DOI: 10.21917/ijsc.2018.0251

ENSEMBLE CLASSIFICATION BASED MICROARRAY GENE RETRIEVAL SYSTEM

Thomas Scaria[1] and T. Christopher[2]

[1]Department of Computer Science, St. Pius X College, India

[2]Department of Information Technology, Government Arts College, Coimbatore, India

Abstract

Data mining plays an important role in the process of classifying between the normal and the cancerous samples by utilizing microarray gene data. As this classification process is related to the human lives, greater sensitivity and specificity rates are mandatory. Taking this challenge into account, this work presents a technique to classify between the normal and cancerous samples by means of efficient feature selection and classification. The process of feature selection is achieved by Information Gain Ratio (IGR) and the selected features are forwarded to the classification process, which is achieved by ensemble classification. The classifiers being employed to attain ensemble classification are k-Nearest Neighbour (k-NN), Support Vector Machine (SVM) and Extreme Learning Machine (ELM). The performance of the proposed approach is analysed with respect to three different datasets such as Leukemia, Colon and Breast cancer in terms of accuracy, sensitivity and specificity. The experimental results prove that the proposed work shows better results, when compared to the existing techniques.

Keywords:

Data Mining, Classification, Feature Selection

1. INTRODUCTION

Data mining is the technology that aims to gain knowledge from the datasets and the gained knowledge is represented in an intelligible fashion. Most of the research processes involve data analysis to come up with a research solution. Due to the advancement in medical science and technology, most of the biological researches rely on data analysis. Data analysis in biomedical field helps the experts to study and understand the patterns of normal and abnormal cases without any special effort.

Cancer is the most life-threatening disease, which is found prevalently these days. Ignorance is the major cause for the progression of this disease. Several computer based technologies contribute in detecting different kinds of cancer. Among them, image processing and data mining techniques prove remarkable contribution in detecting cancer at an earlier stage. Advanced image processing techniques deal with digital images to detect the stage of the cancer. On the other hand, data mining techniques utilize biological data to detect the cancer.

The interrelationship between the gene expression and the health condition of an individual is greater. Hence, this work utilizes the microarray gene expression datasets to distinguish between the normal and the abnormal genes. The normal and the cancerous gene expressions are studied and classified by machine learning algorithms. With the help of these microarray gene expressions, the type of cancer can be determined. The microarray gene expression data can be analysed in three different ways, which are supervised, unsupervised and semi-supervised techniques.

Hence, the microarray technology utilizes the data mining approaches for analysing the microarray data. The microarray technology relies on the DNA (Deoxyribo Nucleic Acid), which is present in the nucleus of the cell. The DNA possesses two different segments called coding and non-coding segments. The coding segment is popularly known as genes, which is a vital component. The advancements of genetic engineering help in visualizing the entire structure of the cell. With the help of DNA microarrays, the genes can be analysed effectively and the expression level of genes is utilized to study the nature of disease. This is accomplished with the help of retrieving similar expression profiles.

The supervised learning technique relies on prior knowledge that has been gained from the dataset. The classification system is trained with the dataset initially, followed by which the classification techniques can differentiate between the classes of the gene expression. Some of the popular classifiers for achieving the task of classification are k-Nearest Neighbour (k-NN) [1], Support Vector Machine (SVM) [2] and Relevance Vector Machine (RVM) [3]. The unsupervised techniques do not require any prior knowledge about the dataset and the related gene expressions are grouped together. Some of the popular gene based clustering techniques are observed in the works of [4-9]. Semi-supervised techniques are the combination of supervised and unsupervised techniques.

This research article aims to present a system for retrieving the microarray gene data of a particular class by employing supervised algorithm based on ensemble classification. The dimensionality of gene expression data is greater, such that it consumes more time to process the data and involves computational complexity. In order to make the microarray gene retrieval system efficient in terms of computation and time consumption, this work reduces the feature dimension by means of Information Gain Ratio (IGR). The IGR selects the necessary features from the dataset and the classification is carried out by ensemble classifier. The classifiers being utilised are k-NN, SVM and Extreme Learning Machine (ELM) respectively. The main contributions of this work are listed below.

