Volume 5, Issue 5, May 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Comparative Study on Classification Algorithms in Data Mining Using Microarray Dataset of Colon Cancer

R. Porkodi*
Assistant Professor
Department of Computer Science
Bharathiar University, India

G. Suganya
PG Student
Department of Computer Science
Bharathiar University, India

Abstract*— Data mining is the non-trivial extraction of implicit, previously unknown and potentially useful information from the data. The Research area in data mining techniques includes Classification, Clustering, Association rule and Neural Network, Regression, etc. Classification is the process of finding a model that describes and distinguishes data classes or concepts. The purpose of being able to use the model is to predict the class of objects whose class label is unknown. The classification process uses data of any kind, provided the dataset has some class variables. The classification of microarray datasets in biomedical field plays a vital role in the past decade of data mining research. This paper classifies the colon cancer microarray dataset in bioinformatics using five different classification algorithms namely Naïve Bayesian, K-Nearest Neighbors, Support Vector Machine, Random Forest and Neural Network. The performance of these classification algorithms are calculated based on the Performance measures namely Classification Accuracy (CA), Precision and Recall. The experimental result shows that the highest accuracy is found in both KNN and Neural Network classifier among all other classification algorithms.

Keywords*— Classification, Naive Bayes, KNN, Neural Network, Random Forest, SVM

I. INTRODUCTION

Data mining refers to extracting or "mining" knowledge from large amounts of data. Data mining is the search for the relationships and global patterns that exist in large databases but are hidden among large amounts of data. Data mining is becoming strategically important tool for many organizations.

A microarray database is a repository containing microarray gene expression data. The key uses of a microarray database are to store the measurement data, manage a searchable index, and make the data available to other applications for analysis and interpretation. The models that determine to solve a problem are classified as Predictive and Descriptive. The essential process of Knowledge Discovery is the conversion of data into knowledge in order to help in decision making, referred to as data mining. Knowledge Discovery process consists of an iterative sequence of data cleaning, data integration, data selection, data mining, pattern recognition and knowledge presentation. Classification is the process of finding a model that describes and distinguishes data classes or concepts. The purpose of being able to use the model is to predict the class of objects whose class label is unknown. In this project the Models are built for Traditional Classification, Multi classification. Traditional classification is the application of classifiers. Multi classifications are used to combine two classifiers. The performance of these two models is calculated based on the Performance measures Accuracy (CA), Precision and Recall.

The paper organized as follows: section 1 describes the introduction on data mining with colon cancer, section 2 describes the literature review, section 3 describes the various classification and algorithms, section 4 gives the summary of breast cancer dataset, and section 5 discusses the experimental results and finally the paper is concluded in section 6.

II. LITERATURE REVIEW

D.S.V.G.K.Kaladhar [1] had proposed that Classification of colon cancer dataset, in which Logistics, Ibk, Kstar, NNge, ADTree ,Random Forest Algorithms show 100 % correctly classified instances, followed by Navie Bayes and PART with 97.22 %, Simple Cart and Zero R has shown the least with 50 % of correctly classified instances. Kappa Statistic for Logistics, Ibk, Kstar, NNge, ADTree, and Random Forest has shown Maximum. Mean absolute error and Root mean squared error are shown low for Logistics, Kstar and NNge. Using various Classification algorithms the cancer dataset can be easily analyzed.

Ramadevi Yellasiri, C.R.Rao [2] had proposed a new classification model called Rough Set Classifier for classifying the voluminous protein data based on structural and functional properties of protein. This model is fast and accurate and it can be used as an efficient classification tool than the others. This Classifier provides 97.7% accuracy. It is a hybridized tool comprising Sequence Arithmetic, Concept Lattice and Rough Set Theory. It can reduce the domain search space to 9% without losing the potentiality for the classification of proteins. The information about the family is identified using special arithmetic and utilizes it for reducing the domain search space is proposed. The rules are generated and stored in Sequence Arithmetic database.

Huilin Xiong And Xue –Wen Chen [3] says the new approach called kernel function, which improves the performance of the classifier in genetic data. The efficiency of a kernel approach has been probed in which it is depends upon on optimizing a data -dependent kernel model. The K-nearest-neighbor (KNN) and support vector machine (SVM) could be used as a classifier for performance analysis.

K. Srinivas, B. Kavitha Rani and Dr. A. Goverdhann [4] had developed the classification based data mining techniques such as Rule Based, Decision tree, Naïve Bayes and Artificial Neural Network to the massive volume of healthcare data. Using an age, sex, blood pressure and blood sugar medical profiles it can predict the likelihoodof patients getting a heart disease.

David B.fogel et al. [5] had presented the evolving neural networks for detecting breast cancer and the related works used for breast cancer diagnosis using back propagation method with multilayer perceptron. In contrast to back propagation found that evolution computational method and algorithms were used often, perform more classic optimization techniques.

Dr.S.Santhosh baboo and S.Sasikala [6] had done a survey on data mining techniques for gene selection classification. This article dealt with most used data mining techniques for gene selection and cancer classification; particularly they have focused on four main emerging fields. They are neural network based algorithms, machine learning algorithms, genetic algorithm and cluster based algorithms and they have specified future improvement in this field.

