APPENDIX – A

PUBLISHED PAPERS
Paper 1


Publication Type: International Journal
Year of Publishing: 2013
Publisher: Published by Foundation of Computer Science, New York, USA
Complexity-Reduced Tumor Classification System using Microarray Gene Expression Dataset

N.Gopala Krishna Murthy  
Associate Professor  
Department of CSE, 
G.V.V.R.Institute of Technology, 
Andhra Pradesh

O.NagaRaju  
Head of the Department,  
Department of Computer Science, 
Government Degree College,  
Andhra Pradesh

Allam Appa Rao  
Director, CR Rao Advanced Institute of Mathematics, Statistics & Computer Science (AIMSCS),  
University of Hyderabad Campus,  
Andhra Pradesh

ABSTRACT
The classification of cancer based on gene expression data is the advancements in DNA Microarray technology and genome sequencing. The important feature is to predict the genes for various diagnosis purposes using such micro-array gene expression dataset and also the gene expressions that are relevant to a particular type of genes. Lot of research works have been carried out to produce a better solution to improve the prediction accuracy of cancer gene prediction. But the analysis results are not up to the convincing level artificial intelligence is exploited to improve the prediction accuracy meanwhile state-of-the-art insists necessary enhancements which are essential in the classification module instead of the features module. The enhanced classifier called Principal component analysis used in latter researches is used in this work for the performances comparison of the classifier as a conventional prediction methodology.

This work intends to apply the developed classifier and dominant gene prediction methodology to predict extensive set of cancer expression datasets. The experimental study will be carried out by considering the techniques on CNS tumor, colon tumor and ALL_AML Leukemia. The prediction performance of the proposed methodology will be compared against the conventional prediction methodologies and the results will be validated extensively. The method will be implemented in the working platform of MATLAB and the performance will be analysed.

Keywords  
Artificial Neural Network, Artificial Bee Colony algorithm, CNS tumor, colon tumor and ALL_AML Leukemia.

1. INTRODUCTION
In Microarray technology, vast quantities of data are collected because of the quantity of information brought forth from every experiment [1]. Microarray technology has become one of the indispensable tools that several biologists utilize to monitor genome broad expression levels of genes in a provided organism [2]. Generally, a microarray is a glass slide to which DNA molecules are fixed in an orderly manner at particular locations named spots. By the presence of genetic variations [3] [4] nearly all human genetic diseases such as cancer and developmental abnormalities are classified. To explore significant normal and abnormal biological phenomena, the discovery and analysis of gene expression patterns of several model organisms shows a fascinating opportunity. Within a particular mRNA sample [5], DNA microarrays can simultaneously measure the expression level of thousands of genes. In order to: 1) identify diagnostic or prognostic biomarkers; 2) classify diseases; 3) monitor the response to therapy; and 4) understand the mechanisms involved in the genesis of disease processes [7], such high-throughput expression profiling can be used to compare the level of gene transcription in clinical conditions.

Many microarray technologies that supervise the level of expression of a multitude of genes have lately come forth [6]. In several ways to measure gene expression, Microarrays may be utilized, but one of the most popular applications is to compare expression of a set of genes from a cell maintained in a specific condition to the same set of genes from a reference cell maintained under normal conditions [8]. Provided DNA-microarray information for a set of cells classified by a rendered phenotype and for a set of control cells, an significant issue is to recognize “patterns” of gene expression that can be applied to predict cell phenotype. It is possible to examine the expression of thousands of genes at once with microarrays. Checking for elevated expression of sure genes can help in foretelling cancer [9].

As an imperative tool, Micro array technology has come forth in the tracking of genome-wide expression levels of gene [10]. In diverse organs applying micro array technologies [11], separate genes, gene ensembles, and the metabolic ways fundamental to the structurally workable organization of an organ and its physiological function are brought out by the analysis of the gene expression profiles. The application of micro array technology can automate the diagnostic task and improve the precision of conventional diagnostic techniques. By micro array technology [12], simultaneous analysis of thousands of gene expressions is alleviated. Gene expression microchip likely the most speedily developing tool of genome analysis allows supervising of the expression levels of tens of thousands of genes under various experimental conditions simultaneously. In their atmosphere, this shows a rich tool in the analysis of united gene reaction to variations, and renders suggestions related to the organization of the concerned gene networks [13]. The usage of fuzzy logic for interpretation of gene expression information has not been searched considerably. As a result of the fuzzification of measurements, initial investigations proposed that poor quality clusters are made. Estimating the generalizability of these experiments is not simple because calculating the quality of clusters goes forward to be an extremely subjective task, and various fuzzification systems have not been attempted [14].

To diminish the dimensionality of gene expression information [15] many methods have been suggested earlier. Applying micro array information, numerous machine learning techniques have been effectively used to cancer classification [16] [17]. But, categorization in micro array technology is considered to be extremely hard, because of the high dimensionality and insignificant sample size of the gene expression information. For the successful classification of gene expression data, lots of researches have been performed. In the following section [19] [20], a few recent works available in the literature are reviewed.
2. RELATED WORKS AND RESEARCHES

Three various classification methods have been focused on by Jayashree Dev et al. [21]: BPN, FLANN and PSO-FLANN and discovered that the integrated approach of Functional Link Artificial Neural Network (FLANN) and Particle Swarm Optimization (PSO) could foretell the disease as compared to other method. The nonlinearity of the classification issue is overcome by this suggested technique. In order to classify various kinds of cancer genes from vast amount of DNA microarray gene expression information, this suggested algorithm could be developed.

For gene expression information analysis, an algorithm has been suggested by Alok Sharma et al. [22]. The algorithm first separates genes into subsets, the sizes of which are comparatively small, then chooses informative smaller subsets of genes from a subset and merges the selected genes with another gene subset to modify the gene subset. Till all subsets were merged into one informative subset, it iterated this process. For all the test datasets, they depicted promising classification accuracy. By examining three distinct gene expression datasets, they illustrated the effectiveness of the suggested algorithm.

Classification method for microarray gene expression has been suggested by A. Castano F. et al. [23], obtained by the light reflection analysis of genomic material. This suggested algorithm was in two-stages, in the first stage; from thousands of genes, two filter algorithms recognize prominent expression genes. The suggested methodology was performed, in the second stage, utilizing chosen gene subsets as new input variables. The methodology was composed of a combination of Logistic Regression (LR) and Evolutionary Generalized Radial Basis Function (EGRBF) neural networks which have depicted to be extremely precise in previous research in the modeling of high-dimensional patterns. The results received were contrasted, at last, with nonparametric statistical tests and verify good synergy between EGRBF and LR models.

An algorithm has been suggested by Chhanda Ray et al. [24] to examine DNA microarray gene expression patterns efficiently for vast amount of DNA microarray information. On the basis of the variations of DNA microarray gene expression patterns of the same organism by concurrently supervising the expression of thousands of genes, this development method was identified. Finally, based on the distribution probability of codes, classification of cancer genes was also focused.

