two different solution algorithms: Locally Weighted Regression and Locally Weighted Projection Regression. He did some successful application of Locally Weighted Learning in the field of Robot learning. He found Locally Weighted Regression is well suited for tasks that need a very accurate prediction and Locally Weighted Projection Regression is well suited for tasks with high-dimensional data, redundant input dimensions and continuous data streams. The biggest strength of LWPR is the combination of the high accuracy of the

prediction and the low computational costs through the model structure and the dimensionality reduction with PLS. Another advantage is the adaption over time, which is useful when the system changes over time. Finally, he also added his conclusion as the Locally Weighted Learning provides some powerful methods that are well suited for many different tasks and the results are comparable to current state of the are global function approximation methods.

Prem Melville and Raymond J. Mooney introduced a new method [15] for generating ensembles that directly constructs diverse hypotheses using additional artificially-constructed training examples. In their approach they present a new meta-learner that uses an existing "strong" learner (one that provides high accuracy on the training data) to build an effective diverse committee in a fairly simple, straightforward manner. This is accomplished by adding different randomly constructed examples to the training set when building new committee members. These artificially constructed examples are given category labels that *disagree* with the current decision of the committee, thereby easily and directly increasing diversity when a new classifier is trained on the augmented data and added to the committee. In this paper, they claim that DECORATE's chief advantage over Bagging and AdaBoost is the focus on maximizing diversity. In their conclusion they claimed that by manipulating artificial training examples, DECORATE is able to use a strong base learner to produce an effective, diverse ensemble. Their experimental results demonstrate that the approach is particularly effective at producing highly accurate ensembles when training data is limited, outperforming both bagging and boosting low on the learning curve. In general, the idea of using artificial or unlabeled examples to aid the construction of effective ensembles seems to be a promising approach worthy of further study.

Chapter 3

3. Algorithms Studied

In this dissertation, total five classification algorithms were studied for the analysis of cancer datasets.

3.1 Voted Perceptron

Voted Perceptron algorithm was introduced by Yoav Freund and Robert E. Schapire[13]. This is an algorithm for linear classification which is based on the well-known perceptron algorithm of Rosenblatt (1958, 1962) and a transformation of leave-one-out method of Helmbold and Warmuth(1995). In this algorithm, more information is stored during learning and then this elaborated information is used to generate better predictions on the test data.

The algorithm details of Voted-Perceptron algorithm is given below:

Training

Input: a labeled training set $\{(x_1,y_1),\ldots,(x_m,y_m)\}$

Number of epochs T

Procedure:

• Initialize: $k := 0$, $v_1 := 0$, $c_1 := 0$.

• Repeat T times:

– For i = 1,...,m:

* Compute prediction: $y' := \text{sign}(v_k \cdot x_i)$

* If

$y' = y$ then $c_k := c_k + 1$.

else

$v_{k+1} := v_k + y_i x_i$;

$c_k+1 := 1$;

$k := k + 1$.

Prediction

Given: the list of weighted perceptrons: $\{(v_1, c_1), \ldots, (v_k, c_k)\}$

an unlabeled instance: x

Compute a predicted label y' as follows:

$s = \sum_{i=1}^{k} c_i \text{ sign}(v_i \cdot x)$: $y' = \text{sign}(s)$

3.2 LWL (Locally Weighted Learning)

Locally Weighted Learning (LWL) [14] is the classic approach to solve the function approximation problem locally. It is also called Memory-Based Learning, because all training data is kept in memory to calculate the prediction. In this method, prediction is done by using an approximated local model around the point of interest.

The basic idea behind Locally Weighted Learning is that instead of building a global model for the whole function space, for each point of space a local model is created based on the neighboring data of query point. For this purpose, each data point becomes a weighting factor which express the influence of the data point for the prediction. In general, data points which are in the close neighborhood to the current query point are receiving a higher weight than data points which are far away. It is also called Lazy learning because the processing of the training data is shifted until a query point needs to be answered. This approach makes LWL a very accurate function approximation method where it is easy to add new training point.

The algorithm details of Locally Weighted Learning algorithm is given below:

Input:

Query point x_q

N training points $\{x_i, y_i\}$

Procedure:

Build matrix $X = (x_1, x_2, \ldots, x_n)^T$ Where $x_i = [x_i^T \ 1]^T$

Build vector $y = (y_1, y_2, \ldots, y_n)^T$

Compute diagonal weight matrix W:

$$W_{i,I} = \exp(\frac{-1\ (xi-xq)T\ D(xi-xq)}{2})$$

Calculate regression coefficient:

$$B_q = (X^T W X)^{-1}\ X^T W_y$$

Predict

$$Y_q = [x_i^T\ \ 1)\ B_q$$

3.3 DECORATE (Diverse Ensemble Creation by Oppositional Relabeling of Artificial Training Examples)

In DECORATE [15], an ensemble is generated iteratively, learning a classifier at each iteration and adding it to the current ensemble. We initialize the ensemble to contain the classifier trained on the given training data. The classifiers in each successive iteration are trained on the original training data and also on some artificial data. In each iteration a specified number of artificial training examples are generated from the data distribution. All training examples are generated from the data distribution. The labels for these artificially generated training examples are chosen so as to differ maximally from the current ensemble's predictions. The construction of the artificial data is explained in greater detail in the following section. We refer to the labeled artificially generated training set as the diversity data. We train a new classifier on the union of the original training data and the diversity data. If adding this new classifier to the current ensemble increases the ensemble training error, then we reject this classifier, else we add it to the current ensemble. This process it repeated until we reach the desired committee size or exceed the maximum number of iterations.

The algorithm details of DECORATE algorithm is given below:

Input:

BaseLearn- Base learning algorithm

T- set of *m* training example $<(x_1,y_1),\ldots (x_m,y_m)>$

With lebels y_i € Y

C_{size}- desired ensemble size

I_{max}- maximum number of iterations to build an ensemble

R_{size}- number of examples to generate at each iteration

Procedure:

i-1

trials=1

Ci=BaseLearn(T)

Initialize ensemble, C*={Ci}

Compute ensemble error, $\mathcal{E} = \frac{\sum xj \mathcal{E}\ T:C*(xj) \neq yi}{m}$

While i<Csize and trials<Imax

Generate Rsize training examples, R, based on distribution of training data

Label examples in R with probability of class lebels inversely proportional to C*'s predications

T=T$\cup$R

C'=BaseLearn(T)

C*=C*$\cup${ C'}

T=T-R, remove artificial data

Compute training error , $\mathcal{E}$

If $\mathcal{E}'$<= $\mathcal{E}$

I=i+1

$\mathcal{E}$= $\mathcal{E}'$

Else

C*=C*-{C'}

Trials=trials+1

3.4 Random Forest

A random forest is a collection of unpruned decision trees. It is typically made up of many decision trees. Each decision tree is built from a random subset of the training dataset. In building each decision tree model based on a different random subset of the training dataset a random subset of the available variables is used to choose how best to partition the dataset at each node. Each decision tree is built to its maximum size, with no pruning performed [16].

Input:

A Table with training set $\{(x_1,y_1),\ldots,(x_m,y_m)\}$

Procedure:

TreeGrowing (S,A,y)

Where:

S - Training Set

A - Input Feature Set

y - Target Feature

Create a new tree T with a single root node.

IF One of the Stopping Criteria is fulfilled THEN

Mark the root node in T as a leaf with the most

common value of y in S as a label.