D. Lavanya [7] had considered Decision tree classifier-CART with and without feature selection in terms of accuracy, time to build a model and size of the tree on various Breast Cancer Datasets are observed. From the results it is clear that, though we considered only breast cancer datasets, a specific feature selection may not lead to the best accuracy for all Breast Cancer Datasets. The best feature selection method for a particular dataset depends on the number of attributes, attribute type and instances.

V. Krishnaiah et al [8] had developed a prototype lung cancer disease prediction system using data mining classification techniques. The most effective model to predict patients with Lung cancer disease appears to be Naïve Bayes followed by IF-THEN rule, Decision Trees and Neural Network. For Diagnosis of Lung Cancer Disease Naïve Bayes observes better results and fared better than Decision Trees.

R. Geetha Ramani had considered the data mining application on medical research for a predicting and discovering pattern base on detected symptom on health condition for process take a mammography, dermatology, orthopedic thyroids for data pre-processing execute classification for clinical test data lode test data for verification for classifier Malady classification. They support decision tree generate by the quinlan's algorithm is smaller than the decision tree by the random tree classification technique.

Endo et al [9] had implemented common machine learning algorithms to predict survival rate of breast cancer patient. This study is based upon data of the SEER program with high rate of positive examples (18.5 %). Logistic regression had the highest accuracy; artificial neural network showed the highest specificity and J48 decision trees model had the best sensitivity.

Manaswini Pradhan [10] had proposed an Artificial Neural Network (ANN) based classification model as one of the powerful method in intelligent field for classifying diabetic patients into two classes. For achieving better results, genetic algorithm (GA) is used for feature selection. The designed models are also compared with the Functional Link ANN (FLANN) and several classification systems like NN (nearest neighbor), kNN(k-nearest neighbor), BSS(nearest neighbor with backward sequential selection of feature, MFS1(multiple feature subset) , MFS2(multiple feature subset) for Data classification accuracies. It is revealed from the simulation that our suggested model is performing better compared to NN, kNN, BSS, MFS1, MFS2 and FLANN model and it can be a very good candidate for many real time domain applications as these are simple with good performances.

Sonali Agarwal, G. N. Pandey, and M. D. Tiwari [11] had proposed Support Vector Machines (SVM) is established as a best classifier with maximum accuracy and minimum root mean square error (RMSE). The study also includes a comparative analysis of all Support Vector Machine Kernel types and in this the Radial Basis Kernel is identi fied as a best choice for Support Vector Machine. A Decision tree approach is proposed which may be taken as an important basis of selection of student during any course program. The paper is aimed to develop a faith on Data Mining techniques so that present education and business system may adopt this as a strategic management tool.

Tina R. Patil, Mrs. S. S. Sherekar [12] had proposed that to make comparative evaluation of classifiers NAIVE BAYES AND J48 in the context of bank dataset to maximize true positive rate and minimize false positive rate of defaulters rather than achieving only higher classification accuracy using WEKA tool. The result shows that the efficiency and accuracy of j48 is better than that of Naïve bayes.

III. CLASSIFICATION TECHNIQUES USED IN HEALTH CARE

Classification is one of the most extensively used methods of data mining in healthcare. The classification algorithms on Colon Cancer data can be useful to predict the outcome of some diseases or discover the genetic performance of tumor. Classification model is build relating a predefined set of classes or ideas. The model is constructed by analysing database tuples described by attributes. The classification is used to predict categorical class labels and classify data based on the training set.

Classification techniques in data mining are capable of processing a huge quantity of data. It can predict categorical class labels, classifies data based on training set and class labels. Hence it can be used for classifying newly available data. Thus it can be out lined as a predictable part of data mining and is gaining more popularity. Classification contains two phase such as training phase and testing phase, in training phase every

sample in the training set is assumed to belong to a predefined class. The testing phase is unknown test samples are measured to classify using the model build using the training set. This paper gives the detailed description of five algorithms namely Naïve Bayes, K-Nearest Neighbors, Support Vector Machine, Random Forest and Neural Network, and also presents the comparative study on above mentioned algorithms using colon cancer microarray dataset.

A. Naive Bayes Classifier

Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables in a learning problem. Maximum-likelihood training can be done by evaluating a closed-form expression, which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers.

Bayesian classification is based on Baye's Theorem. Bayesian classifiers are the statistical classifiers. Bayesian classifier is able to predict class membership probabilities such as the probability that a given tuple belongs to a particular class. Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of independence between every pair of features. Given a class variable y and a dependent feature vector $x_1$through x_n, Bayes' theorem states the following relationship:

$$P(y|x_{1,\dots}\ x_{n)=} \quad \frac{P(y)P(x_1,\dots\ x_n|y)}{P(x_1,\dots x_n)}$$