Microarray analysis or gene expression profiling has been discussed by Venkatesh et al. [25] that gives techniques to examine thousands of genes in a single sample. By providing large amount of data, Micro array analysis was providing challenges in various fields, which could be processed to get helpful data. In this study, the gene samples received from biopsy samples are gathered from colon cancer patients. They brought in a learning vector quantization method that determines anti-fact states and separate malignant genes from regular genes. Finally, organism was identified based on the variations of DNA microarray gene expression patterns of the same organism.

3. ARTIFICIAL BEE COLONY (ABC) ALGORITHM

Artificial Bee Colony (ABC) algorithm is based on the representation of the bees searching behavior. Generally the bee colony consists of a single queen bee that is responsible for laying eggs, thousands of male bees called drones and thousands of worker bees, which are the sterile bees, and the young bee larvae are called broods. The artificial bee colony algorithm consists of 3 types of bees namely the employed bee, onlooker bee and the scout bee. Scout bee is responsible for carrying out random searches in the environment. A bee who visits the source visited by him previously is called an employed bee and the bee that waits in the beehive for decision making is called the onlooker bee. In the ABC algorithm it is assumed that the colony consists of an equal number of employed bees and onlooker bees and for every solution there is an employed bee in the hive. An onlooker waits and decides on whether an optimal solution is acceptable or not. The fitness of the species is given by the profitability of the solution it describes. Information survives by continuing to circulate within the nest, and is capable of reproducing itself by recruiting new foragers who become informed of the food source and come back to the nest and share their information.

4. CLASSIFICATION OF TUMOR GENES

The tumor genes are classified in to subsets using the different classifiers. The different classifiers used here are Principal component analysis (PCA) and Artificial Neural network (ANN). This can be done to reduce the dimension of the datasets to make the analysis easier. Dimension reduction has the approaches to find a subset of the original variables and transforms the data in the high-dimensional space to a space of fewer dimensions.

4.1 Principal Component Analysis

PCA is a linear transformation technique to form a simplified data set retaining the characteristics of the original data set. The transformation is given by

\[ Y = E^T . X \]  

Consider the original data matrix contains \( d \) dimensions and \( n \) observations which have to be reduced to \( k \) dimensional subspace. Here \( E_{d \times k} \) is the projection matrix which contains \( k \) Eigen vectors corresponds to the \( k \) highest Eigen values, and where \( X_{d \times n} \) is a mean centred data matrix.

4.2 Artificial Neural Network

The dimensionality reduced microarray gene dataset is utilized for training the feed forward Neutral network with back propagation. The network maps a set of input values to a set of output values and the graphical representation of a parametric function is supposed. The neural network to train the gene dataset is shown in the fig (1). The steps to be followed in this process is given below

Step 1: Set the input weights of every neuron network with \( N_g \) input layers, a \( N_g \) hidden layers and an output layer are designed.

Step 2: The designed NN is weighted and biased. The developed NN is shown in the Fig.1.

Step 3: The basis function and the activation function which is chosen for the designed NN are shown in Eqn (2) and Eqn (3) respectively whereas \( \bar{M}_c \) is the dimensionality reduced microarray gene data, \( w_i \) is the weight of the neuron, \( \alpha \) is the bias and \( y \) ranges [0,1].
where the network), the...

or gets minimized. Normally, it is

output layer, but with the

parameters.

Step 5: After adjusting the weights, steps (2) and (3) are

repeated until the BP error is minimized. Normally, it is

repeated till the criterion, \( E < 0.1 \) is satisfied.

When the error gets minimized to a minimum value it is

construed that the designed ANN is well trained for its further

testing phase and the BP algorithm is terminated. Thus, the

neural network is trained by using the samples.

5. DOMINANT GENE PREDICTION
USING ABC ALGORITHM

The ABC algorithm is applied on the classified test sequence

and then this test sequence is evaluated and the dominant genes

are predicted. This can be done using the random selection of

population using the search algorithm. The random populations

are the indices of the test sequence which are classified as CNS

tumor (or) colon tumor (or) ALL_AML Leukemia. The fitness

of the selected population is evaluated using (7) where \( N_{out} \)

is the network output obtained from the ANN \( N_{fit} \) in (6) is the

fitness value of the initially generated population.

\[
N_{fit} = \frac{1}{(1 - \mu_{net})^F} \sum_{k=0}^{SN - 1} N_{out} - |k| \quad \text{in case of CNS tumor}
\]

\[
\mu_{net} = \begin{cases} 
0 & \text{if test sequence is class1} \\
1 & \text{if test sequence is class2}
\end{cases} \quad \text{in case of colon tumor}
\]

\[
\begin{cases} 
0 & \text{if test sequence is positive} \\
1 & \text{if test sequence is negative}
\end{cases} \quad \text{in case of ALL_AML Leukemia}
\]

The steps involved in the ABC algorithm is discussed below

Step1: Initialize the population of solutions \( X_{ij} \).

\( i = 1 \ldots SN \) and \( j = 1 \ldots D \) where \( SN \): the

number of food source and \( D \) is the number of optimization

parameters.

Step2: Evaluate the population by the means of fitness

Step 2: In case of training perform output calculation based

on two functions i.e. Basis function (the product of weights and

inputs) and Activation function (non-linear).

Step 3: In order to determine the BP error using Eq. (2), the

training gene data sequence is given to the NN. Eq. (2), Eq. (3)

and Eq. (4) show the basis function and transfer function.

Step 4: The weights of all the neurons are adjusted when the BP

error is determined as follows,

\[
w_{ij} = w_{ij} + \Delta w_{ij}
\]

The change in weight \( \Delta w_{ij} \) given in Eq. (5) can be determined

as

\[
\Delta w_{ij} = \gamma E_{ij} Y_{bij} \quad \text{where } E \text{ is the BP error and } \gamma \text{ is the}

learning rate; normally it ranges from 0.2 to 0.5.

Step 5: After adjusting the weights, steps (2) and (3) are

repeated until the BP error gets minimized. Normally, it is

repeated till the criterion, \( E < 0.1 \) is satisfied.

The pre-processing steps for predicting dominant genes are

explained in the following steps in the fig (2).

4.3 Minimization of Error by BP Algorithm

The Back Propagation algorithm is used to train the network

and to minimize the total errors. The steps involved in Back

Propagation Algorithm are discussed below:

Step 1: Assign random weights to the links range \([0, 1]\) in the

Artificial Neural network above

\[
Y_i = \alpha + \sum_{j=0}^{N_y-1} w_{ij} M_{ejj} , \quad 0 \leq i \leq N_y - 1
\]

\[
g(y) = \frac{1}{1+e^{-y}}
\]

\[ (2) \]

\[ (3) \]

\[ (4) \]

\[ (5) \]
Step 3: Produce new solutions \( v_{ij} \) for the employed bees by using (8) and evaluate them

\[
v_{ij} = x_{ij} + \phi_{ij}(x_{ij} - x_{kj})
\]  

(8)

Step 4: Apply greedy selection process using the probability values \( P_i \) for the solutions \( x_{ij} \) using (2) where \( fit_i \) is the fitness value of the solution \( i \) which is proportional to the nectar amount of the food source in the position \( i \) and \( SN \) is the number of food sources which is equal to the number of employed bees.

\[
P_i = \frac{fit_i}{\sum_{n=1}^{SN} fit_n}
\]  

(9)

Step 5: Produce the new solutions \( v_{ij} \) for the onlookers from the solutions \( x_{ij} \) selected depending on \( P_i \) and evaluate them.