ELSE

Find a discrete function f(A) of the input

attributes values such that splitting S

```
        according to f(A)'s outcomes (v1,...,vn) gains
        the best splitting metric.
IF best splitting metric > treshold THEN
        Label t with f(A)
        FOR each outcome vi of f(A):
        Set Subtreei= TreeGrowing (σf(A)=viS,A,y).
        Connect the root node of tT to Subtreei with
        an edge that is labelled as vi
        END FOR
ELSE
        Mark the root node in T as a leaf with the most
        Common value of y in S as a label.
        END IF
        END IF
        RETURN T
END IF
```

3.5 RIDOR

Ripple Down Rule learner (RIDOR) is a direct classification method. It constructs the default rule [17]. An incremental reduced error pruning is used to find exceptions with the smallest error rate, finding the best exceptions for each exception, and iterating. The most excellent exceptions are created by each exceptions produces the tree-like expansion of exceptions. The exceptions are a set of rules that predict classes other than the default.

Input:
A relational database D with target
Relation Rt that contains P positive and N negative tuples

Procedure:
Rule set R= empty
If $|R_t|$< MIN_SUP then return
Ruler=empty rule
Set R_t active
Repeat
Find a rule in active relation
Learn except branch and if not branch
Set relation of r to active
R=R+r
X=X-r
Until (X=NULL)
Set all active relations into inactive
Return R
End

Chapter 4

4. Implementation

4.1 Tools used

All the algorithms are implemented in Java language using NetBeans IDE 8.0 with the partial use of WEKA's libraries.

4.1.1 Programming language

For the implementation of proposed algorithm Java Programming Language is used. Java is a general-purpose, concurrent, class-based, object-oriented computer programming language that is specifically designed to have as few implementation dependencies as possible. One characteristic of Java is portability, which means that computer programs written in the Java language must run similarly on any hardware/operating-system platform. This is achieved by compiling the Java language code to an intermediate representation called Java bytecode, instead of directly to platform-specific machine code. Java bytecode instructions are analogous to machine code, but they are intended to be interpreted by a virtual machine written specifically for the host hardware. End-users commonly use a Java Runtime Environment installed on their own machine for standalone Java applications, or in a Web browser for Java applets.

Java is a robust language. It provides many safeguards to ensure reliable code. It has strict compile time and run time checking for data types. It is designed as a garbage-collected language ease the programmers virtually all memory management problems. Java also incorporates the concepts of exception handling which captures series errors and eliminates any risk of crashing the system.

4.1.2 NetBeans IDE

NetBeans is an integrated development environment for Java which contains base workspace and an extensible plug-in system for customizing the environment. NetBeans SDK is free and open source software mostly written in Java. The initial software development can extend its ability by installing plug-ins written for NetBeans Platform, such as development toolkits for other programming languages, and can write and contribute their own plug-in modules.

The NetBeans SDK includes the Eclipse Java development tools, offering an IDE with a built-in incremental Java compiler and a full model of the Java source files. This allows advanced

refactoring techniques and analysis. It provides the Rich client platform for developing general purpose applications.

4.1.3 WEKA Workbench

The WEKA workbench is a collection of state-of-the-art machine learning algorithms and data preprocessing tools [18]. It includes virtually all the ML algorithms. It provides extensive support for the whole process of experimental data mining, including preparing the input data, evaluating learning schemes statistically, and visualizing the input data and the result of learning. As well as a variety of learning algorithms, it includes a wide range of preprocessing tools. This diverse and comprehensive toolkit is accessed through a common interface so that its users can compare different methods and identify those that are most appropriate for the problem at hand.

WEKA was developed at the University of Waikato in New Zealand; the name stands for *Waikato Environment for Knowledge Analysis*. The system is written in Java and distributed under the terms of the GNU General Public License. It runs on almost any platform and has been tested under Linux, Windows, and Macintosh operating systems—and even on a personal digital assistant. It provides a uniform interface to many different learning algorithms, along with methods for pre- and post-processing and for evaluating the result of learning schemes on any given dataset.

Chapter 5

5. Data Collection and Result Analysis

5.1 Data Collection

Three different types of cancer datasets were collected from author of different previous papers. The datasets have been chosen such that they differ in size, mainly in terms of number of instances and number of attributes [5].

5.1.1 Data set 1:

The first data set is a small Breast Cancer Micro-array Gene Expression data used in an earlier study [10]. The data set contains 9 attributes apart from the class attribute with 286 instances.

5.1.2 Data set 2:

The second data set is a medium sized data set with Micro- array Gene Expression data of Lymphoma patients [11]. The data set has a total of 4,026 attributes and 45 instances.

5.1.3 Data set 3:

The large data set 3 is also a Micro-array Gene Expression data of Leukemia with 7,129 attributes and 34 instances [12].

5.2 Comparison Criteria

The comparative analysis or the result is made on the basis of the following criteria.

5.2.1 Execution Time

This is the actual Time taken to build model and classify the given dataset. The algorithm is run for 10 times and average execution time is taken as result.

5.2.2 Classification Accuracy

All classification result could have an error rate and from time to time it will either fail to classify correctly, or classify wrongly. So accuracy can be calculated as follows:

*Accuracy = (Instances Correctly Classified/Total number of Instances)*100%*

5.2.3 Mean Absolute Error

Mean absolute error is the average of the difference between predicted and actual value in all test cases; it is the average prediction error

$$MAE = (|a_1-c_1| + |a_2-c_2| + \ldots + |a_n-c_n|)/n \text{ (where } a=\text{actual output, } c=\text{expected output)}$$

5.2.4 Root Mean Square Error

Root mean square error is frequently used measure of differences between values predicted by a model or estimator and the values actually observed from the thing being modeled or estimated. It is just the square root of mean square error.

$$\text{Square root of } \{(a_1-c_1)^2 + (a_1-c_1)^2 + \ldots + (a_1-c_1)^2/n\}$$

5.3 Result Analysis

In this study, the accuracy of all five algorithms mentioned in chapter 4 is compared for the three different dimensional datasets mentioned in chapter 5.1 which is compared based on execution time, classification accuracy, mean absolute error and root mean square error. The result were achieved using whole data as its training set and also for prediction.

All the result sets of this study are mentioned below:

5.3.1 Comparison result of classifiers for dataset 1

Table 5-1 provides the summary output for comparison of all five algorithms studied over Dataset 1(i.e. Breast Cancer Micro-array Gene Expression data). Based on Table 5-1, we can clearly see that the Random Forest has the highest accuracy and the lowest is RIDOR, the DECORATE algorithm has the highest execution time and LWL has the lowest. It is discovered that the highest mean value error is found in Voted Perceptron and lowest is found in Random Forest. Moreover, RIDOR has the highest Root Mean Square Error and Random Forest has the lowest.

Classifier	Execution Time	Correctly Classified Instances	Incorrectly Classified Instances	Mean Absolute Error	Root Mean Square Error
Voted Perceptron	0.02	217(75.8741%)	69(24.1259%)	0.4443	0.4914
LWL	0.01	227(79.3706%)	59(20.6294%)	0.3353	0.3971
DECORATE	0.47	157(89.8601%)	29(10.1399%)	0.3655	0.3808

Random Forest	0.21	280(97.9021%)	6(2.0979%)	0.1412	0.1871
RIDOR	0.02	208(72.7173%)	78(27.2727%)	0.2727	0.5222

Table 5-1 : Result of Classifier of data set 1

The Result of Table 5-1 is partitioned into for several sub items for easier analysis and evaluation. As Classification accuracy, execution time and error rate.

From the Figure 5-1, we can clearly see that the higher accuracy belongs to the Random Forest with a value of 97.9021% followed by DECORATE with value 89.8601% and subsequently LWL, Voted Perceptron and RIDOR with the values of 79.3706%, 75.8741% and 72.7173%.