- The reduction of feature dimensionality helps in attaining reduced computational and time complexity.
- The feature dimensionality of the microarray gene dataset is reduced by IGR.
- The incorporation of ensemble classification results in enhanced accuracy rates and reduced misclassification results.

The remaining sections of the paper are organized as follows. Related review of literature with respect to microarray gene classification is presented in section 2. The proposed microarray gene retrieval system is described in section 3. The performance of the proposed approach is analysed and the experimental results are presented in section 4. Finally, the concluding points about the work are presented in section 5.

2. REVIEW OF LITERATURE

This section reviews the state-of-the-art related literature with respect to microarray gene classification techniques.

In [10], a multiple objective model is designed on the basis of analytical hierarchy process. A multi-objective heuristic algorithm is proposed, which is an enhancement of Univariate Marginal Distribution Algorithm. This model works by framing two important rules, which are 'Higher rule' 'Fewer rule' and 'Forcibly decrease rule'. The higher and lower rule intends to analyse and sort the gene data. The 'Forcibly decrease rule' produces the better individuals with maximum classification accuracy. This work gives more importance to the classification accuracy, rather than to reduce the gene count.

A centroid based feature discrimination principle for selecting better genes is presented in [11]. The centroid of the class is computed by the kernel based expectation. The feature selection problem is designed as the L1-regularized optimization problem by considering the linear discriminant analysis principle. The centroid of the class is computed by the kernel based technique, which can define the between and within class separability.

In [12], a work to classify between the cancer types is proposed. The proposed technique is based on information gain and genetic algorithm. The features are selected by information gain, followed by the employment of genetic algorithm to reduce features and finally, the classification of cancer is attained by genetic programming. This work concludes that the performance of genetic algorithm maximizes the accuracy rates.

In [13], an unsupervised technique based on Ant Colony Optimization (ACO) algorithm is proposed for selecting genes. This technique applies the ACO algorithm in the filter based approach, in order to maximize the gene relevance and minimize the gene redundancy rates. Additionally, the fitness function being proposed by this work does not require any prior knowledge about the gene dataset. The classification is attained by the filter based approach and is proven to be better than SVM, naïve bayes and decision trees.

A hybrid classification technique is proposed for classifying the gene microarray data [14], which is based on Principal Component Analysis (PCA) and Brain Emotional Learning (BEL) network. This work states that BEL is suitable for high dimensional features. The proposed work proves better results in terms of accuracy. In [15], a technique to classify between the microarray data for drug response is proposed and is based on feature selection and classification. The feature selection of this work is carried out by a metaheuristic approach, which selects the top ranking relevant genes based on max relevance and min redundancy is proposed. The metaheuristic approaches being employed are Particle Swarm Optimization (PSO), Cuckoo Search (CS) and Artificial Bee Colony (ABC) algorithms. The classifiers being employed by this work are k-NN and SVM. This work concludes that the CS algorithm works better than PSO and ABC algorithms respectively.

In [16], a two stage classification model which relies on pre-processing and classification is proposed. ReliefF is utilized in the pre-processing stage, such that the top ranking genes are selected. The selected genes are mapped into a dissimilarity space and the classification process is carried out. The classification results of this work are compared against artificial neural network, SVM and Fisher's linear discriminant classifier by varying the number of genes. In [17], a faster feature selection technique is proposed that is based on recursive feature elimination by simulated annealing and square root. This approach eliminates several features instead of removing a single feature, hence the count of features being removed as the iteration progresses. This technique concludes that this approach of feature reduction does not affect the classification accuracy of the work. In [18], a hybrid feature selection algorithm is proposed for microarray gene expression data. The features are selected from the mutual information maximization and adaptive genetic algorithm.

In [19], a novel gene selection technique is proposed for cancer classification, which is based on genetic algorithm and artificial intelligence. Initially, the dimensionality of the features is reduced by integer coded genetic algorithm. Laplacian and Fisher score are employed to compute similarity between the features. This work is applied over several classifiers and the performance of the proposed gene selection technique is tested. In [20], the microarray gene dataset is classified by means of kernel ridge regression techniques such as wavelet kernel ridge regression and radial basis kernel ridge regression. The features of the microarray gene dataset are reduced by means of modified cat swarm optimization technique. The performance of this approach is compared against simple ridge regression, SVM and random forest classifer.