Step 6: Apply the greedy selection process.

Step 7: Determine the abandoned solution for the scout, if exists, and replace it with a new randomly produced solution \( x_{ij} \) by using (10)

\[
x_{ij} = x_{min}^j + \text{rand}(0,1)(x_{max}^j - x_{min}^j)
\]  

(10)

Step 8: Memorize the best solution achieved and cycle = cycle + 1 until the maximum no. of cycle reached.

6. EXPERIMENTAL RESULTS

The proposed classification technique is implemented in the MATLAB platform and it is evaluated using the microarray gene expression data. The training dataset is of dimension \( N_g = 7129 \) and \( N_s = 60 \) in case of CNS tumor, \( N_g = 2000 \) and \( N_s = 62 \) in case of colon tumor, \( N_g = 7129 \) and \( N_s = 72 \) in case of ALL_AML [26]. This high dimensional training dataset is subjected to dimensionality reduction using ANN and so a dataset of dimension 50×60, 50×62, 50×72 for CNS tumor, colon tumor and ALL_AML respectively is obtained.

This training dataset is utilized to design the ANN and then the test input sequence is tested through the ABC algorithm. Testing is done under the evaluation of sensitivity and specificity values. These values are among the terms true positive (TP), true negative (TN), false positive (FP), false negative (FN) [27] which are given in the Table (1) below

<table>
<thead>
<tr>
<th>Data sets</th>
<th>Dimension reduction using PPCA</th>
<th>Dimension reduction using ABC with ANN</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNS tumor</td>
<td>TP 6</td>
<td>TN 12</td>
</tr>
<tr>
<td>Colon tumor</td>
<td>TP 0</td>
<td>TN 25</td>
</tr>
<tr>
<td>ALL_AML</td>
<td>TP 2</td>
<td>TN 12</td>
</tr>
</tbody>
</table>

The graphical representation for the Sensitivity, Specificity, FPR, accuracy, PPV, NPV, FDR, MCC for the PPCA and GA with ANN classifiers under different datasets are shown below.
Fig 6 Graphical representation of false positive rate for PCA and GA with ANN classifiers under different datasets

Fig 7 Graphical representation of positive prediction value for PPCA and GA with ANN classifiers under different datasets

Fig 8 Graphical representation of negative prediction value for PCA and ABC with ANN classifiers under different datasets

Fig 9 Graphical representation of false discovery rate for PCA and ABC with ANN classifiers under different datasets

Fig 10 Graphical representation of Matthew’s correlation coefficient for PCA and ABC with ANN classifiers under different datasets

DISCUSSION
Considering the values in the Table (1), sensitivity, specificity, accuracy, positive prediction value, negative prediction value, false positive rate, false discovery rate and Matthews’s correlation coefficient can be found out. The graphical representations of the testing values are plotted in the figures 3 – 10.

In case of sensitivity in fig (3), the cancer prediction rate using the classifier ABC with ANN is better than PCA whereas ABC with ANN performs with 50%-65% but PCA has only 40%. The specificity in fig (4) also yields a better outcome with 50% - 100% in case of ABC with ANN but PCA with 35% - 60%. The accuracy rate in fig (5) gives 35% in PCA and 70% in ABC with ANN. The false positive rate in fig (6), which is recognized as an error, is observed as average rate of 32% in PCA and 29% in ABC with ANN. In fig (7), positive predictive value which is the proportion of the positive results has been identified as maximum of 30% in PCA and 95% in ABC with ANN. In fig (8) the negative prediction value, the proportion of negative results, has 60% in PCA and 95% in ABC with ANN. The false discovery rate in fig (9) which is the identification of false results is analyzed as minimum of 46% in PCA and 43% in ABC with ANN. The Matthew’s correlation coefficient has been used for the identification of the results, which hold its value from the range of 0 to 1. The value 1 concludes as a correct identification, while analyzing the fig (10) it is concluded that ABC with ANN has better results than PPCA. So the cancer prediction can be done easily using the proposed classifier ABC with ANN.

7. CONCLUSION
Most literary works had shown interest on developing better classifiers to improve the classification accuracy of microarray gene expression dataset. A mechanism to predict the cancer patients using artificial neural network is presented. The training and testing of the three datasets under consideration is undergone using classification function (sensitivity and specificity measures) respectively. Results presented in this paper shows that the cancer prediction is easily done using the proposed ABC with ANN mechanism while comparing to the conventional PCA method. Even the accuracy level of the proposed classifier is encouraging ANN needs some enhancement to improve the level. Some adaptive techniques with the ANN will be used for the better improvement in the future works.
REFERENCES


[27] en.wikipedia.org/wiki/Sensitivity_and_specificity

IJCA™: www.ijcaonline.org
Paper 2


Publication Type: International Journal
Year of Publishing: 2013
Publisher: Science and Information (SAI)Organization, New York, USA.
Performance Analysis and Evaluation of Different Data Mining Algorithms used for Cancer Classification

Gopala Krishna Murthy Nookala  
Department of Computer Science & Engineering  
Grandhi Varalakshmi Venkatarao Institute of Technology  
Bhimavaram, Andhra Pradesh, India

Nagaraju Orsu  
Department of Computer Science  
Goverment Degree College  
Macherla, Andhra Pradesh, India

Bharath Kumar Pottumuthu  
Centre for Bioinformatics Research & S/w Development  
Grandhi Varalakshmi Venkatarao Institute of Technology  
Bhimavaram, Andhra Pradesh, India

Suresh B. Mudunuri*  
Centre for Bioinformatics Research & S/w Development  
Grandhi Varalakshmi Venkatarao Institute of Technology  
Bhimavaram, Andhra Pradesh, India

* Corresponding Author

Abstract—Classification algorithms of data mining have been successfully applied in the recent years to predict cancer based on the gene expression data. Micro-array is a powerful diagnostic tool that can generate handful information of gene expression of all the human genes in a cell at once. Various classification algorithms can be applied on such micro-array data to devise methods that can predict the occurrence of tumor. However, the accuracy of such methods differ according to the classification algorithm used. Identifying the best classification algorithm among all available is a challenging task. In this study, we have made a comprehensive comparative analysis of 14 different classification algorithms and their performance has been evaluated by using 3 different cancer data sets. The results indicate that none of the classifiers outperformed all others in terms of the accuracy when applied on all the 3 data sets. Most of the algorithms performed better as the size of the data set is increased. We recommend the users not to stick to a particular classification method and should evaluate different classification algorithms and select the better algorithm.