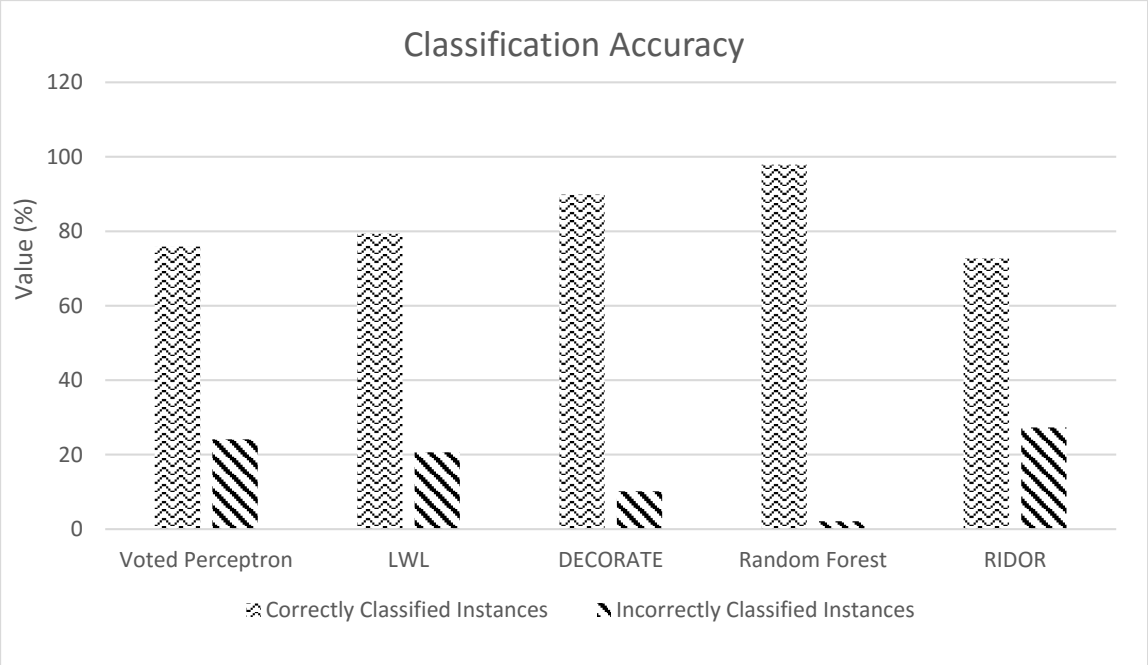

Figure 5-1 : Classification accuracy of dataset 1

In this study, from figure 5-2, we can say that LWL algorithm requires the shortest time which is around 0.01 seconds compared to others. DECORATE algorithm requires the longest model building time which is around 0.47 seconds. The second in the list is Voted Perceptron and RIDOR with same time value 0.02 seconds. The third in the list is Random Forest algorithm with 0.21 seconds.

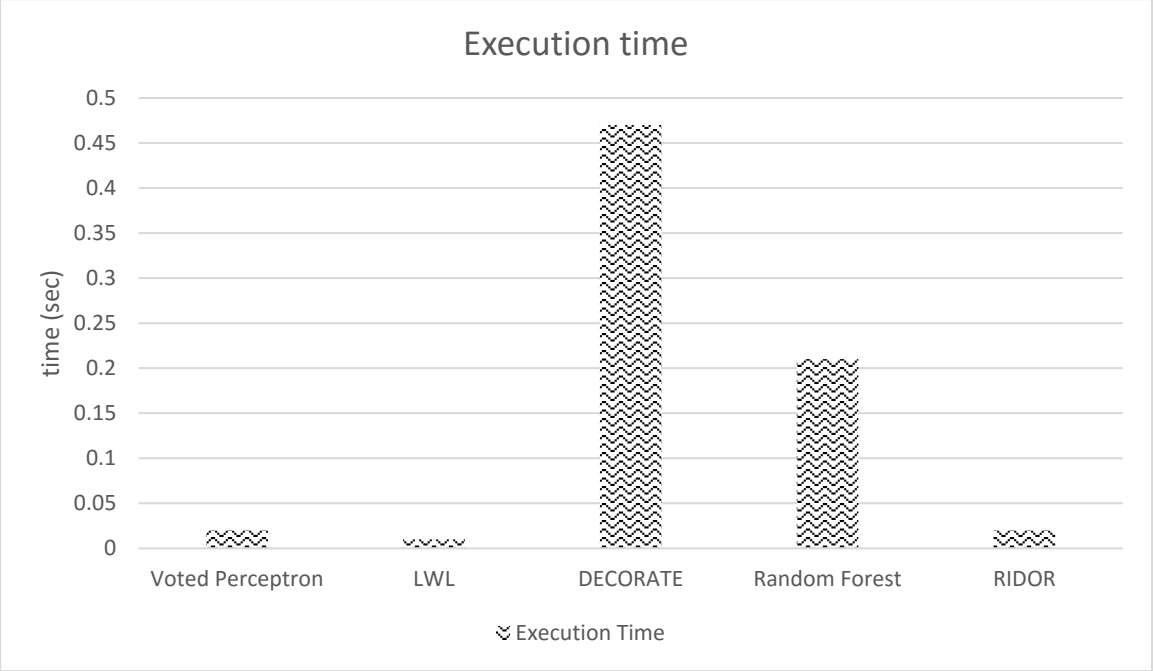

Figure 5-2 : Execution Time to build classification model of dataset 1

From Figure 5-3, it is discovered that lowest mean value error is found in Random Forest with value of 0.1871. RIDOR has the second lowest mean value error of 0.2727 followed by LWL, DECORATE and Voted Perceptron with values 0.3353, 0.3655 and 0.4443. Moreover, in the case of root mean square error, we observed that Random Forest gives the lowest root mean square error of 0.1871, second lowest value is 0.3808 of DECORATE algorithm, third is 0.3971 of LWL, fourth lowest 0.4914 of Voted perceptron and the highest error rate is 0.5222 which is belongs to RIDOR.

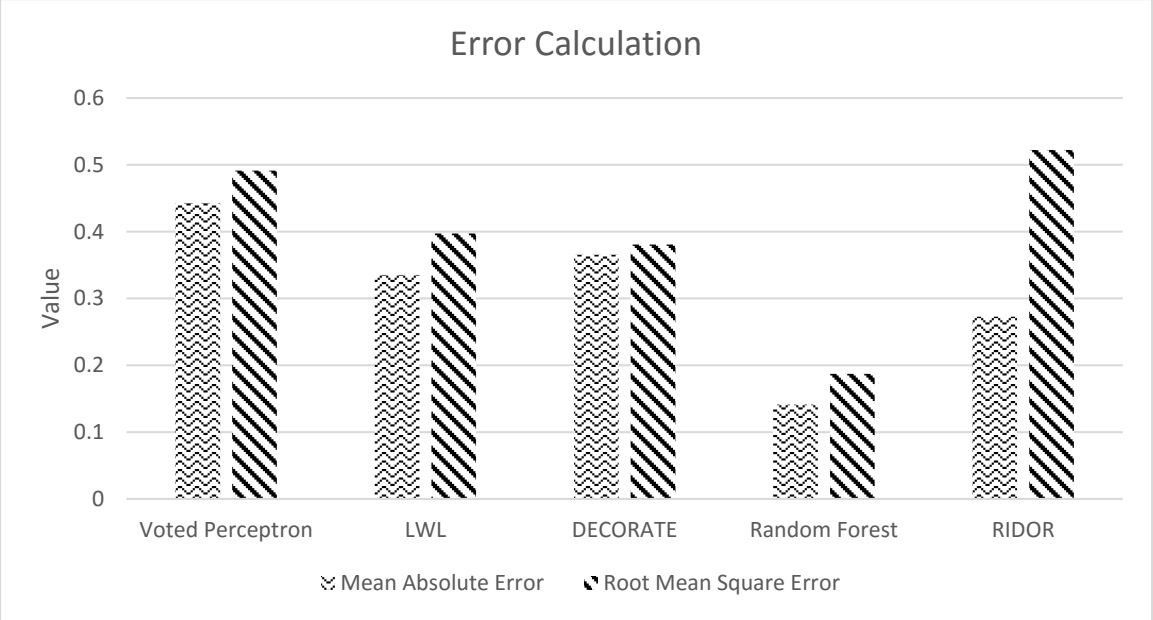

Figure 5-3 : Error calculation of dataset 1

5.3.2 Comparison result of classifiers for dataset 2

Table 5-2 provides the summary output for comparison of all five algorithms studied over Dataset 2 (i.e. Micro- array Gene Expression data of Lymphoma). Based on Table 5-2, we can clearly see that the Random Forest and DECORATE has the highest accuracy and the lowest is RIDOR, the DECORATE algorithm has the highest execution time and Voted Perceptron has the lowest. It is discovered that the highest mean value error is found in Random Forest and lowest is found in Voted Perceptron. Moreover, RIDOR has the highest Root Mean Square Error and DECORATE has the lowest.