In [24], the microarray gene retrieval system that relies on Local Fisher Discriminant Analysis (LFDA) and Support Vector Machine (SVM) is proposed. A supervised microarray gene retrieval system based on Kernel Local Fisher Discriminant Analysis (KLFDA) and Extreme Learning Machine (ELM) classifier is presented in [25].

Motivated by the above related works, this article aims to present a gene retrieval system for microarray gene data. The dimensionality of microarray gene data is huge and hence the dimensionality of the data is reduced by selecting the necessary features by means of IGR. The selected features are then processed to classify between the cancer types by means of ensemble classification. The following section elaborates the proposed approach in addition to the overview of the proposed approach.

3. PROPOSED MICROARRAY GENE DATA CLASSIFICATION SYSTEM

The intention of this work is to present a microarray gene data classification system based on Information Gain Ratio (IGR) and the process of classification is carried out by ensemble classifiers. The entire work is segregated into two phases, which are feature selection and classification. The feature selection process selects the relevant and necessary features, which makes the classification process to the point. The IGR is an improvisation of information gain and the main reason for the choice of IGR is that it takes all the unique entities into account. This idea weeds out all the redundant and irrelevant entities from the feature set and conceives all the necessary set of features. This kind of feature selection improves the classification accuracy and reduces the overall time consumption of the work.

The classification process relies on ensemble classifiers, which are k-NN, SVM and ELM classifiers. k-NN is the classical classifier which takes certain number of samples for training and the classification results are returned by taking the nearest neighbours into account. This k-NN classifier is the simplest classifier with minimal computational complexity. SVMs are the promising classifiers which consider a reasonable set of samples for training and SVM can perform non-linear classification. ELM is chosen as one of the ensemble classifiers, owing to its faster learning ability and efficiency. The decisions of all the individual classifiers are collected and the maximum occurred decision is declared as the final decision. This way of classification increases the accuracy rate. The following sub-sections of the work describe the feature selection and classification process.

3.1 FEATURE SELECTION

This phase of the work intends to select the necessary features from the high dimensional dataset. The main reason is to avoid redundant and irrelevant features, which can reduce time and computational complexity considerably. The IGR is employed for selecting the features, which considers the unique entities into account and is computed as follows. In order to calculate IGR, information gain is computed.

$$IG_S = -\left[\frac{r \cdot f(c_1, S)}{|S|}\right] \log_2 \left[\frac{r \cdot f(c_1, S)}{|S|}\right] \tag{1}$$

In the above equation, $\left[\frac{r \cdot f(c_1, S)}{|S|}\right]$ is the probability of the repetitive occurrence of a sample, which is present in the class c_1. Let the feature f contains q unique values, which can be represented as $\{f_1, f_2, f_3, \ldots, f_q\}$. For a dataset with f features, the training dataset is formed as follows $\{c_1, c_2, c_3, \ldots, c_q\}$ and the information gain of the feature is computed by the following formula,

$$IG(F) = \left[\frac{|F_i|}{|F|}\right] \times IG(F_i). \tag{2}$$

The *IGR* of the feature is then computed with the help of information gain and is presented below,

$$IGR(F) = \left[\frac{|IG_S - IG(F)|}{|IG_S + IG(F)|}\right] \times 100. \tag{3}$$

The IGR makes it possible to extract the features with high correlation degree, which makes sense that the features required to distinguish between the samples alone are extracted. Hence, the IGR selects the genes with high correlation degree and the rest are not considered. As soon as the feature selection process is over, the classification process is triggered and is carried out by ensemble classification.

3.2 ENSEMBLE CLASSIFICATION

Ensemble classification is achieved by clubbing different classifiers in order to achieve greater classification accuracy. This kind of classification is efficient, as the classification decision depends on multiple classifiers, rather than a single classifier. Ensemble classification is reliable and effective. This work utilizes k-NN, SVM and ELM classifiers to make the final classification decision.