Algorithm

- Use notation P(A) as the probability of an event A and P(A/B) denotes the probability of A conditional on another event B
- H is the hypothesis and E is the evidence then
  P (H/E)= p(E/H).p(H) / P (E)
- Eg. Let H be 'yes' and E is the combination of the attribute values for new day: outlook=sunny, temp=cool, humidity=high, windy=true. Call these for pieces E1,E2,E3 and E4 are independent then
  P (H/E)= p(E1/H).p(E2/H).p(E3/H).p(E4.H).p(H) / p(E)
- Denominator can be eliminated as the final normalizing step when we make the probabilities of different pieces the sum is 1. Thus, p(H/E)=p(E1/H).p(E2/H).p(E3/H).p(E4/H).p(H)

Maximum a Posteriori (MAP) estimation to estimate P(y) and $P(x_i|y)$; the former is then the relative frequency of class y in the training set. The different naive Bayes classifiers differ mainly by the assumptions they make regarding the distribution of $P(x_i|y)$. It includes the parameters such that set the values for Prior as Relative Frequency and Parameter for m-estimate as 2.0. The size of LOESS window is 0.5 and LOESS sample points as 100.

B. K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) classification divides data into a test set and a training set. For each row of the test set, the K nearest training set objects are found, and the classification is determined by majority vote with ties broken at random. If there are ties for the K^{th} nearest vector, all candidates are included in the vote. It uses the Backward Elimination.KNN technique assumes that the entire training set includes not only the data in the set but also the desired classification for each item. When a classification is to be made for a new item, its distance to each item in the training set must be determined. Only the K closest entries in the training set are considered further. The new item is then placed in the class that contains the most items from this set of K closest items. The Euclidean distance between $X=(x_1, x_2, x_3,\dots x_n)$ and $Y=(y_1,y_2, y_3,\dots y_n)$ is defined as:

$$D(X,Y) = \sqrt{\sum (x_1 - y_1)^2}$$

Algorithm

The training phase for KNN consists of simply storing all known instances and their class labels. A tabular representation or a specialized structure can be used. If we want to tune the value of 'k' and/or perform feature selection, n-fold cross-validation can be used on the training dataset. The testing phase for a new instance 't', given a known set 'I' is as follows:

1. Compute the distance between 't' and each instance in 'I'
2. Sort the distances in increasing numerical order and pick the first 'k' elements
3. Compute and return the most frequent class in the 'k' nearest neighbors, optionally weighting each instance's class by the inverse of its distance to 't'

In *k-NN classification*, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its *k* nearest neighbors. If $k = 1$, then the object is simply assigned to the class of that single nearest neighbor. It includes the parameters such that set the number of neighbors as 5 and in the metrics, set as Euclidean distance algorithm and initialized the attributes type as the normalize continuous attributes.

C. Support Vector Machine

Support Vector Machines are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs

into high-dimensional feature spaces. A support vector machine constructs a hyper plane or set of hyper planes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyper plane that has the largest distance to the nearest training-data point of any class, since in general the larger the margin the lower the generalization error of the classifier.

Algorithm

Step 1: Let "N" be the number of train in g cases, and let "M" be the number of variables in the classifier. Choose as input variables, to be used to determine the decision at a node of the tree; m should be much less than M.

Step 2: Recurse a training set for this tree by choosing N times with replacement from all N available training cases. Rest of the cases to be estimated as error of the tree by predicting their classes.

Step 3: For each node in the tree, randomly choose m variables, which should be based on the decision at that node.

Step 4: Calculate the best split based on these m variables in the training set. The value of m remains to be constant during forest growing. Random forest is sensitive to the value of m.

Step 5: Each tree is grown to the largest extent possible, into many classification trees without pruning, in constructing a normal tree classifier. It include the parameters such that, set the SVM type as C-SVM and from the kernel select RBF, set the value for g as 0.000000 and set the numerical tolerance as 0.0010.

D. Random Forest

Random forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean prediction of the individual trees. Random forests correct for decision trees' habit of over fitting to their training set.

Algorithm

Decision trees are a popular method for various machine learning tasks. Tree learning "come[s] closest to meeting the requirements for serving as an off-the-shelf procedure for data mining", say Hastie *et al.*, because it is invariant under scaling and various other transformations of feature values, is robust to inclusion of irrelevant features, and produces inspect able models. However, they are seldom accurate.

In particular, trees that are grown very deep tend to learn highly irregular patterns: they over fit their training sets, because they have low bias, but very high variance. Random forests are a way of averaging multiple deep decision trees, trained on different parts of the same training set, with the goal of reducing the variance. This comes at the expense of a small increase in the bias and some loss of interpretability, but generally greatly boosts the performance of the final model. It includes the parameters such that set the number of trees in forest as 10 and the stop splitting nodes with as 5 of fewer instances.

E. Artificial Neural Network

Artificial neural networks are a family of statistical learning algorithms inspired by biological neural networks and are used to estimate or approximate functions that can depend on a large number of inputs and are generally unknown. Artificial neural networks are generally presented as systems of interconnected "neurons" which can compute values from inputs, and are capable of machine learning as well as pattern recognition thanks to their adaptive nature.

Network function

An ANN is typically defined by three types of parameters: (i) The interconnection pattern between the different layers of neurons (ii) The learning process for updating the weights of the interconnections (iii) The activation function that converts a neuron's weighted input to its output activation. Mathematically, a neuron's network function f(x) is defined as a composition of other functions $g_i(x)$, which can further be defined as a composition of other functions. This can be conveniently represented as a network structure, with arrows depicting the dependencies between variables.