Keywords—Weka; Cancer Classification; Micro-array; Data-mining; Classification Algorithms; Gene Expression Data;

I. INTRODUCTION

Advancement of Information Technology led to huge data accumulation in the recent years in several domains including banking, retail, telecommunications and medical diagnostics. The data from all such domains includes valuable information and knowledge which is often hidden. Processing the huge data and retrieving meaningful information from it is a difficult task. Data Mining is a wonderful tool for handling this task. The term Data Mining, also known as Knowledge Discovery in Databases (KDD) refers to the non trivial extraction of implicit, previously unknown and potentially useful information from data in databases [1]. Data mining in cancer research has been one of the important research topics in biomedical science during the recent years [2].

They are several different data mining techniques like Pattern Recognition, Clustering, Association and Classification [3]. Classification has been identified as an important problem in the emerging field of data mining [4] as they try to find meaningful ways to interpret data sets. Classification of data is very typical task in data mining. There are large number of classifiers that are used to classify the data namely Bayes, Function, Rule’s based, Tree based classification etc. The goal of classification is to correctly predict the value of a designated discrete class variable, given a vector of predictors or attributes [5]. In the age of bioinformatics, cancer data sets have been used for the cancer diagnosis and treatment that can improve human aging [6].

Cancer is a disease characterized by uncontrolled growth and spread of the abnormal cells and the capability to invade other tissues that can be caused by both external factors like radiation, chemicals, tobacco etc., and internal factors like inherited mutations, hormones, immune conditions, etc. There are more than 100 different types of cancers. Most of the cancers are named after the organ or type of cell in which they appear e.g., Melanoma, Colon Cancer, Breast Cancer etc.

All cancers begin in cells which are the structural and functional units of the body. These cells grow and divide in a controlled way to produce more cells as they are needed to keep the body healthy. When cells become old or damaged, they die and are replaced with new cells. However, sometimes life cycle of the cells fails or gets disturbed due to many reasons. When this happens, cells do not die as expected and new cells are formed even when the body does not need them. These extra cells may form a mass of tissue called a tumor. Tumors can be either benign or malignant. Some cancers do not form tumors. For example, leukemia is a cancer of the blood that does not form tumors.

Gene expression analysis of cancer is used to study 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.

In the recent years, tumor classification is frequently studied by applying various data mining classification algorithms on cancer gene expression micro-array data sets so as to predict the presence of cancer. However, the availability of several algorithms in data mining for classification often leads to confusion over the selection of the right algorithm. In this study, we have made a comparative analysis of the performances of various classification algorithms on different cancer micro-array data sets.

II. MATERIALS AND METHODS

We have used the popular, open-source data mining tool Weka (version 3.6.6) for this analysis. Three different data sets have been used and the performance of a comprehensive set of classification algorithms (classifiers) has been analyzed. The analysis has been performed on a HP Windows system with Intel® Core ™ i3 CPU, 2.40 GHz Processor and 4.00 GB RAM. The data sets have been chosen such that they differ in size, mainly in terms of the number of attributes.

A. Data set 1:

   The first data set is a small Breast Cancer Micro-array Gene Expression data used in an earlier study [7]. The data set contains 9 attributes apart from the class attribute with 286 instances.

B. Data set 2:

   The second data set is a medium sized data set with Micro-array Gene Expression data of Lymphoma patients [8]. The data set has a total of 4,026 attributes and 45 instances.

C. 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 [9].

D. Classifiers Used:

   A total of 14 classification algorithms have been used in this comparative study. The classifiers in Weka have been categorized into different groups such as Bayes, Functions, Lazy, Rules, Tree based classifiers etc. 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 & Simple Cart (from Trees). The following sections explain a brief about each of these algorithms.

1. Bayes Net

   Bayes Nets or Bayesian networks are graphical representation for probabilistic relationships among a set of random variables. A Bayesian network is an annotated Directed Acyclic Graph (DAG) that encodes a joint probability distribution [10].

2. Naive Bayesian

   Naive Bayesian classifier is developed on bayes conditional probability rule used for performing classification tasks, assuming attributes as statistically independent; the word Naive means strong. All attributes of the data set are considered as independent and strong of each other [11].

3. Simple Logistics

   It is a classifier used for building linear logistic regression models. LogitBoost with simple regression functions are base learners used for fitting the logistic models. The optimal number of LogitBoost iterations to perform is cross-validated, which leads to automatic attribute selection [12].

4. Multilayer Perceptron

   Multilayer Perceptron is a nonlinear classifier based on the Perceptron. A Multilayer Perceptron (MLP) is a back propagation neural network with one or more layers between input and output layer. The following diagram illustrates a perceptron network with three layers [13].

5. SMO

   Sequential Minimal Optimization (SMO) is used for training a support vector classifier using polynomial or RBF kernels. It replaces all missing the values and transforms nominal attributes into binary ones [14]. A single hidden layer neural network uses exactly the same form of model as an SVM.

6. IBk

   IBk is a k-nearest-neighbor classifier that uses the same distance metric. k-NN is a type of instance based learning or lazy learning where the function is only approximated locally and all computation is deferred until classification. In this algorithm an object is classified by a majority vote of its neighbors [15].

7. KStar (K*)

   Aha, Kibler & Albert describe three instance-based learners of increasing sophistication. IB1 is an implementation of a nearest neighbor algorithm with a specific distance function. IB3 is a further extension to improve tolerance to noisy data. Instances that have a sufficiently bad classification history are forgotten and only instances that have a good classification history are used for classification. Aha [16] described IB4 and IB5, which handle irrelevant and novel attributes.
8. NNge

Instance-based learners are “lazy” in the sense that they perform little work when learning from the data set, but expend more effort classifying new examples. The simplest method, nearest neighbor, performs no work at all when learning. NNge does not attempt to out-perform all other machine learning classifiers. Rather, it examines generalized exemplars as a method of improving the classification performance of instance-based learners [17].

9. PART

PART uses the separate-and-conquer strategy, where it builds a rule in that manner and removes the instances it covers, and continues creating rules recursively for the remaining instances. Where C4.5 and RIPPER does global optimization to produce accurate rule sets, this added simplicity is the main advantage of PART [18].

10. ZeroR

ZeroR is the simplest classification method which depends on the target and ignores all predictors. ZeroR classifier simply predicts the majority category (class). Although there is no predictability power in ZeroR, it is useful for determining a baseline performance as a benchmark for other classification methods [19].

11. ADTree

Alternating Decision Tree is one of the classification method used in Machine learning which consists of decision nodes and prediction nodes. An instance is classified by an ADTree for which all decision nodes are true and summing any prediction nodes that are traversed. This makes it different from basic classification tree models that follow only one path through the tree [20].

12. J48

The J48 algorithm is WEKA’s implementation of the C4.5 decision tree learner. The algorithm uses a greedy technique to induce decision trees for classification and uses reduced-error pruning [21].

13. Random Forest

Random forest is an ensemble classifier which consists of many decision tree and gives class as outputs i.e., the mode of the class’s output by individual trees. Random Forests gives many classification trees without pruning [22].

14. Simple Cart

CART is a recursive and gradual refinement algorithm of building a decision tree, to predict the classification situation of new samples of known input variable value. Breiman et. al., 1984 provided this algorithm and is based on Classification and Regression Trees (CART) [23].

In our study, we have applied all the above classifiers on the 3 different cancer data sets and the results have been analyzed.