The reason for selecting three classifiers is that the time conservation is improved, as the count of classifiers increases. Hence, this work employs three reliable classifiers and the reasons for the choice of classifiers are presented above. All the classifiers gain knowledge from the input database during the process of training and the classifiers distinguish between the microarray data during the testing phase. The following sub-sections present the working principles of the three different classifiers. The classification is diagrammatically represented in the following figure.

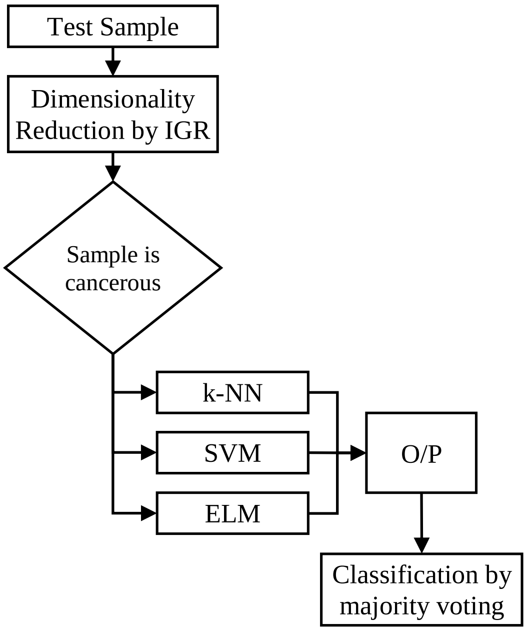
Test Sample
Dimensionality Reduction by IGR
Sample is cancerous
k-NN
SVM
ELM
O/P
Classification by majority voting

Fig.1. Overall flow of ensemble classification

3.2.1 k-NN Classifier:

The k-NN classifier is one of the basic classifiers, which distinguish between the normal and abnormal gene data. The k-NN classifier distinguishes between the normal and the abnormal data by computing Euclidean distance between the data, which is computed by,

$$E_{Dis} = \sum_{i=1}^{N} \sqrt{x_i^2 - y_i^2} \tag{4}$$

The efficiency of this classifier depends on the choice of the value '*k*'. The value of *k* decides the classification accuracy of the classifier and choosing the right value of *k* is a challenging problem. Additionally, the value of *k* differs for every dataset and prior knowledge about the dataset is necessary for fixing the value of *k*. This makes it ineffective and this work utilizes *k* fold cross validation, which automatically chooses the *k* value. In the process of *k* value fixation is carried out by decomposing the training data into multiple *k* samples, among which a single sample is treated as the test sample and the remaining samples are considered as the training samples. This operation is repeated for *k* times, till all the samples are treated as testing sample atleast once. When this operation is over, the mean value of the computed *k* results and the mean value is fixed as *k*. This technique is optimal and the need to choose the value of *k* by the users is eliminated. Additionally, this technique does not require any prior knowledge about the dataset and is feasible for any kind of dataset.

3.2.2 SVM Classifier:

SVM is a promising classifier and is trained with a set of training samples. The SVM classifies between the microarray gene data by means of a separating margin. Let the group of training samples are to be classified as either normal or abnormal. The samples are included in the corresponding class by means of a hyperplane. This hyperplane is necessary to separate the samples belonging to the normal and abnormal classes. The differentiation between the classes can be represented as follows.

$$\psi.j_i + b \geq +\text{ve for } cl_i = \text{Positive} \tag{5}$$

$$\psi.j_i + b \leq -\text{ve for } cl_i = \text{Negative} \tag{6}$$

The distance between the hyperplane decides the accuracy of the classification results.

$$\text{Distance between the hyperplanes}=\frac{2}{\|\psi\|}. \tag{7}$$

The lesser the value of $\|\psi\|$, the better is the classification results. Hence, the distance between the hyperplane is optimized by means of Lagrange's function.

$$lf(x) = \sum_{i=1}^{Q} \alpha_i \psi_i (j_i \cdot j) + th \tag{8}$$

In Eq.(8), α_i is the lagrange multiplier which partitions the hyperplane $\psi_i(j_i{\cdot}j)$ and *th* is the threshold to partition the hyperplane. Hence, if the value of $lf(x)$ is greater than 0, then the sample is abnormal, otherwise normal.