The first view is the functional view: the input x is transformed into a 3-dimensional vector h, which is then transformed into a 2-dimensional vector g, which is finally transformed into f this view is most commonly encountered in the context of optimization.

The second view is the probabilistic view: the random variable F= f(g) depends upon the random variable G=g(H), which depends upon H=h(X), which depends upon the random variable X. This view is most commonly encountered in the context of graphical models. The two views are largely equivalent. In either case, for this particular network architecture, the components of individual layers are independent of each other. This naturally enables a degree of parallelism in the implementation. It include the parameters such that set the hidden layer neurons as 20, regularization factor as 1.0 and max iterations as 300.

IV. EXPERIMENTAL RESULTS

DATASET

The Colon Cancer dataset is used for this project work is taken from gene expression of Princeton University, New Jersey, USA used in study of Transcriptional Gene Expression Profiles of Colerectal Adenoma, Adenocarcinoma and Normal Tissue by Daniel A.Notterman et al, 2001[16]. Table 1 describes that the dataset contains the information about 62 instances and 2000 attributes are used for this classification task.

TABLE 1: DATASET DESCRIPTION

Dataset name	No. of. Attributes	No. of. Instances	No.of Class
Colon Cancer	2000	62	2(Negative, Positive)

The Table 2 describes the Performance comparison for the Negative class using all the five classifier techniques with the help of Accuracy, Precision and Recall. Based on the experimental results, it is observed that the KNN and Neural Network are considered as good classifier and the accuracy is obtained by 1.The next highest Accuracy obtained by Random Forest and Naïve Bayes and the accuracy values are 0.9516 and 0.9194. The lowest accuracy obtained by Support Vector Machine and its value is 0.8226. Similarly the Precision and Recall of the KNN and Neural Network are 1 and apart from Support Vector Machine all have the Precision values as 1. The Recall of Random Forest and Support Vector Machine are 0.925. The lowest Precision value obtained by Naïve Bayes is 0.875.

TABLE 2: PERFORMANCE COMPARISON FOR NEGATIVE CLASS

Techniques	CA	Sensitivity	Specificity	AUC	Precision	Recall
Naive Bayes	0.9194	0.875	1	0.9625	1	0.875
KNN	1	1	1	1	1	1
Random Forest	0.9516	0.925	1	0.9943	1	0.925
Neural Network	1	1	1	1	1	1
SVM	0.8226	0.925	0.6364	0.9545	0.8222	0.925

The Table 2 gives the details of the different classification techniques Accuracy, Sensitivity, Specificity, and Area under the ROC Analysis, Precision and Recall. The Fig. 1 represents the Performance comparison for Negative which is listed in Table 1.

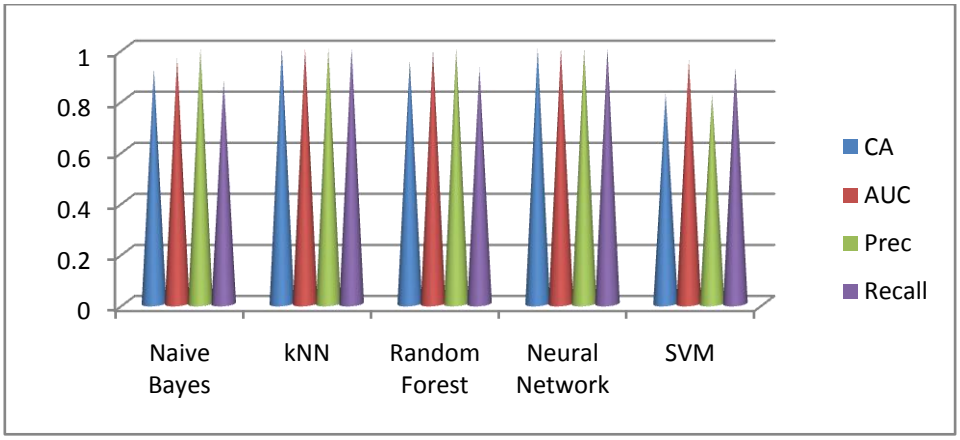

Fig. 1 Performance comparison for Negative class

The Table 3 describes that Performance comparison for Positive, Accuracy of KNN and Neural Network are 1 and the other algorithm Accuracy of Random Forest is 0.9516 and Naïve Bayes is 0.9194 and Support Vector Machine is 0.8226.Similary the Precision and Recall of the KNN and Neural Network are 1 and expect from Support Vector Machine all have the Recall values as 1. The lowest Recall value is Support Vector Machine and its value is 0.6364.The next highest value of Precision is Random Forest and its value is 0.88 and Support Vector Machine is 0.8235. The lowest Precision is Naïve Bayes and its value is 0.8148.