Classifier	Execution Time	Correctly Classified Instances	Incorrectly Classified Instances	Mean Absolute Error	Root Mean Square Error
Voted Perceptron	0.16	44(97.7778%)	1(2.2222%)	0.0164	0.109
LWL	0.26	44(97.7778%)	1(2.2222%)	0.0695	0.1437
DECORATE	6.83	45(100%)	0(0%)	0.0243	0.0336
Random Forest	0.43	45(100%)	0(0%)	0.1342	0.1421

RIDOR	0.25	41(91.1111%)	4(8.8889%)	0.0889	0.2981

Table 5-2 : Result of Classifier of data set 2

The Result of Table 5-2 is partitioned into for several sub items for easier analysis and evaluation. As Classification accuracy, execution time and error rate.

From the Figure 5-4, we can clearly see that the higher accuracy belongs to the Random Forest and DECORATE with a value of 100% accuracy followed by Voted Perceptron and LWL with value 97.7778% and subsequently RIDOR with the least value of 91.1111%.

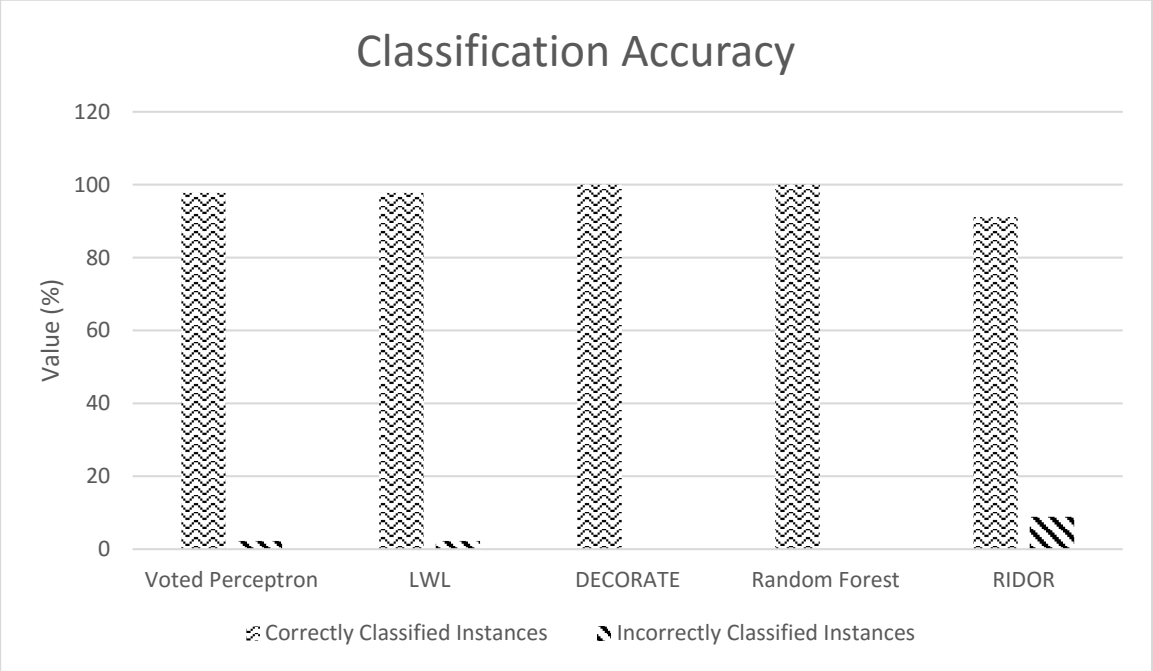

Figure 5-4 : Classification accuracy of dataset 2

In this study, from figure 5-5, we can say that Voted Perceptron algorithm requires the shortest time which is around 0.16 seconds compared to others. DECORATE algorithm requires the longest model building time which is around 6.83 seconds. The second in the list is RIDOR with time value 0.25 seconds. The third in the list is LWL algorithm with 0.26 seconds and fourth is Random Forest with value 0.43 seconds.

Figure 5-5 : Execution Time to build classification model of dataset 2

From Figure 5-6, it is discovered that lowest mean value error is found in Voted Perceptron with value of 0.0164. DECORATE has the second lowest mean value error of 0.0243 followed by LWL, RODOR and Random Forest with values 0.0695, 0.0889 and 0.1342. Moreover, in the case of root mean square error, we observed that DECORATE gives the lowest root mean square error of 0.0336, second lowest value is 0.109 of Voted Perceptron algorithm, third is 0.1421 of Random Forest, fourth lowest 0.1437 of LWL and the highest error rate is 0.2981 which is belongs to RIDOR.

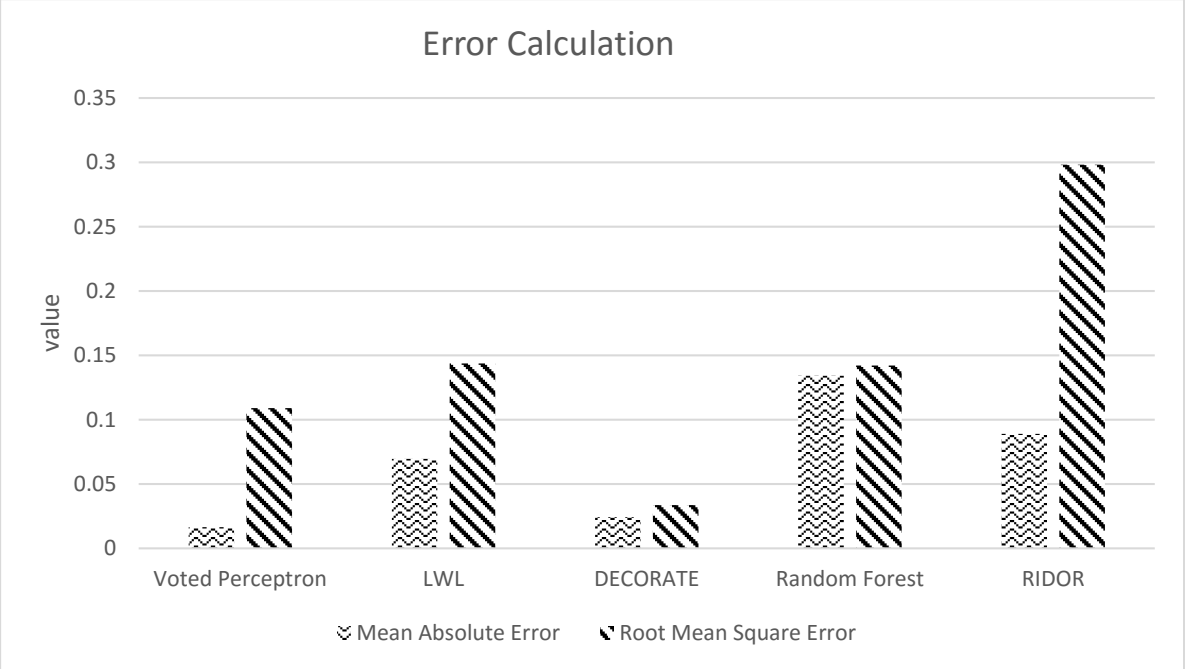

Figure 5-6 : Error calculation of dataset 2

5.3.3 Comparison result of classifiers for dataset 3

Table 5-3 provides the summary output for comparison of all five algorithms studied over Dataset 3 (i.e. Micro-array Gene Expression data of Leukemia). Based on Table 5-3, we can clearly see that the LWL, Random Forest and DECORATE has the highest accuracy and the lowest is Voted Perceptron, the DECORATE algorithm has the highest execution time and Voted Perceptron has the lowest. It is discovered that the highest mean value error is found in Voted Perceptron and lowest is found in LWL. Moreover, Voted Perceptron has the highest Root Mean Square Error and LWL has the lowest.