3.2.3 ELM Classifier:

The striking point about this classifier is its faster learning ability [20]. Consider the microarray gene data is represented by (a_i,b_i), where $a_i = [a_{i1},a_{i2},\ldots,a_{in}]^T \in D^n$ and a_{in} is the i^{th} training sample in *n* dimension. The i^{th} label in c^{th} dimension of the training dataset is represented by $b_i = [b_{i1},b_{i2},\ldots,b_{ic}]^T \in D^c$ and *c* is the total count of classes in the process of classification. As far as this work is taken into consideration, the total count of class is two. The Single hidden Layer Feed Forward Neural Network (SLFN) is formed by

$$\sum_{j=1}^{N} \gamma_i q\left(w_j \cdot a_i + y_j\right) = y_i ; i = 1,2,\ldots,n \tag{9}$$

In Eq.(9), w_j is the weight of the samples that are represented by $[w_{j1},w_{j2},..w_{jn}]^T$. The w_j is responsible for interconnecting the j^{th} neuron with the input neurons and the *j* is presented by $j = [j_1,j_2,\ldots,j_c]^T$. Additionally, the w_j links the hidden neuron with the output neurons. The bias of the j^{th} hidden neuron is given by y_j. Let HL be the hidden layer output matrix of the classifier, in which the j^{th} column of HL signifies the j^{th} hidden neurons output vector by taking the input a_{i1}, a_{i2},…, a_{in} into account.

$$HL = \begin{bmatrix} a_{fn}\left(w_1 \cdot a_1 + y_j\right) & \cdots & a_{fn}\left(w_N \cdot a_1 + y_N\right) \\ \vdots & \ddots & \vdots \\ a_{fn}\left(w_1 \cdot a_n + y_j\right) & \cdots & a_{fn}\left(w_N \cdot a_n + y_N\right) \end{bmatrix} \tag{10}$$

$$\gamma = \begin{bmatrix} \gamma_1^T \\ \vdots \\ \gamma_N^T \end{bmatrix} \tag{11}$$

$$R = \begin{bmatrix} r_1^T \\ \vdots \\ r_N^T \end{bmatrix} \tag{12}$$

The matrix format of the SLFN is represented as follows.

$$HL\gamma = R \tag{13}$$

The output weights are computed by normalized least-square solution as follows

$$\gamma = HL^{\dagger}R \tag{14}$$

where $HL^{\dagger}$ is the HL's Moore-Penrose generalized inverse. When the training process is initialized, the total count of classes, hidden neurons and activation function a_{fn} are fed into the classifier. The training samples are given by $\{a_i,b_i\}$ and the classifier is made to get trained by γ, as given in Eq.(14). By this way, the ELM classifies between the normal and abnormal data.

As soon as the decisions of all the classifiers are obtained, the decision with maximum count is computed. For instance, any two of the classifiers may classify the sample as normal and one classifier may arrive at the decision as abnormal. In this case, the sample is declared as normal as per the maximum voting strategy. As this work employs three different classifiers, there can be nine different combinations of solutions and the final decision is made accordingly. Each classifier can end up the decision with normal and abnormal. The abnormal class is denoted by 1 and the normal class is denoted by 0. Each column of the decision matrix (D_{mat}) denotes the classifier. The first column of the D_{mat} indicates the k-NN classifier, the second and third column of the matrix denotes the SVM and ELM classifiers. Based on the D_{mat}, the final decision is computed by taking the repeatedly occurring decision.

$$D_{mat} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix} \Rightarrow \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 0 \\ 1 \\ 1 \end{bmatrix} \tag{15}$$

Thus, the final decision is made by taking the D_{mat} into consideration. This way of classification is not prone to misclassification and so the classification accuracy of this technique is greater. The following section analyses the performance of the proposed approach in terms of standard performance measures.

4. RESULTS AND DISCUSSION

The performance of the proposed approach is analysed in terms of classification accuracy, sensitivity and specificity. The datasets being utilized for evaluating the performance of the proposed approach are leukemia [21], colon [22] and breast cancer [23] datasets. The leukemia dataset consists of 3571 genes, which are derived from 72 individuals. Colon dataset contains 2000 genes collected from 62 instances. In this dataset, forty samples are normal and the remaining twenty two samples are abnormal. The breast cancer dataset contains 24481 genes with 78 samples. Out of the 78 samples, 34 samples are considered as abnormal and the remaining 44 are considered as normal.