TABLE 3: PERFORMANCE COMPARISON FOR POSITIVE CLASS

Techniques	CA	Sensitivity	Specificity	AUC	Precision	Recall
Naive Bayes	0.9194	1	0.875	0.9625	0.8148	1
KNN	1	1	1	1	1	1
Random Forest	0.9516	1	0.925	0.9943	0.88	1
Neural Network	1	1	1	1	1	1
SVM	0.8226	0.6364	0.925	0.9545	0.8235	0.6364

Table 3 gives the details of the different classification techniques Accuracy, Sensitivity, Specificity, and Area under the ROC Analysis, Precision and Recall. The Fig. 2 represents the Performance comparison for Positive which is listed in Table 3.

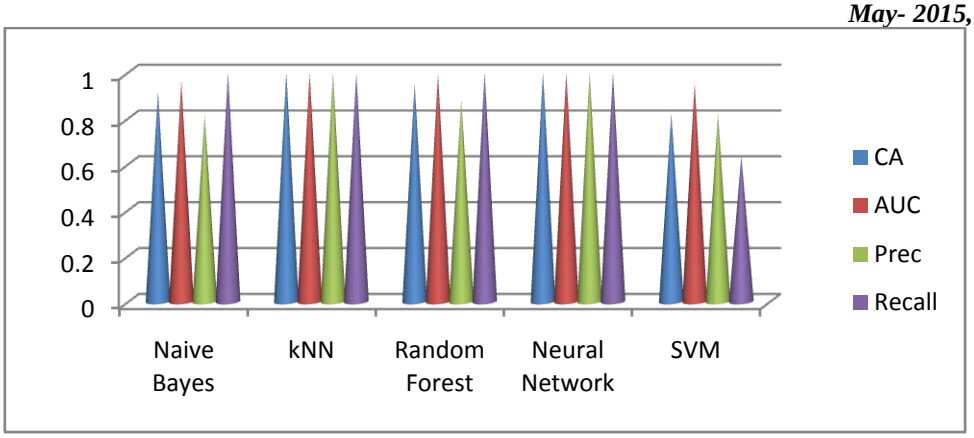

Fig. 2 Performance comparison for Positive class

Table 4 shows the result of Confusion Matrix for five different classification algorithms. In K-Nearest Neighbors and Random Forest produces the accurate confusion matrix for the Colon Cancer dataset.

TABLE 4: CONFUSION MATRIX

Confusion Matrix	**Naïve Bayes**		**KNN**		**Random Forest**		**Neural Network**		**SVM**	
Positive	22	0	22	0	22	0	22	0	14	8
Negative	5	35	0	40	3	37	0	40	3	37

Fig. 3 shows the result of Correctly Classified instances for both positive and negative class of all the five different classification algorithms. From this the KNN and Neural Network classifier produces the 100% of correctly classified instances and the next highest correctly classified instances are Random Forest and Support Vector Machine is 95% and the lowest correctly classified instances is Naïve Bayes classifier has the value as 90%.

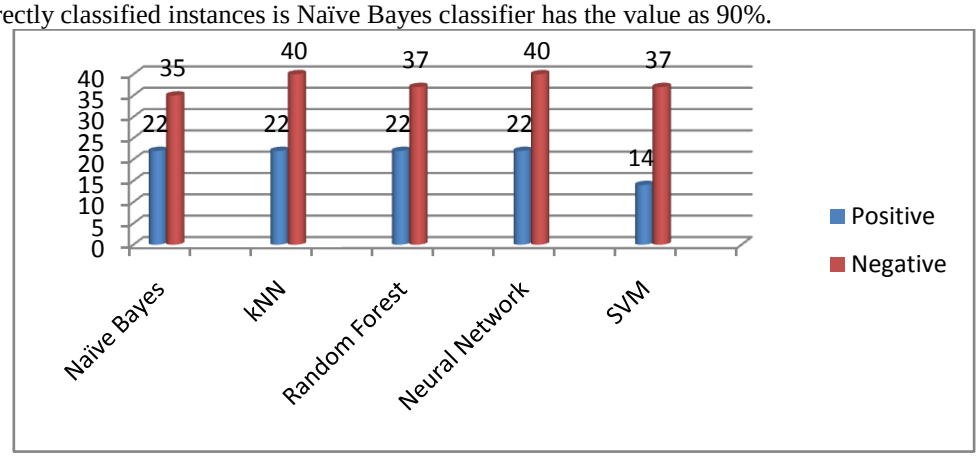

Fig. 3 Correctly classified instances for positive and negative classes

Fig. 4 shows the result of Incorrectly Classified instances for both positive and negative class of all the five different classification algorithms. From this the KNN and Neural Network classifier produces the 0% of incorrectly classified instances and the next highest incorrectly classified instances for positive and negative class of Support Vector Machine is 8 and 3. Incorrectly classified instance for Negative class of Random Forest is 3 and no positive incorrectly classified instances. The Naïve Bayes classifier of Negative class is 5.