III. RESULTS AND DISCUSSION

The data sets have been submitted to a set of classification algorithms of Weka. We have used the 'Explorer' option of the Weka tool. Certain comparative studies conducted earlier [24][25][26][27][28] have shown that a particular algorithm has performed better on their data set and their conclusions however differ from each other. The studies either have used a very minimal set of classifiers or have used data sets that are not diverse resulting in an advantage or bias for a particular algorithm. Keeping that in mind, we have included a good number of classifiers in our analysis and used data sets that are diverse (in terms of size). The following sections describe the results obtained in our analysis.

A. Classification of Data set 1

The data set 1 is a small data set of micro-array gene expression data of Breast Cancer with 10 attributes and 286 instances. 5 out of the 14 algorithms got an accuracy of more than 95% where as the remaining algorithms reported the classification accuracy between 70% and 80%. Table 1 shows the results obtained in the analysis on data set 1.

The results in Table 1 indicate that the classifiers Multilayer Perceptron (ANN), IBk, KStar, NNge, and Random Forest performed better than the remaining algorithms. The Multilayer perceptron however took more time (11.68 secs) for classification whereas the remaining algorithms took almost less than 1 second. The kappa statistic for these 5 algorithms has been almost the same (~0.9). It should be noted that except IBk and KStar (Lazy classifiers), the classifiers among the better performers do not belong to the same group.

B. Classification of Data set 2

When a medium size data set (Lymphoma data set with 4,026 attributes and 45 instances) has been classified, the performance of the classifiers has significantly improved. All the classifiers (except ZeroR) reported more than 97% accuracy. Table 2 gives a summary report of the performances of all the classifiers when applied on Lymphoma data set.

10 out of 14 classifiers have got 100% accuracy as they correctly classified all the 45 instances. Though the number of instances decreased from 268 instances (from data set 1) to 45, the performance of the classifiers has been very good. The data set 2 has more number of attributes than data set 1 that resulted in better accuracy. The multilayer perceptron besides classifying all the instances correctly has however took a longer time (890.2 seconds) to get the results and hence, the accuracy of multi-layer perceptron can be ignored.

C. Classification of Data set 3

Finally, the large data set of Leukemia with 7,129 attributes and 34 instances has been used. The classifiers have achieved accuracies similar to the classification of medium size data set. However, the classifiers KStar and ZeroR underperformed. Rest of the classifiers achieved accuracies close to 100%. As expected, Multilayer perceptron took very long time to generate results. Table 3 gives a summary report of the performances of all the classifiers when applied on Leukemia data set.
TABLE I. Comparison of different classifiers using Breast Cancer Micro-array Gene Expression Data set with 10 attributes and 286 instances.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Time Taken</th>
<th>Correctly Classified Instances</th>
<th>Incorrectly Classified Instances</th>
<th>Kappa statistic</th>
<th>Mean absolute error</th>
<th>Root mean squared error</th>
<th>Confusion Matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bayes Net</td>
<td>0.02 Sec</td>
<td>217 (75.9%)</td>
<td>69 (24.1%)</td>
<td>0.3958</td>
<td>0.3018</td>
<td>0.4284</td>
<td>a b 173 28</td>
</tr>
<tr>
<td>Naive bayes</td>
<td>0.03 Sec</td>
<td>215 (75.2%)</td>
<td>71 (24.8%)</td>
<td>0.3693</td>
<td>0.3012</td>
<td>0.4278</td>
<td>a b 174 27</td>
</tr>
<tr>
<td>Multi layer Perceptron</td>
<td>11.7 Sec</td>
<td>276 (96.5%)</td>
<td>10 (3.5%)</td>
<td>0.9157</td>
<td>0.0482</td>
<td>0.1567</td>
<td>a b 197 4</td>
</tr>
<tr>
<td>Simple Logistics</td>
<td>0.87 Sec</td>
<td>218 (76.2%)</td>
<td>68 (23.8%)</td>
<td>0.32</td>
<td>0.3535</td>
<td>0.4183</td>
<td>a b 191 10</td>
</tr>
<tr>
<td>SMO</td>
<td>0.11 Sec</td>
<td>218 (76.2%)</td>
<td>68 (23.8%)</td>
<td>0.3615</td>
<td>0.2378</td>
<td>0.4876</td>
<td>a b 183 18</td>
</tr>
<tr>
<td>IBk</td>
<td>0 Sec</td>
<td>280 (97.9%)</td>
<td>6 (2.1%)</td>
<td>0.9491</td>
<td>0.0253</td>
<td>0.1053</td>
<td>a b 200 1</td>
</tr>
<tr>
<td>KStar</td>
<td>0 Sec</td>
<td>280 (97.9%)</td>
<td>6 (2.1%)</td>
<td>0.9494</td>
<td>0.0747</td>
<td>0.1399</td>
<td>a b 199 2</td>
</tr>
<tr>
<td>NNge</td>
<td>0.27 Sec</td>
<td>278 (97.2%)</td>
<td>8 (2.8%)</td>
<td>0.933</td>
<td>0.028</td>
<td>0.1672</td>
<td>a b 197 4</td>
</tr>
<tr>
<td>PART</td>
<td>0.21 Sec</td>
<td>229 (80.1%)</td>
<td>57 (19.9%)</td>
<td>0.4825</td>
<td>0.299</td>
<td>0.3866</td>
<td>a b 184 17</td>
</tr>
<tr>
<td>ZeroR</td>
<td>0 Sec</td>
<td>201 (70.3%)</td>
<td>85 (29.7%)</td>
<td>0</td>
<td>0.4183</td>
<td>0.457</td>
<td>a b 201 0</td>
</tr>
<tr>
<td>ADTree</td>
<td>0.08 Sec</td>
<td>223 (78.0%)</td>
<td>63 (22.0%)</td>
<td>0.4522</td>
<td>0.3659</td>
<td>0.4024</td>
<td>a b 175 26</td>
</tr>
<tr>
<td>J48</td>
<td>0.02 Sec</td>
<td>217 (75.9%)</td>
<td>69 (24.1%)</td>
<td>0.2899</td>
<td>0.3658</td>
<td>0.4269</td>
<td>a b 194 7</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.24 Sec</td>
<td>278 (97.2%)</td>
<td>8 (2.8%)</td>
<td>0.9326</td>
<td>0.1439</td>
<td>0.204</td>
<td>a b 193 8</td>
</tr>
<tr>
<td>Simple Cart</td>
<td>1.1 Sec</td>
<td>201 (70.3%)</td>
<td>85 (29.7%)</td>
<td>0</td>
<td>0.4177</td>
<td>0.457</td>
<td>a b 201 0</td>
</tr>
</tbody>
</table>