Classifier	Execution Time	Correctly Classified Instances	Incorrectly Classified Instances	Mean Absolute Error	Root Mean Square Error
Voted Perceptron	0.07	23(75.8741%)	11(24.1259%)	0.3329	0.56663
LWL	1.08	34(100%)	0(0%)	0	0
DECORATE	5.98	34(100%)	0(0%)	0.0104	0.0173

Random Forest	0.3	34(100%)	0(0%)	0.1279	0.1374
RIDOR	0.09	32(94.1176%)	2(2.0979%)	0.0588	0.2425

Table 5-3 : Result of Classifier of data set 3

The Result of Table 5-3 is partitioned into for several sub items for easier analysis and evaluation. As Classification accuracy, execution time and error rate.

From the Figure 5-7, we can clearly see that the higher accuracy belongs to the LWL, DECORATE and Random Forest with a value of 100% followed by RIDOR with value 94.1176% and subsequently LWL, Voted Perceptron with the values 75.8741%.

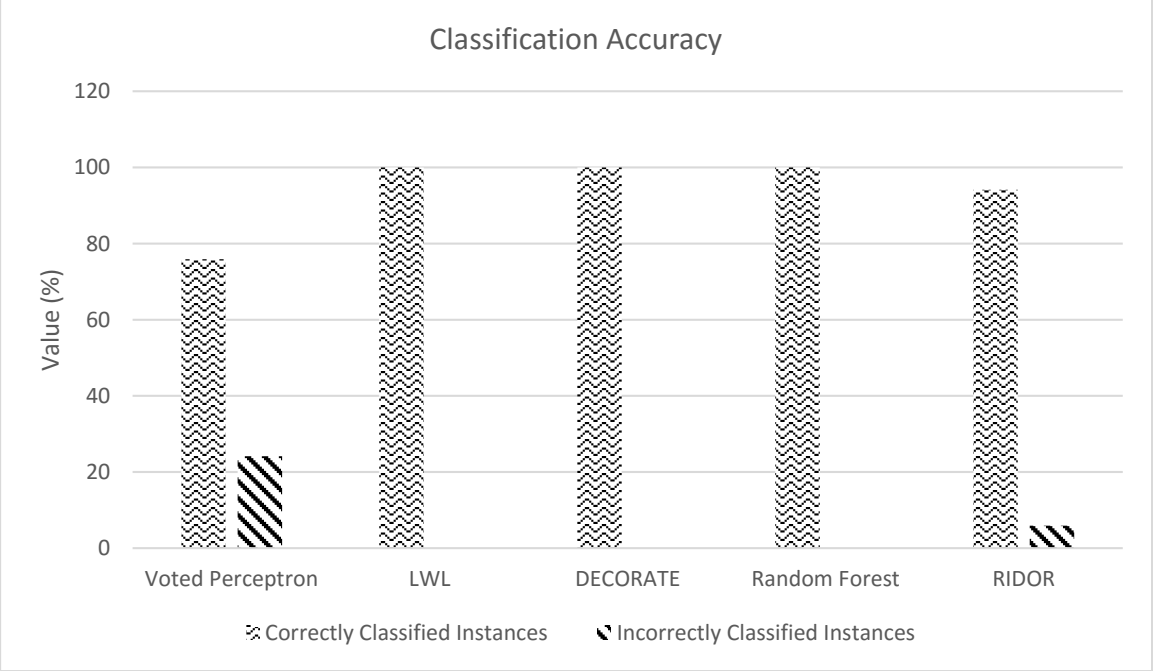

Figure 5-7 : Classification accuracy of dataset 3

In this study, from figure 5-8, we can say that Voted Perceptron algorithm requires the shortest time which is around 0.07 seconds compared to others. DECORATE algorithm requires the longest model building time which is around 5.98 seconds. The second in the list is RIDOR with time value 0.09 seconds. The third in the list is Random Forest algorithm with 0.3 seconds and fourth is LWL with 1.08 seconds.

Figure 5-8 : Execution Time to build classification model of dataset 3

From Figure 5-9, it is discovered that lowest mean value error is found in LWL with value of 0. DECORATE has the second lowest mean value error of 0.0104 followed by RIDOR, Random Forest and Voted Perceptron with values 0.0588, 0.1279 and 0.3329. Moreover, in the case of root mean square error, we observed that LWL gives the lowest root mean square error of 0, second lowest value is 0.0173 of DECORATE algorithm, third is 0.1374 of Random Forest, fourth lowest 0.2425 of RIDOR and the highest error rate is 0.5666 which is belongs to Voted Perceptron.

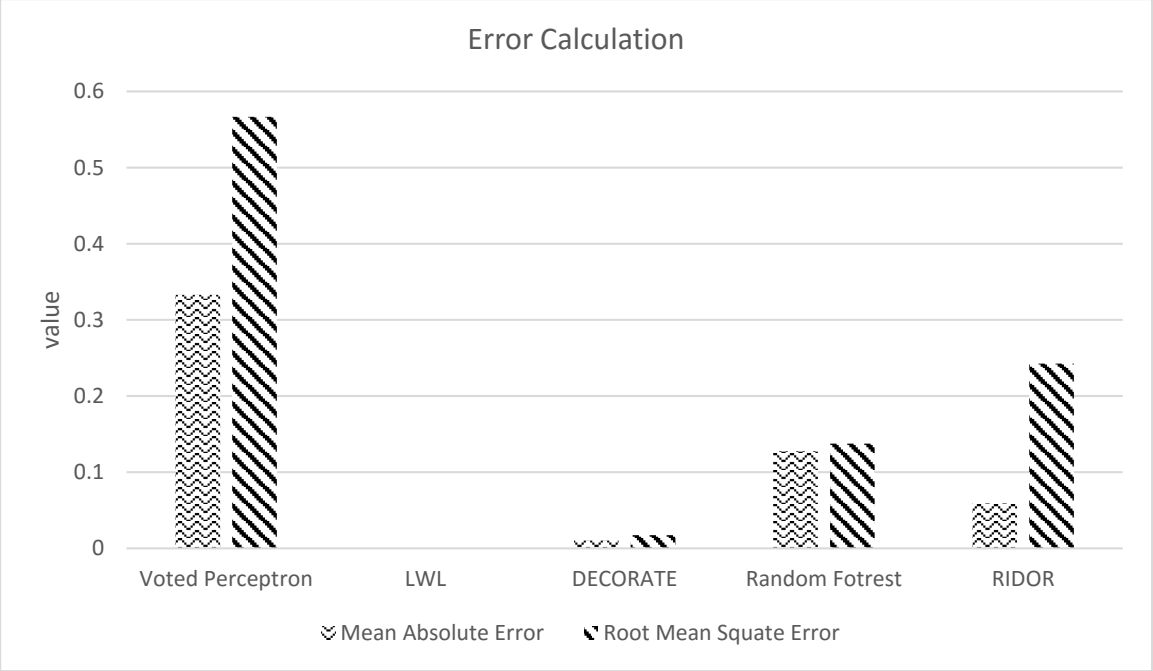

Figure 5-9 : Error calculation of dataset 3

5.3.4 Comparison result of classifiers for all datasets on the basis of Classification accuracy

Figure 5-10 provides the output for comparison of classification accuracy off all studied algorithm over all three datasets. From this figure we can clearly see that all the algorithms except Voted Perceptron has better accuracy with the increase in attribute of cancer datasets, whereas, voted perceptron has inconsistent result of accuracy with increase in attributes. Random forest has the better accuracy in all datasets.