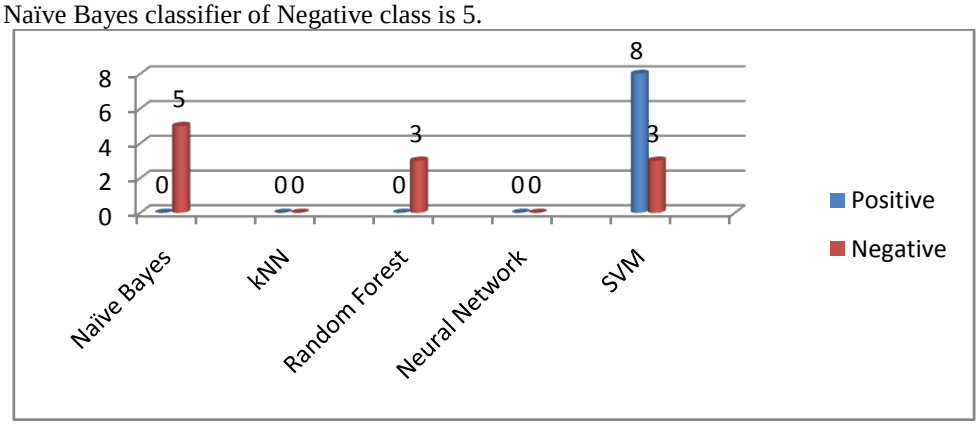

Fig 4: Incorrectly classified instances for positive and negative classes

The Table 5 represents the classification tree for Negative class for Colon Cancer dataset. In the Colon Cancer dataset has 62 instances and it has the P (Class) value as 0.645 and its P (Target) value as 0.645. If the value of attribute1671 <=59.828 then it have 14 instances of Positive class and it has the P (Class) value as 1 and its P (Target) value as 0. If the attribute1671 > 59.828 then it have 48 instances of Negative class and it has the P (Class) value as 0.833 and its P (Target) value as 0.833. In the attribute590 <=289.948 then it has the 18 instances of negative class and it has the P (Class) value as 0.556 and its P (Target) values as 0.556. In the attribute472<=485.623 then it has the 11 instances of negative class and it has the P (Class) value as 0.909 and its P (Target) values as 0.909. In the attribute472 >485.623 then it has the 7 instances of positive class and it has the P (Class) value as 1 and its P (Target) values as 0. If the attribute590 >289.948 then it has the 30 instances of negative class and it has the P (Class) value as 1 and its P (Target) values as 0.

TABLE 5: CLASSIFICATION TREE FOR NEGATIVE CLASS

Classification Tree	Class	P(Class)	P(Target)	# Inst	Distribution (rel)	Distribution (abs)
	Negative	0.645	0.645	62	0.355:0.645	22:40
attribute1671 <=59.828	Positive	1	0	14	1.000:0.000	14:00
attribute1671 >59.828	Negative	0.833	0.833	48	0.167:0.833	8:40
attribute590 <=289.948	Negative	0.556	0.556	18	0.444:0.556	8:10
attribute472 <=485.623	Negative	0.909	0.909	11	0.091:0.909	1:10
attribute472 >485.623	Positive	1	0	7	1.000:0.000	7:00
attribute590 >289.948	Negative	1	1	30	0.000:1.000	0:30

Table 5 gives the classification Tree for Negative class contains Class, P (Class), P (Target), Number of instances and Distribution value for both real and absolute. The microarray data visualization is the next predominant research area in data mining. The "Orange" data mining tool provides many forms of visualizations for micro array dataset. The data profile widget in Orange provides the visualized representation of gene expression profiles in dataset. The Fig. 5 represents Data Profile for Negative class which is listed in Table 1.

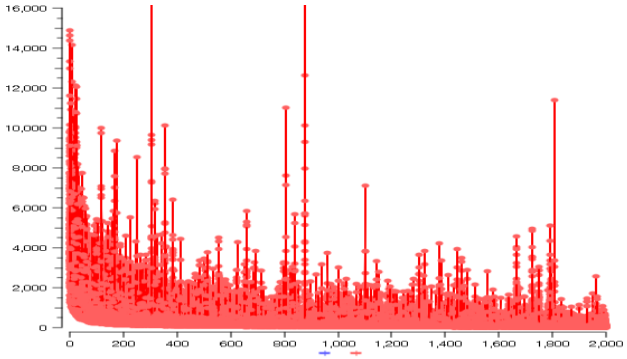

Fig. 5 Data Profile- Negative

The Table 6 represents the classification tree for Positive class for Colon Cancer dataset. In the Colon Cancer dataset has 62 instances and it has the P (Class) value as 0.645 and its P (Target) value as 0.355. If the attribute1671 <=59.828 then it have 14 instances of Positive class and it has the P (Class) and P (Target) value are 1. If the attribute1671 > 59.828 then it have 48 instances of Negative class and it has the P (Class) value as 0.833 and its P (Target) value as 0167. In the attribute590 <=289.948 then it has the 18 instances of negative class and it has the P (Class) value as 0.556 and its P (Target) values as 0.444. In the attribute472<=485.623 then it has the 11 instances of negative class and it has the P (Class) value as 0.909 and its P (Target) values as 0.091. In the attribute472 >485.623 then it has the 7 instances of positive class and it has the P (Class) and P (Target) values are 1. In the attribute590 >289.948 then it has the 30 instances of negative class and it has the P (Class) value as 1 and its P (Target) values as 0.