TABLE II. Comparison of different classifiers using Lymphoma Cancer Micro-array Gene Expression Data set with 4,026 attributes and 45 instances.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Time Taken</th>
<th>Correctly Classified Instances</th>
<th>Incorrectly Classified Instances</th>
<th>Kappa statistic</th>
<th>Mean absolute error</th>
<th>Root mean squared error</th>
<th>Confusion Matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bayes Net</td>
<td>0.27 Sec</td>
<td>45 (100%)</td>
<td>0 (0%)</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>a b 22 0</td>
</tr>
<tr>
<td>Naive bayes</td>
<td>0.24 Sec</td>
<td>45 (100%)</td>
<td>0 (0%)</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>a b 22 0</td>
</tr>
<tr>
<td>Multi layer Perceptron</td>
<td>890.2 Sec</td>
<td>45 (100%)</td>
<td>0 (0%)</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>a b 22 0</td>
</tr>
<tr>
<td>Simple Logistics</td>
<td>5.92 Sec</td>
<td>45 (100%)</td>
<td>0 (0%)</td>
<td>1</td>
<td>0.0641</td>
<td>0.0985</td>
<td>a b 22 0</td>
</tr>
<tr>
<td>SMO</td>
<td>0.18 Sec</td>
<td>45 (100%)</td>
<td>0 (0%)</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>a b 22 0</td>
</tr>
</tbody>
</table>
### TABLE III. Comparison of different classifiers using Leukemia Cancer Micro-array Gene Expression Data set with 7,129 attributes and 34 instances.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Time Taken</th>
<th>Correctly Classified Instances</th>
<th>Incorrectly Classified Instances</th>
<th>Kappa statistic</th>
<th>Mean absolute error</th>
<th>Root mean squared error</th>
<th>Confusion Matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bayes Net</td>
<td>1.78 Sec</td>
<td>34 (100%)</td>
<td>0 (0%)</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>a b 20 0</td>
</tr>
<tr>
<td>Naive bayes</td>
<td>0.41 Sec</td>
<td>34 (100%)</td>
<td>0 (0%)</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>a b 20 0</td>
</tr>
<tr>
<td>Multi layer Perceptron</td>
<td>1313.87 Sec</td>
<td>33 (97.1%)</td>
<td>1 (2.9%)</td>
<td>0.9038</td>
<td>0.376</td>
<td>0.0267</td>
<td>a b 20 0</td>
</tr>
<tr>
<td>Simple Logistics</td>
<td>9.5 Sec</td>
<td>34 (100%)</td>
<td>0 (0%)</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>a b 20 0</td>
</tr>
<tr>
<td>SMO</td>
<td>0.19 Sec</td>
<td>34 (100%)</td>
<td>0 (0%)</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>a b 20 0</td>
</tr>
<tr>
<td>IBk</td>
<td>0.01 Sec</td>
<td>34 (100%)</td>
<td>0 (0%)</td>
<td>1</td>
<td>0.0278</td>
<td>0.0278</td>
<td>a b 20 0</td>
</tr>
<tr>
<td>KStar</td>
<td>0 Sec</td>
<td>20 (58.8%)</td>
<td>14 (41.2%)</td>
<td>0</td>
<td>0.5</td>
<td>0.5</td>
<td>a b 20 0</td>
</tr>
<tr>
<td>NNge</td>
<td>1.48 Sec</td>
<td>34 (100%)</td>
<td>0 (0%)</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>a b 20 0</td>
</tr>
<tr>
<td>PART</td>
<td>0.32 Sec</td>
<td>34 (100%)</td>
<td>0 (0%)</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>a b 20 0</td>
</tr>
<tr>
<td>ZeroR</td>
<td>0 Sec</td>
<td>20 (58.8%)</td>
<td>14 (41.2%)</td>
<td>0</td>
<td>0.4853</td>
<td>0.4922</td>
<td>a b 20 0</td>
</tr>
<tr>
<td>ADTree</td>
<td>1.5 Sec</td>
<td>34 (100%)</td>
<td>0 (0%)</td>
<td>1</td>
<td>0.0142</td>
<td>0.0145</td>
<td>a b 20 0</td>
</tr>
<tr>
<td>J48</td>
<td>0.52 Sec</td>
<td>34 (100%)</td>
<td>0 (0%)</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>a b 20 0</td>
</tr>
</tbody>
</table>
The results from the above 3 tables have been analyzed manually and they indicate that the classifiers work better when there is an increase in the number of attributes in the data set. But, none of the classifiers outperformed the others in terms of the accuracies. The classifiers Multilayer perceptron, IBk, NNge, and Random Forest have performed better on all 3 data sets. However, the performance of Multilayer Perceptron should not be considered because of the huge execution time taken by the classifier to generate results. The algorithm KStar reported around 58% accuracy for the large data set whereas the classifier ZeroR did not perform well on all 3 data sets. The remaining classifiers (except KStar and ZeroR) performed better on large data sets which are expected. The other statistics like kappa statistic and errors seem to be more or less same among all the classifiers in all three tests and are based on the accuracy of the prediction.

IV. CONCLUSION

This study focuses on finding the right algorithm for classification of data that works better on diverse data sets. However, it is observed that the accuracies of the tools vary depending on the data set used. It should also be noted that classifiers of a particular group also did not perform with similar accuracies. Overall, the results indicate that the performance of a classifier depends on the data set, especially on the number of attributes used in the data set and one should not rely completely on a particular algorithm for their study. So, we recommend that users should try their data set on a set of classifiers and choose the best one.

V. FUTURE WORK

We would like to develop web based software for performance evaluation of various classifiers where the users can just submit their data set and evaluate the results on the fly.

ACKNOWLEDGMENT

Bharath Kumar would like to thank Dr. DSVGK Kaladhar of GITAM University for his timely co-operation and guidance in completing the analysis. The authors greatly acknowledge the management and staff of GVVIT for their support and for providing lab facilities.

REFERENCES

[16] A. Jemal, DVM, PhD; Freddie Bray, PhDCA CANCER J CLIN 2011;61:69–90
[17] B. Martin. Instance - Based Learning: Nearest Neighbour with generalisation, Department of Computer Science, University of Waikato, Hamilton, New Zealand


AUTHORS PROFILE

Mr. Gopala Krishna Murthy Nookala obtained his B.Sc degree from Andhra University, Visakhapatnam, India. He obtained his Master’s degree in Computer Applications from Bharathiar University, M.Tech in Information Technology from Andhra University and also obtained Master of Philosophy in Computer Science from Alagappa University. He is pursuing his PhD in Computer Science & Engineering in Acharya Nagarjuna University since 2008. Currently, he is working as a Professor in the Department of Computer Science and Engineering, G.V.V.R. Institute of Technology, Bhimavaram, Andhra Pradesh. His specialties include Bioinformatics, Data mining, Networking, and Software Engineering. His current research interests are complexity reduced tumor classification using micro-array data-sets, usage of bioinformatics tools and optimization techniques for ant based networks using swarm intelligence.

Mr. Bharath Kumar Pottumutthu obtained his M.Sc in Bioinformatics from GITAM University, Visakhapatnam, India and B.Sc from Andhra University, Visakhapatnam. He is currently working as a Research Assistant in Centre for Bioinformatics and Software Development (CBRSD) at GVVR Institute of Technology, Bhimavaram. His current research interests include Comparative Genomics, Bioinformatics and Machine Learning.