The proposed algorithm is applied in Matlab environment with version 8.1. All the experiments are carried out in a standalone system with 16GB RAM and 7th generation Intel core processor with 4MB cache, 3.5GHz. This work divides the dataset into ten parts and each part act as a testing part, while the remaining nine parts are the training parts. Hence, all the parts of the dataset are tested. The performance of the proposed approach is tested with respect to classification accuracy, sensitivity, specificity and misclassification rate.

The accuracy rate of the classification approach is very important, as the correct classification between normal and cancerous cases is necessary. As this work is related to human disease diagnosis, the accuracy rates are given more importance. The accuracy rate is computed by taking the True Positive (*TP*), True Negative (*TN*), False Positive (*FP*) and False Negative (*FN*) rates. The formula for computing accuracy rate is presented as follows.

$$Ac = \frac{TP+TN}{TP+TN+FP+FN} \times 100 \tag{16}$$

The sensitivity and specificity measures are other important performance measures of a classification algorithm. The formulae for computing sensitivity and specificity are presented in the following equations.

$$Sen = \frac{TP}{TP+FN} \times 100 \tag{17}$$

$$Spec = \frac{TN}{FP+TN} \times 100 \tag{18}$$

$$MC = 100 - Ac \tag{19}$$

The sensitivity and specificity rates of the classification algorithm must be greater. It is important to achieve greater sensitivity and specificity rates rather than to achieve better accuracy rates. The accuracy rates consider all the *TP*, *TN*, *FP* and *FN* rates but the sensitivity and specificity rates consider *FN* and *FP* rates respectively. The impact of wrong classification is serious when it comes to diagnosis of disease. The sensitivity rates of the algorithm increase, when the false negative rates decrease. False negative rates are dangerous because the abnormal sample is classified as normal and this is serious because the next process of the treatment may be delayed.

The specificity rates of the work can be improved, provided the *FP* rates are reduced. In this scenario, the abnormal samples are classified as normal, which may hold back the treatment procedure of the affected case. Hence, it is necessary to minimize *FP* and *FN* rates, as much as possible. The greater the sensitivity and specificity rate, the more reliable is the classification system. The performance of this work is evaluated in two aspects. The first aspect aims to prove the efficiency of ensemble classification over the application of individual classifiers. The second aspect of performance analysis compares the performance of the proposed approach with the related state-of-the-art techniques.

4.1 PERFORMANCE ANALYSIS BY VARYING THE CLASSIFIERS

The aim of this section is to justify the choice of ensemble classifier over several individual classifier such as k-NN, SVM and ELM. However, it is not good to determine that the performance of ensemble classifier is the best, on applying it over a single dataset. Considering this, the performance of ensemble classifier is tested over three different datasets and all the analysis prove that ensemble classifier shows better performance than the individual classifiers. As shown below, the ensemble classifier shows greater accuracy, sensitivity and specificity rates. The ensemble classifier proves better results for all the datasets, as the classification decision does not based on a single classifier. The performance of the proposed approach is evaluated for all three datasets by varying the feature selection and classification techniques. The experimental results of the proposed ensemble classification are tabulated in Table.1.