Figure 5-10 : Classification accuracy of studied algorithms for all three datasets

5.3.5 Comparison result of classifiers for all datasets on the basis of Execution Time

From Figure 5-11, we can see that DECORATE algorithm has the longest model building time for all the datasets compared to other algorithms studied and has very large execution time difference. Voted Perceptron has the best average execution time for all algorithms. The average execution time to classify dataset 2 is larger than others except in the case of LWL where we can see LWL took more time to classify dataset 3 having largest number of attributes.

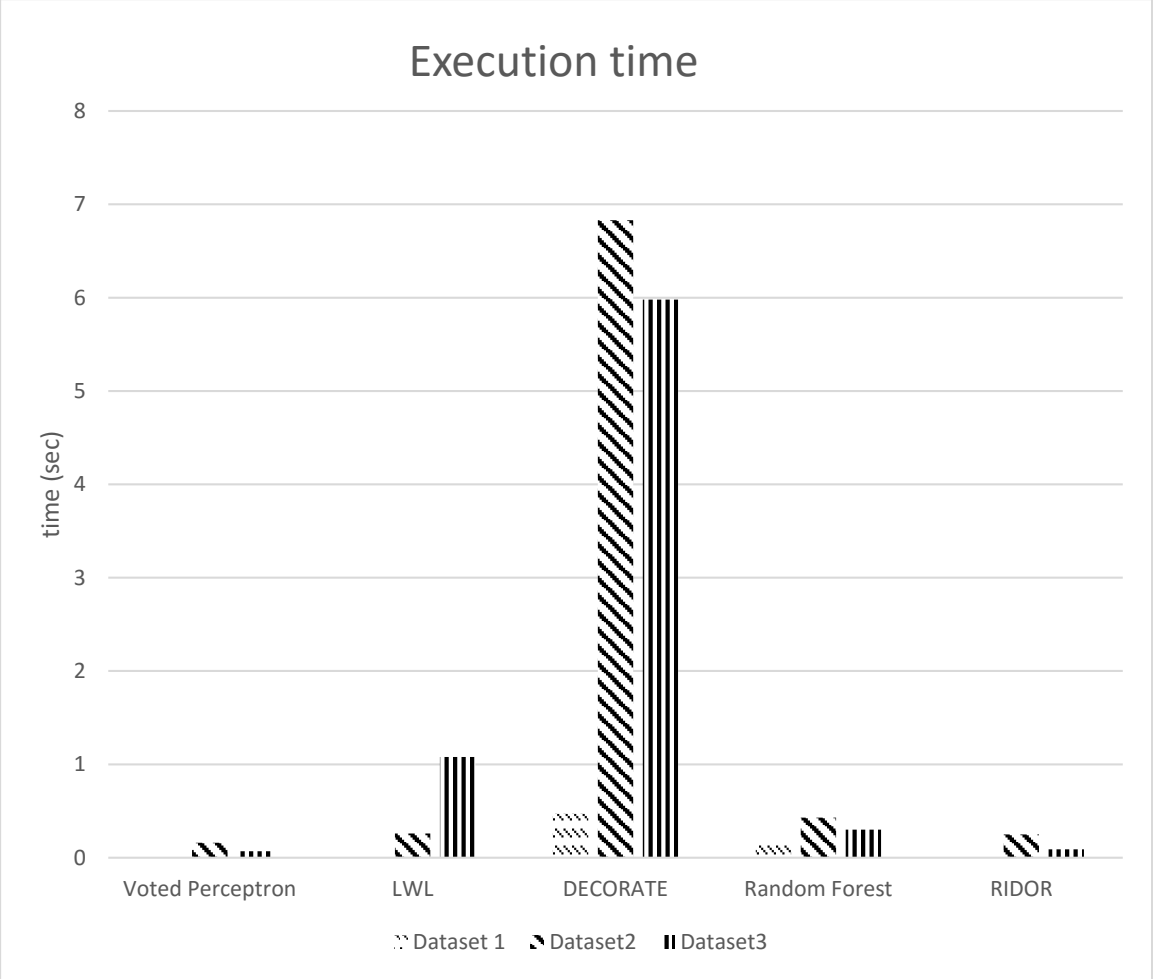

Figure 5-11 : Execution Time to build classification model of all three datasets

5.3.6 Comparison result of classifiers for all datasets on the basis of Mean Absolute Error

From figure 5-12, it is observed that Mean absolute error of all algorithms is higher for the dataset 1 and the lowest for the dataset 3 except in case of Voted Perceptron where error value of dataset 3 is greater than dataset is greater than the dataset 2. The highest generating algorithm is Voted perceptron for dataset 1 and dataset 3, however, in case of dataset 2 Random Forest generates higher error value. LWS has the negligible error value for dataset 3 as DECORATE. But in average analysis, Random forest generates average error rate for all the datasets.

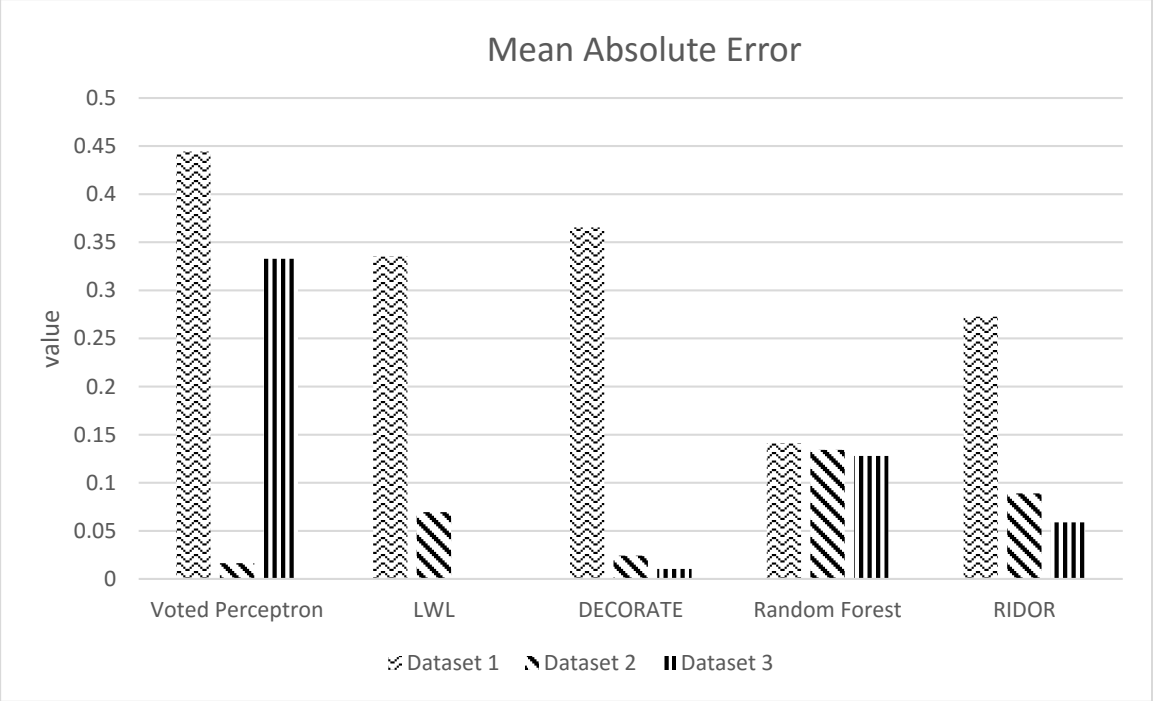

Figure 5-12 : Error calculation of classifiers for all three datasets

5.3.7 Comparison result of classifiers for all datasets on the basis of Root Mean Square Error

Based on figure 5-13, we can clearly that Root Mean Square error of all algorithms is closely same as the mean absolute error i.e. Higher for the dataset 1 and the lowest for the dataset 3 except in case of Voted Perceptron where error value of dataset 3 is the greatest. The highest generating algorithm is Voted perceptron for dataset 1 and dataset 3, however, in case of dataset 2 Random Forest generates higher error value. LWL has the negligible error value for dataset 3 as DECORATE. Random forest generates average error rate for all the datasets.