Dr. NagaRaju Orsu received Ph.D. degree in Computer Science in the faculty of Engineering in 2011 from Acharya Nagarjuna University, Guntur, India and Masters Degree in Computer Science & Engineering from Acharya Nagarjuna University. He has a satisfactory and consistent academic track record. He is currently working as Head of the Department of Computer Science, Government Degree College, Macherla. He has occupied various academic positions like practical examiner, project adjudicator for various universities and colleges in and around in Andhra Pradesh. His current research is focused on Bioinformatics, Software Engineering, Image Processing & Database Management System. He has published several papers in National & International Journals.

Dr. Suresh B. Mudunuri received his PhD in Computer Science & Systems Engineering from Andhra University College of Engineering, Visakhapatnam, India in the year 2012. He received his BCA and MS degrees from Andhra University, and International Institute of Information Technology (IIIT), Hyderabad in 2002 and 2005 respectively. During 2005 and 2007, he worked on a bioinformatics research project in the Lab of Computational Biology (LCB) at Centre for DNA Fingerprinting and Diagnostics (CDFD), Hyderabad. Later, he has also worked as an Associate Professor in CSE departments at Aditya Engineering College, Surampalem and Vishnu Institute of Technology, Bhimavaram. He is currently working as Head of Centre for Bioinformatics Research & Software Development (CBRSD) at GVVR Institute of Technology, Bhimavaram. He is passionate to do research in the areas of Bioinformatics, Computational Biology and Data mining. He has published good number of quality research papers in international journals and conferences.
Paper 3


Publication Type: International Conference
Year of Publishing: 2013
Publisher: ACM
Appendix B

Glossary
A

Adenine
One of the 4 bases found in DNA and RNA; Pairs with Thymine (T) in DNA and Uracil (U) in RNA.

Algorithm
A procedure that is used to solve a mathematical or computational problem, especially to address a data processing issue.

Aminoacids
Organic compounds containing an amino group as well as a carboxylic acid group. Proteins are composed of a sequence of 20 different aminoacids.

Artificial Neural Network (ANN)
It is a learning algorithm that is stimulated by the structure and functional characteristics of biological neural networks.

Artificial Bee Colony(ABC)
ABC algorithm simulates the intelligent foraging behavior of a honey bee swarm.

ALL_AML Leukemia
Acute myeloid leukemia (AML) is a cancer of the myeloid line of blood cells, characterized by the rapid growth of abnormal white blood cells that accumulate in the bone marrow and interfere with the production of normal blood cells. Acute lymphocytic leukemia (ALL) is a cancer that starts from white blood cells called lymphocytes in the bone marrow (the soft inner part of the bones, where new blood cells are made).

B

Bacteria
Single-celled, microorganisms that are found in air, soil, and in living bodies. Some of them are known to cause diseases or infections while some are very useful.

Bioinformatics
A field which involves the collection, organization, storage, analysis, and integration of large amounts of biological data using networks of computers and databases.

Biomolecule
Any organic molecule that is produced by a living organism. Examples: Protein, DNA, RNA, metabolites.

Biological Data
Data or measurements collected from biological sources, which is stored or exchanged in a digital form. Biological data is commonly stored in files or databases. Examples of biological data are DNA / RNA base-pair sequences, protein data and population data used in ecology.

Biological Process
An organic process that occurs in living organisms such as metabolism, reproduction, etc.
Biology
A natural science concerned with the study of life and living organisms, including their structure, function, growth, origin, evolution, distribution, and taxonomy.

Back propagation
Back propagation or feed forward, is a method of determining the parameters required for an efficient neural network.

C
cDNA
Complimentary DNA; Synthesized in the laboratory from the mRNA sequence back to DNA.

Cell
The basic and functional unit of any living organism. Some of the organisms are unicellular (single celled) while some are multicellular (made up of many cells).

Chromosome
A thread like structure found in cells of all organisms that is made up of a single long DNA molecule. The entire DNA of a cell is wrapped into a set of chromosomes.

Codon
A sequence of three adjacent nucleotides which encode for a specific aminoacid during protein synthesis.

Compliment (DNA)
The preferential binding of bases A with T / U and G with C in a DNA / RNA sequence.

Classification
Classification is a supervised learning that learns a process for predicting the class instance from pre labeled (classified) instances.

Clustering
Clustering is an unsupervised learning that finds natural grouping of instances given unlabeled data.

CNS tumor
The central nervous system tumor (CNS) is defined as an abnormal growth of cells within the brain or the central spinal canal.

Colon tumor
Colorectal cancer is a cancer from uncontrolled cell growth in the colon or rectum (parts of the large intestine), or in the appendix.

Cancer Classification
Cancers may be classified by their primary site of origin or by their histological or tissue types.

Classification Algorithms
A Classification Algorithm is a procedure for selecting a hypothesis from a set of alternatives that best fits a set of observations.
D
Database
An electronic collection of records that allows the data to be manipulated, retrieved, and analyzed.

Data mining
Field of computational and statistical algorithms for automatic searching of large databases to find patterns.

Data Warehouse
A collection of data from a variety of sources stored at a particular place and is used for analysis and report generation.

Deletion (Mutation)
A type of mutation where one or more bases are deleted or removed from a DNA / RNA sequence.

Domain (Protein)
A part of the protein sequence that determines the function of the protein.

Diagnosis
The process of identification of the nature and cause of an illness. Disease / Disorder Abnormal condition of an organism that impairs the normal functions. It may be caused by external factors, bacteria, viruses or can be genetic.

Divergence
The variation that deviates the original sequence with the mutated sequence.

DNA
De-oxyribo Nucleic Acid (DNA) is a double helical structure formed by nucleotides that contains the genetic instructions for the development and functioning of an organism.

DNA Chip
A microchip that holds DNA segments to be used mainly for micro-array analysis.

DNA Computing
A form of computing that uses DNA, Molecules and Chemicals aimed at developing the next generation computers.

Drug
A chemical compound that, when absorbed by the organism, alters the normal function of the body. It is a substance that is used as a medicine to treat illness.

Drug Design (Computer-aided)
The process of computer modeling to modify an existing drug or design a new drug that interacts with a target molecule.

E
Enzyme
Form of proteins that catalyzes (speeds up) the biochemical processes and reactions.

Evolution
The change in the inherited characteristics of a population of organisms over a period of successive generations. Evolution leads to the development of different species / organisms due to the changes that happen over a period of time in their genetic material.

**F**
Feature (Software)
A characteristic part of a software; Option or functional capability of a software.

**Fuzzy**
Being indistinct.

**G**
Gene
The segment of DNA that is involved in the formation of protein. Genes are the units of heredity that are transferred from one generation to another.

Gene Density
Number of genes per unit genome.

Gene Expression
The process by which a protein is produced from a gene. Also known as Protein Synthesis.

Gene Expression Data
Gene expression data is highly informative of disease state particularly in the area of oncology where accurate and early diagnosis followed by appropriate treatment.

Gene Flow
Exchange of genes between different, usually related, populations.

Gene Regulation
The processes that turn the gene expression, the process of producing protein, ON and OFF.

Genetic Code
The set of rules by which the information encoded in a DNA sequence (especially gene regions) are translated into an aminoacid sequence (protein). It is a conversion table that specifies what aminoacid is coded for each triplet codon in the DNA sequence.