TABLE 6: CLASSIFICATION TREE FOR POSITIVE CLASS

Classification Tree	Class	P(Class)	P(Target)	# Inst	Distribution (rel)	Distribution (abs)
	Negative	0.645	0.355	62	0.355:0.645	22:40
attribute1671 <=59.828	Positive	1	1	14	1.000:0.000	14:00

	attribute1671 >59.828			Negative	0.833	0.167	48	0.167:0.833	8:40
		attribute590 <=289.948		Negative	0.556	0.444	18	0.444:0.556	8:10
			attribute472 <=485.623	Negative	0.909	0.091	11	0.091:0.909	1:10
			attribute472 >485.623	Positive	1	1	7	1.000:0.000	7:00
		attribute590 >289.948		Negative	1	0	30	0.000:1.000	0:30

Table 6 gives the classification Tree for Positive class contains Class, P (Class), P (Target), Number of instances and Distribution value for both real and absolute. The microarray data visualization is the next predominant research area in data mining. The "Orange" data mining tool provides many forms of visualizations for micro array dataset. The data profile widget in Orange provides the visualized representation of gene expression profiles in dataset. The Fig 6 represents Data profile for Positive class which is listed in Table 1.

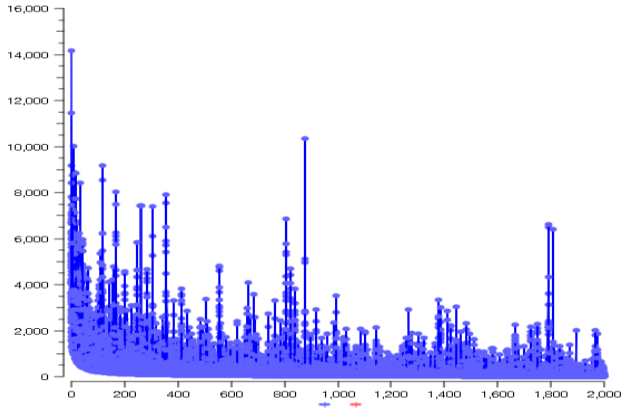

Fig. 6 Data Profile- Positive

Table 1 gives the Colon Cancer Dataset for both Positive class and Negative class instances. The Fig. 7 represents the Performances of Both Negative and Positive class instances which are listed in Table 6 and Table 7.

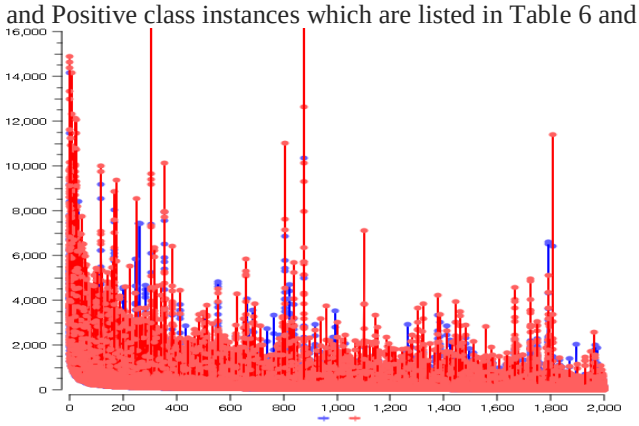

Fig. 7 Data Profile- Both Positive and Negative

Table 1 gives the number of instances and attributes lists. The next visualization widget in orange is differential expression, which is used to plot the differentially expressed genes in Colon Cancer data as shown in Fig.7

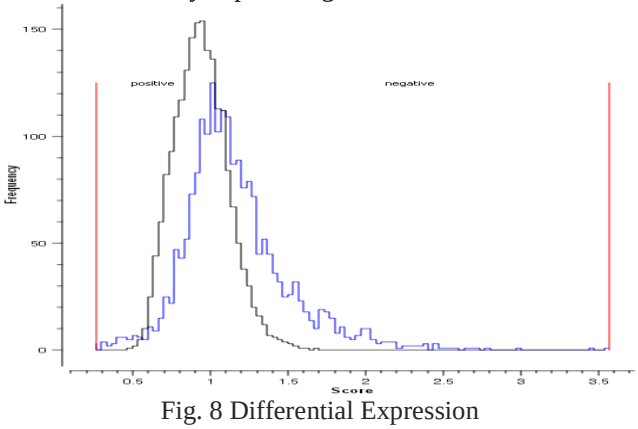

Fig. 8 Differential Expression

V. CONCLUSION

In this paper five classification methods were used to build the classifier models. The work demonstrates the advantages of applying the data mining techniques including traditional algorithm and multi class classifier to the colon dataset for the purpose of finding the best classifier for thus dataset. The colon data set have been applied to the classification algorithms and the best performing algorithm is chosen from the analysis to find the best classification algorithm based on Accuracy (CA), Precision and Recall.

This paper compared on various supervised learning algorithms to predict the best classifier and the highest accuracy of the classifier is identified. The experimental result shows that the highest accuracy is found in both KNN and Neural Network classifier gives the result as 1 and it produces 100% of correctly classified instances among all other classification algorithms. The next highest accuracy is found in Random Forest and its accuracy value is 0.9516 with the lower error rate 30%. The Naïve Bayes classifier produce the lowest incorrectly classified instances for Negative class with 50 %. It has the next highest accuracy value is 0.9194.The Support Vector Machine classifier has the lowest accuracy value with 0.8226 and it has the 20% of incorrectly classified instances of Positive class and 25% of Negative class. KNN and Neural Network obtained better accuracy for Colon Cancer dataset. The KNN and Neural Network classifier produces the good accuracy than the Support Vector Machine, Random Forest and Naïve Bayes classifier.