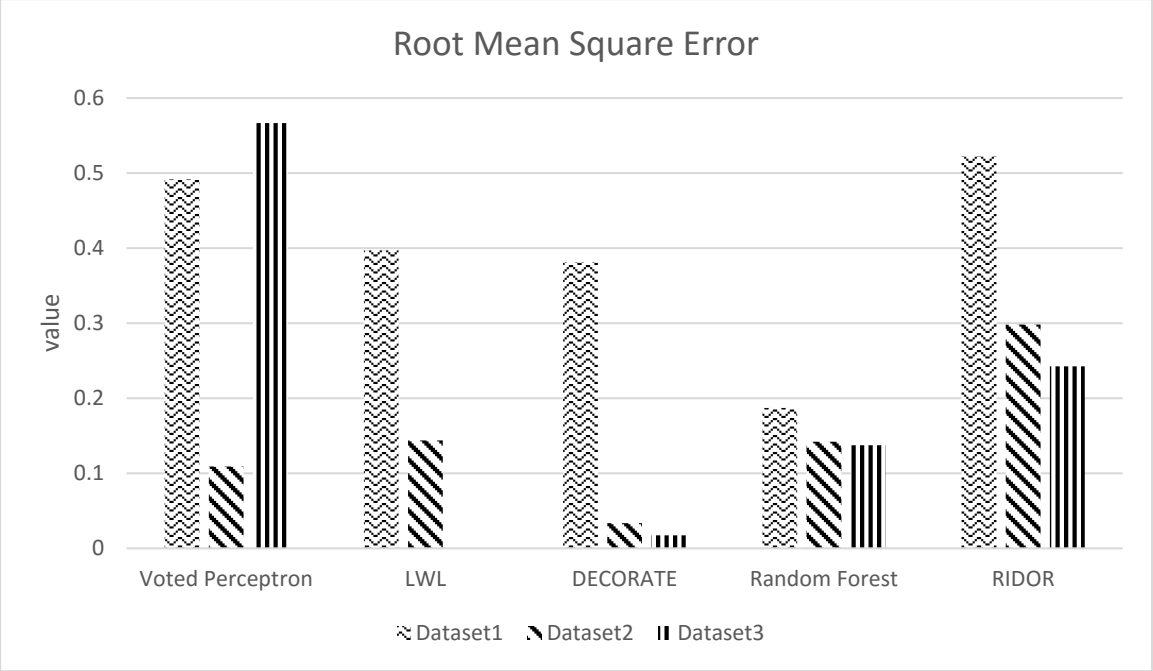

Figure 5-13 : Root mean square Error calculation of classifiers for all three datasets

Chapter 6

6. Conclusion and Future Works

6.1 Conclusion

In this study, the comparative study of classification algorithms (i.e. Voted Perceptron, LWL, DECORATE, Random Forest and RIDOR) using various measures like classification accuracy, Execution Time, Mean Absolute Error an Root Mean Square Error over three different cancer datasets with different dimensions are evaluated. The result suggest some conclusions. Though, Voted Perceptron has lowest Execution Time, it produces higher error rate with inconsistent accuracy. DECORATE has higher accuracy but Execution Time is drastically higher than other algorithms and also has bigger error rate over some datasets. LWL provides average accuracy with average execution time but error rate is very inconsistent over different datasets. RIDOR gives less accuracy value with higher error rate, however execution time is very less.

On balance scale, Random Forest algorithm has predicted better result than other algorithms studied for all datasets with consistent lower error rate. However, this algorithm has higher execution time than Voted Perceptron and RODOR.

6.2 Future Works

More algorithms from the classification can be incorporated for the future study to the studied datasets or other datasets which are associated with other human aging diseases. Moreover some algorithms can be customized for the specific domain so that the prediction could have more accurate and reliable.

References

[1] D. Grossman and P. Domingos. “Learning Bayesian Network Classifiers by Maximizing Conditional Likelihood”. *Proceedings of the 21st International Conference on Machine Learning*, Banff, Canada, 2004.

[2] Christoph Bock, Thomas Lengauer, “Computational epigenetics,” Bioinformatics, Vol. 24, No.1, pp. 1-10, 2008.

[3] Cancer: Types, Symptoms and Causes, http://www.medicalnewstoday.com/info/cancer-oncology, 2015

[4] Gene expression profiling, https://en.wikipedia.org/wiki/Gene_expression_profiling_in_cancer, 2015.

[5] Gopala Krishna Murthi Nookala, Bharath kumar Pottumuthu, Nagaraju Orsu, Suresh B Mudunuri , Performance Analysis And Evaluation of Different Data Mining Algorithms used for Cancer Classification, (IJARAI) International Journal of Advanced Research in Artificial Intelligence, Vol. 2, No 5, 2013.

[6] D.S.V.G.K. Kaladhar and B. Chandana “Data mining, inference and prediction of Cancer datasets using learning algorithms, International Journal of Science and Advanced Technology, Volume 1 No 3 May 2011.

[7] M. Fauzi, B. Othman,T M S Yau. "Comparison of different classification techniques using WEKA for breast cancer." 3rd Kuala Lumpur International Conference on Biomedical Engineering 2006. Springer Berlin Heidelberg, 2007.

[8] R. Arora and Suman "Comparative Analysis of Classification Algorithms on Different Datasets using WEKA." Int. Journal of Computer Applications 54.13, 2012.

[9] D. Delen, G Walker, and A. Kadam. "Predicting breast cancer survivability: a comparison of three data mining methods." Artificial intelligence in medicine, 34.2: 113-128, 2005.

[10] R.S. Michalski, I. Mozetic, J. Hong, and N. Lavrac, “The Multi-Purpose Incremental Learning System AQ15 and its Testing Application to Three Medical Domains”. In *Proceedings of the Fifth National Conference on Artificial Intelligence*, 1041-1045.

[11] A.A. Alizadeh, B. M.B. Eisen, R.E. Davis, C.M., *et. al* “Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.” Nature, Vol 403, No. 3, pp. 503-511, 2000.

[12] T.R. Golub, D. K. Slonim, P. Tamayo, C. Huard *et al*, "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring" *Science*, Vol. 286, pp. 531-537, 15 October 1999.

[13] Yoav Freund, Robert E. Schapire, Large Margin Classification Using the Perceptron Algorithm, 1999.

[14] Peter Englert, Locally Weighted Learning, TU Darmstadt.

[15] Raymond J. Mooney, Prem Melville, Constructing Diverse Classifier Ensembles using Artificial Training Examples, Texas University.

[16] Graham Williams,DATA MINING: Desktop Survival Guide,Random Forests, 2010.

[17] V. VeeraLaxmi, Dr. D. Ramyachitra, Ripple Down Rule learner (RIDOR) Classifier for IRIS Dataset, International journal of computer Science and engineering(IJCSE), 2015.