Genetic Disease
An illness caused by the abnormality in genes or chromosomes.

Genome
The total hereditary / genetic material of an organism or the collection of all genes of an organism. The entire DNA of an organism is also called the Genome.

Genome Sequencing
A laboratory process that determines the complete DNA sequence of an organism's genome at a single time.
Genomics
Study of the genomes or the entire DNA sequence of organisms.

Genotype
The collection of genetic material of an organism that gives rises to its physical properties (phenotype).

High-throughput techniques
Techniques that aid the fast and automated analysis of a variety of substances including chemicals and genes. Example, high-throughput sequencing, high-throughput screening.

In Silico
Experimentation or studies by use of computers to stimulate, process or analyze a biological experiment.

In Vitro
Experimentation or studies of biomolecules in artificial environment. Example: Laboratory studies using test tubes, pippets, petri dish, etc.

In Vivo
Experimentation or studies of biomolecules in live organisms or live cells. Example: Animal Testing and Clinical Trails

Information Technology
Branch of engineering that deals with the use of computers and telecommunications to retrieve and store and transmit information.

Intelligent Databases
A full text database that employs artificial intelligence or other sort of techniques so as to retrieve the most relevant information possible.

Insertion (Mutation)
A type of mutation where one or more bases are added or inserted into a DNA / RNA sequence.

Inversion (Mutation)
A kind of mutation that reverses the order of the genes in a particular location of a chromosome.

Java
A programming language that is widely used by programmers. It is platform independent and object oriented.

Lysosomes
Sphere like structures found in the cytoplasm of a cell that contains enzymes to break down the food materials.
M
Metabolism
Biochemical cellular functions involved in maintaining life such as growth, reproduction, etc.

Metabolites
Substances involved in the metabolic activities of an organism.

Metabolomics
The study of metabolites present in a person's body at different times.

Micro-array
A laboratory technology that is used to study if a large number of genes are switched on or off in a tissue sample. A microarray is also called a DNA chip or a gene chip. It is the technology used to obtain a genomic profile.

Mutation
Change in the DNA sequence that are caused by radiation, viruses, chemicals as well as errors that occur during replication.

Machine learning
Machine learning is the art of science of getting computers to perform without being explicitly programmed.

MATLAB
MATLAB is a high level language and interactive background for programming, numerical computation and visualization. MATLAB provides the environment to develop the algorithms, analyze the data and create the models and applications.

N
Neural Networks
Neural Networks is a method for computational data analysis that mimics the function of a brain and has the capacity of learn something from the existing information.

Nucleotide
Nucleotides are molecules that when joined together form the DNA and RNA. Each nucleotide is made up of a base, a sugar and a phosphate.

Nucleus
A membrane bound compartment inside the eukaryotic cells that contain the genetic material (DNA) of the cell.

O
Organ
A largest part of an organism that is composed of tissues that perform similar functions.

Organism
A living thing or a living system that has the ability to function on its own. Examples: Humans, Plants, Bacteria
Pathogenesis
The origin or development of a disease or disorder.

Peptide
A small protein fragment.

Peptide Bond
The primary linkage of all aminoacids and protein structures. It is a chemical bond between the carboxyl group and the aminogroup that form the protein.

Phenotype
An observable characteristics or traits of an organism such as weight, height, development, presence or absence of a disease, etc. An organism's phenotype is greatly determined by its genotype.

Phylogenetics
The study of relationships among different groups of organisms.

Protein
Protein is a molecule composed of a large, linear sequence of aminoacids. Proteins are the building blocks of life that determines the various processes of a cell.

Protein Synthesis
Protein Synthesis is a process of producing proteins inside a cell. Proteins are formed from the DNA sequence inside the cell through processes called Transcription and Translation.

Proteome
The proteome is the entire set of proteins expressed by a genome, cell, tissue or organism. More specifically, it is the set of expressed proteins in a given type of cells or an organism at a given time under defined conditions.

Proteomics
Proteomics is the study of the entire proteome or all proteins of an organism.

Principal component analysis (PCA)
Principal component analysis (PCA) is a mathematical procedure that uses an orthogonal transformation to convert a set of annotations of possible correlated variables into a set of values of linearly uncorrelated variables.

Regular Expression
A language for specifying patterns that match a sequence of characters.

Regulatory Region
A region of DNA that controls the expression of gene.

Replication
Replication is the process by which a DNA makes a copy of itself during cell division.
Retrotransposon
A transposable or mobile elements that move from location to location inside a genome and has the ability to transcribe its DNA to RNA and vice-versa.

Reverse Transcription
The process of creating a double stranded DNA back from the single stranded RNA.

Ribosomes
Components of cell that are involved in translation. Ribosomes translate the mRNA sequence into its corresponding protein sequence.

RNA
Ribonucleic acid (RNA) is a molecule similar to DNA that contains Uracil instead of Thymine. DNA after transcription results in RNA.

RNA Polymerase
An enzyme that synthesizes the RNA sequence from the template DNA sequence.

Sequence Analysis
The process by the nucleotide or protein sequence is analyzed by various methods such as sequence alignment, database searches, other bioinformatic techniques etc.

Sequencing
The procedure by which the order of the nucleotides (in DNA or RNA) or the aminoacids (in proteins) is determined.

Sperm
A male reproductive cell of an organism.

Supervised learning
Supervised learning produces a function that can map the inputs to desired outputs (also called labels.

Thymine
Thymine (T) is one of the 4 bases and is found only in DNA; Pairs with Adenine (A). Thymine is replaced with Uracil (U) in RNA.

Tissue
A mass of cells specialized to perform a particular function.

Transcription
Transcription or RNA synthesis is the process of forming a mRNA sequence from a template DNA sequence.

Transcriptome
The set of all RNA molecules produced in a group of cells.
Translation
Translation is the process by which the mRNA produced by transcription process is further decoded into an equivalent aminoacid sequence.

Tumor
A tumor is recognized as the existence of an abnormal mass of tissue with a capacity for progressive growth.

U
Ultra-centrifugation
An experiment in microbiology used to separate parts of a cell such as DNA, ribosomes, etc. for their analysis.

Uracil
Uracil (U) is one of the 4 bases and is found only in RNA; Pairs with Adenine (A). DNA contains Thymine in its place.

User Interface
The part of the software application that the user sees and interacts with.

Unsupervised learning
Unsupervised learning represents a set of inputs, such as clustering. In this model, labels are unknown during training.

V
Virus
A small infectious non-living organism that resides in the cells of other organisms.

Viruses are the causative agents of several diseases and cancers.

W
Web server
A software that send out web pages containing information in response for the requests from web browsers.

Weka
The Weka machine learning workbench offers a general-purpose environment for automatic classification, clustering, feature selection and regression extensive data mining problems in bioinformatics research.

Web-Weka
A new web based software developed WEB-WEKA 1.0 for performance evaluation of various classifiers where the users can just submit their data set online and evaluate the results on the fly.
APPENDIX – C

Supplementary Material

- Sample Dataset
- Web-Weka1.0 code