In future, the work can be extended to add other classification algorithms by including several Optimization Techniques. It is also decided to improve the work by identifying correlation of genes present in the Colon Cancer Dataset using go annotation.

REFERENCES

[1] D. S.V.G.K.Kaladhar, V.Nageswara Rao, Varahalarao Vadlamudi, "The Elements of Statistical Learning in Colon Cancer Datasets: Data Mining, Inference and Prediction",

[2] Ramadevi Yellasiri, C.R.Rao, "Rough Set Protein Classifier", Journal of Theoretical and Applied Information Technology, (2009).

[3] Huilin Xiong And Xue-Wen Chen,"Optimized KernelMachines for Cancer Classification Using gene Expression Data", Proceedings Of The 2005 IEEE Symposium On Computational Intelligence in Bioinformatics and Computational Biology, Pp.1-7, 2005.

[4] K. Srinivas , B. Kavitha Rani and Dr. A. Govrdhan, ―Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks‖ International Journal on Computer Science and Engineering (2010).

[5] David B.Fogel, Eugene C, Wasson, Edward M.Boughton "Evolving neural networks for detecting breast cancer". 1995 Elsevier Science Ireland Ltd.

[6] Dr.Santhosh baboo, S.Sasikala "A Survey on data mining techniques in gene selection and cancer classification"-April 2010 International journal of Computer science and information technology.

[7] D.Lavanya and Dr.K.Usha Rani," Analysis of feature selection with classification Breast cancer datasets", Vol.2-No.5, oct-nov: 2011, Pg.no:756-763.

[8] Krishnaiah "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques" International Journal of Computer Science and Information Technologies, Vol. 4 (1) 2013, 39 – 45 www.ijcsit.Com ISSN: 0975-9646.

[9] Endo, T. Shibata and H. Tanaka (2008), Comparison of seven algorithms to predict breast cancer survival, Biomedical Soft Computing and Human Sciences, vol.13, pp.11-[16].

[10] Manaswini Pradhan, Dr. Ranjit Kumar Sahu, "Predict the onset of diabetes disease using Artificial Neural Network (ANN)".

[11] Sonali Agarwal, G. N. Pandey, and M. D. Tiwari ,"Data Mining in Education: Data Classification and Decision Tree Approach".

[12] Tina R. Patil, Mrs. S. S. Sherekar ,"Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification".

[13] N. Poomani, R.Porkodi, "A Comparative Study of Classification Algorithms for Breast Cancer"

[14] Fayyad, Piatetsky-Shapiro, Smyth, and Uthurusamy: "Advances in Knowledge Discovery and Data Mining", AAAI/MIT Press 1996.Jiawei Han and Micheline Kamber,"Data mining: concepts and techniques", San Francisco:Morgan Kaufmann Publishers, 2001.

[15] Xin Yao, Yong Liu "Neural Networks for Breast Cancer Diagnosis" 01999 IEEE.

[16] A. Notterman Daniel,nUri Alon, Alexander J.Sierk, Arnold J. Levine, "Transcriptional gene expression profiles of colorectal adenoma, ad enocarcinoma, and normal tissue examined by oligonucleotide arrays", Ccancer Research.

[17] K. Srinivas , B. Kavitha Rani and Dr. A. Govrdhan, ―Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks‖ International Journal on Computer Science and Engineering (2010).

[18] David B.Fogel, Eugene C, Wasson, Edward M.Boughton "Evolving neural networks for detecting breast cancer". 1995 Elsevier Science Ireland Ltd.

[19] Dr.Santhosh baboo, S.Sasikala "A Survey on data mining techniques in gene selection and cancer classification"-April 2010 International journal of Computer science and information technology.

[20] Aruna, Dr S.P. Rajagopalan and L.V. Nandakishore, 2011 Knowledge Based Analysis Of Various Statistical Tools In Detecting Breast Cancer.

[21] D.Lavanya and Dr.K.Usha Rani," Analysis of feature selection with classification Breast cancer datasets", Vol.2-No.5, oct-nov: 2011, Pg.no:756-763.

[22] Delen Dursun, Walker Glenn and Kadam Amit, "Predicting breast cancer survivability: a comparison of three data mining methods," Artificial Intelligence in Medicine, vol. 34, pgno. 113-127, June 2005.

[23] Krishnaiah "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques" International Journal of Computer Science and Information Technologies, Vol. 4 (1) 2013, 39 – 45 www.ijcsit.Com ISSN: 0975-9646.

[24] ShomonaGracia Jacob, R. GeethaRamani ―Mining of Classification patterns in clinical data through data mining algo‖ access from IEEE.

[25] Padmavati J., "A Comparative study on Breast Cancer Prediction Using RBF and MLP," International Journal of Scientific & Engineering Research, vol. 2, Jan. 2011.