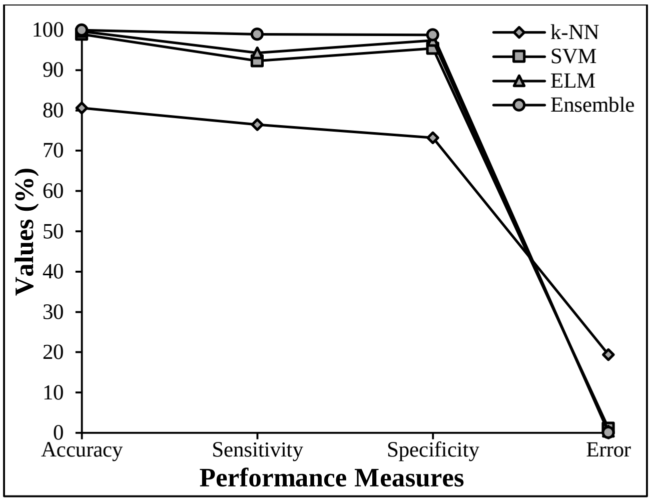

Fig.2. Performance analysis on ALL-AML

Table.1. Performance analysis by varying classification techniques

Dataset – ALL-AML				
Performance Metrics / Classifiers	**k-NN**	**SVM**	**ELM**	**Ensemble**
Accuracy (%)	80.6	98.9	99.4	99.9
Sensitivity (%)	76.5	92.3	94.3	98.9
Specificity (%)	73.2	95.4	97.4	98.7
Error Rate (%)	19.4	1.1	0.6	0.1
Dataset – Colon Tumor				
Performance Metrics / Classifiers	**k-NN**	**SVM**	**ELM**	**Ensemble**
Accuracy (%)	73.6	78.6	98.5	99.4
Sensitivity (%)	60.25	81.7	99.7	99.8
Specificity (%)	65.6	75.4	96.4	98.9

Error Rate (%)	26.4	21.4	1.5	0.6
Dataset – Breast Cancer				
Performance Metrics / Classifiers	**k-NN**	**SVM**	**ELM**	**Ensemble**
Accuracy (%)	66.73	76.8	92.17	98.3
Sensitivity (%)	73.2	74.6	94.2	97.8
Specificity (%)	72.9	79.6	93.28	98.7
Error Rate (%)	33.27	23.2	7.83	1.7

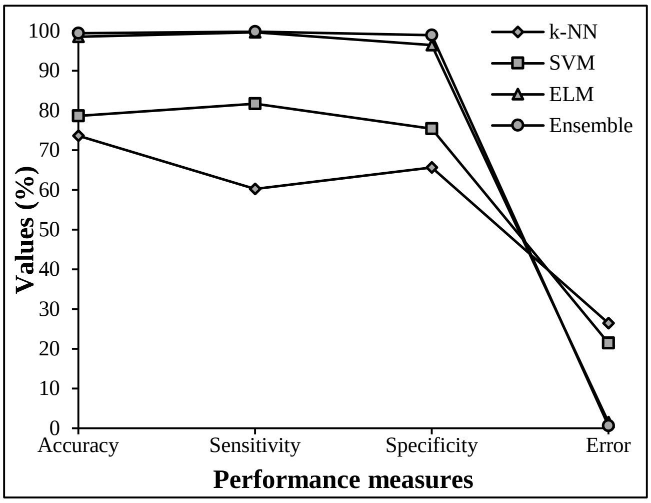

Fig.3. Performance analysis on Colon tumor dataset

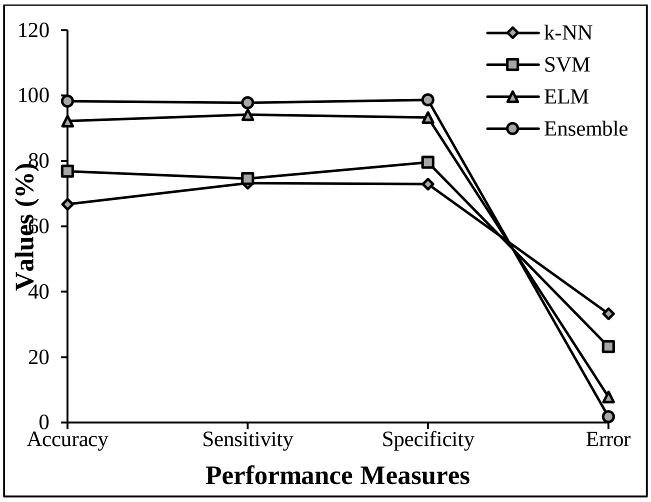

Fig.4. Performance analysis on Breast cancer dataset

The above presented performance analysis proves the efficiency of ensemble classifier in the place of employed individual classifier. The ensemble classifier proves better results with greater sensitivity and specificity rates. The maximum accuracy rate being proven by the ensemble classifier is 99.9% for the Leukemia dataset.