[18] *http://www.cs.waikato.ac.nz/ml/WEKA/*

Bibliography

1. Sushmita Mitra and Tinku Acharya, *"Data Mining: Multimedia, Soft Computing, and Bioinformatics"*
2. Richard O. Duda, Peter E. Hart and David G. Stork *"Pattern Classification", 2nd Edition.*
3. I.H. Witten, E. Frank, and M.A. Hall*, "Data mining practical machine learning tools and techniques", 3rd Edition.*
4. J. Han and M. Kamber, *"Data mining concepts and techniques", 2nd Edition.*
5. O. Maion and L. Rokach*, "Data mining and Knowledge discovery handbook", Springer, 2nd Edition.*
6. P. Harrington, *"Machine learning in action", Manning.*
7. R. Jensen and Q. Shen, *"Computational Intelligence and Feature selection".*
8. S. Marshall, *"Machine learning an algorithmic perspective", CRC Press.*

Appendix

(Sample Data)

1. **Sample data of dataset 1 (Breast Cancer Micro-array Gene Expression data)**

@relation breast_cancer

@attribute class { no-recurrence-events, recurrence-events}

@attribute age {10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99}

@attribute menopause {lt40, ge40, premeno}

@attribute tumor-size {0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59}

@attribute inv-nodes {0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-39}

@attribute node-caps {yes, no}

@attribute deg-malig {1, 2, 3}

@attribute breast {left, right}

@attribute breast-quad {left_up, left_low, right_up, right_low, central}

@attribute irradiat {yes, no}

@data

no-recurrence-events,60-69,lt40,10-14,0-2,no,1,left,right_up,no

no-recurrence-events,50-59,ge40,25-29,0-2,no,3,left,right_up,no

no-recurrence-events,40-49,premeno,30-34,0-2,no,3,left,left_up,no

no-recurrence-events,60-69,lt40,30-34,0-2,no,1,left,left_low,no

no-recurrence-events,40-49,premeno,15-19,0-2,no,2,left,left_low,no

no-recurrence-events,50-59,premeno,30-34,0-2,no,3,left,left_low,no

no-recurrence-events,60-69,ge40,30-34,0-2,no,3,left,left_low,no

no-recurrence-events,50-59,ge40,30-34,0-2,no,1,right,right_up,no

no-recurrence-events,50-59,ge40,40-44,0-2,no,2,left,left_low,no

no-recurrence-events,60-69,ge40,15-19,0-2,no,2,left,left_low,no

no-recurrence-events,30-39,premeno,25-29,0-2,no,2,right,left_low,no

no-recurrence-events,50-59,premeno,40-44,0-2,no,2,left,left_up,no

no-recurrence-events,50-59,premeno,35-39,0-2,no,2,right,left_up,no

no-recurrence-events,40-49,premeno,25-29,0-2,no,2,left,left_up,no

2. **Sample data of dataset 2 (Micro-array Gene Expression data of Lymphoma Cancer)**

@relation DLBCL

@attribute 'GENE1835X ' numeric

@attribute 'GENE1836X ' numeric

@attribute 'GENE1865X ' numeric

…

…

@attribute 'GENE3120X ' numeric

@attribute 'GENE48X ' numeric

@attribute 'GENE47X ' numeric

@attribute Class {germinal,activated}

@data

-0.31,-0.4,-0.46,-0.45,-0.09,0.26,0.05,-0.16,0.25,-0.34,-0.05,-0.03,-0.24,-0.47,-0.36,0.55,-0.07,-0.32,0.59,0.94,-0.17,0.54,-0.21,-0.24,0.19,-0.69,0.37,0.35,0.2,-0.44,-0.17,0.17,0.25,0.9,0.97,0.81,0.17,1.47,1.7,2.33,2.58,2.83,2.41,0.55,0.92,1.01,0.61,0.39,0.34,0.35,0.46,-0.06,0.27,-0.02,0.06,0.56,0.34,-0.19,0.33,-1.07,-0.52,-0.07,0.36,0.13,-0.28,-1.03,-0.13,-0.41,0.01,-0.25,1.27,1.05,0.82,0.27,-0.72,0,0.34,0.05,1.65,1.53,-1.04,0.77,0.5,0.46,-0.27,-1.28,-1.17,-0.78,-0.11,-0.07,-0.09,-0.08,0.27,0.14,0.01,0.58,0.67,0.66,0.74,0.57,0.59,0.42,1.86,0.41,0.3,0.23,0.78,1.11,-0.19,-0.37,0.04,0.43,0.3,0.19,0.43,0.42,0.54,0.06,-0.11,-0.22,0.58,0.74,0.28,0.42,0.11,0.21,0.15,0.32,-0.09,

…

…

0.48,0.07,0.24,0.05,1.13,1.23,1.95,0.08,0.06,0.38,0.44,1.13,0.17,0.02,0.09,0.12,0.09,-0.32,-0.37,0.12,0.08,0.09,0.09,0.13,-0.11,0.43,0.1,0.45,0.19,-0.34,-0.93,-0.03,0.35,-0.15,-0.72,0.24,0.44,0.03,0.64,0.76,0.15,1.03,0.56,0.05,0.19,0.46,0.82,0.13,-0.41,-0.25,0.04,0.62,0.77,1.25,-0.19,0.8,0.31,0.14,0.25,-0.17,-0.26,-0.01,0.2,0.2,-0.11,0,0.2,-0.78,-0.67,-0.31,-0.39,-0.99,-0.9,0.59,-0.2,0.03,-0.33,-0.04,-0.16,-0.12,-0.21,0.03,0.1,-0.28,0.01,-0.67,-0.01,-0.1,germinal

3. Sample data of dataset 2 (Micro-array Gene Expression data of Lymphoma Cancer)

@relation AMLALL

@attribute attribute1 numeric

@attribute attribute2 numeric

@attribute attribute3 numeric

..

..

@attribute attribute7128 numeric

@attribute attribute7129 numeric

@attribute Class {ALL,AML}

@data

-342,-200,41,328,-224,-427,-656,-292,137,-144,48,-591,-622,-342,294,241,-7,-108,45815,57,422,-185,-48,-181,-4,-132,-2,115,41,-50,-202,113,-557,-687,-289,-195,135,267,57,-238,-337,6339,5199,9045,19541,27768,24477,149,19,229,6418,27707,456,622,72,267,-70,-299,193,241,-38,251,98,-260,149,275,2573,168,1151,-885,-883,-106,-685,147,-328,-77,301,111,1016,-60,190,1233,207,0,43,767,195,1214,

..

..

1163,-72,-154,596,28,516,675,190,1805,3104,494,108,123,-31,-16,38,-364,248,123,395,168,79,-31,19,-28,-118,1561,-98,-125,-374,-1209,453,318,533,67,197,125,-337,-33,12,171,127,1209,103,656,5083,695,591,-53,893,-1627,388,-152,335,-67,284,7,41,-176,388,-251,313,280,16,55,3719,1214,835,127,60,287,596,444,1054,917,-265,772,301,840,3782,716,-354,-977,403,-354,-277,-38,1511,156,-224,8540,183,-16,2754,4,36,6022,7159,893,-953,55,-36,900,-181,-36,-135,-103,-4,610,41,1581,-1226,260,137,-463,289,432,2218,449,526,19,663,661,977,31,-21,-388,41,-72,48,9885,977,-45,17858,50,70,-195,70,94,494,330,183,142,-9,-7,-84,-24,429,856,815,429,-605,-2,603,381,2435,20818,12869,835,388,-118,16456,12103,451,3239,-352,41,547,-50,156,41,19,323,420,231,246,533,-101,-451,2112,277,1023,67,214,-135,1074,475,48,168,-70,ALL