The accuracy rates of the colon tumor and breast cancer of the ensemble classification technique are 99.4% and 98.3% respectively. The average sensitivity rate of the ensemble classification technique is 98.83%. On the other hand, the average sensitivity rate of the SVM and ELM are 82.86% and 96.06% respectively. The specificity rate of the proposed approach is 98.76%. The experimental results prove the efficacy of the proposed approach.

4.2 PERFORMANCE ANALYSIS

This section aims to compare the performance of the proposed approach and compares it with the recent exiting approaches such as dissimilarity measure based microarray gene data classification [16], heuristic algorithm based microarray gene data classification [10]. The dissimilarity based microarray gene data classification takes the top ranking gene into account without concerning the features of the gene data. The heuristic based microarray data classification focuses on classification accuracy, rather than the process of gene selection. The experimental results are presented in Table.2.

Table.2. Comparative analysis with the existing approaches

Datasets/Performance metrics	**Colon Tumor**		
	Dissimilarity based	**Heuristics based**	**Proposed**
Accuracy	82.3	89.6	**99.01**
Sensitivity	73.6	81.9	**99.4**
Specificity	77.3	78.96	**97.2**
Misclassification rate	17.7	10.4	**0.99**
Datasets/Performance metrics	**Breast cancer**		
	Dissimilarity based	**Heuristics based**	**Proposed**
Accuracy	82.3	89.6	**99.01**
Sensitivity	73.6	81.9	**99.4**
Specificity	77.3	78.96	**97.2**
Misclassification rate	17.7	10.4	**0.99**
Datasets/Performance metrics	**ALL-AML**		
	Dissimilarity based	**Heuristics based**	**Proposed**
Accuracy	82.3	89.6	**99.01**
Sensitivity	73.6	81.9	**99.4**
Specificity	77.3	78.96	**97.2**
Misclassification rate	17.7	10.4	**0.99**

This work has proven the performance of the proposed approach by comparing with the recent existing techniques. The proposed approach is tested up on three different datasets however, the performance of the proposed approach is stable and promising. The major reason for the better results being shown by the proposed approach is the feature dimensionality reduction and ensemble classification. The feature dimensionality reduction aims in selecting the most prominent features rather than the entire feature set. This helps in achieving better classification results. In addition to this, ensemble classifier takes the final decision by considering the decisions of all the individual classifiers. The maximal occurring decision is declared as the final decision, which improves the classification accuracy.

5. CONCLUSIONS

This paper presents a retrieval system for microarray gene data, which relies on information gain ratio and ensemble classification. As the dimensionality of the gene data is greater, this work intends to select the optimal features by means of IGR and then the classification phase is initiated. Initially the ensemble classifier is trained with the dataset, which makes it able to classify between the normal and the abnormal samples. The efficiency of the proposed approach is evaluated against different datasets and the power of ensemble classification is justified. In future, this work plans to introduce multiclass classification system for microarray gene data.

REFERENCES

[1] D. Coomans and D.L. Massart, "Alternative k-Nearest Neighbour Rules in Supervised Pattern Recognition: Part 1. k-Nearest Neighbour Classification by using Alternative Voting Rules", Analytica Chimica Acta, Vol. 136, pp. 15-27, 1982.

[2] C. Cortes and V. Vapnik, "Support-Vector Networks", *Machine Learning*, Vol. 20, No. 3, pp. 273-297, 1995.

[3] G.B. Huang, Q.Y. Zhu and C.K. Siew, "Extreme Learning Machine: Theory and Applications", *Neurocomputing*, Vol. 70, No. 1, pp. 489-501, 2006.

[4] S. Bandyopadhyay, A. Mukhopadhyay and U. Maulik, "An Improved Algorithm for Clustering Gene Expression Data", Bioinformatics, Vol. 23, No. 21, pp. 2859-2865, 2007.

[5] U. Maulik, A. Mukhopadhyay and S. Bandyopadhyay, "Combining Pareto-Optimal Clusters using Supervised Learning for Identifying Coexpressed Genes", *BMC Bioinformatics*, Vol. 10, No. 1, pp. 20-27, 2009.

[6] A. Mukhopadhyay, S. Bandyopadhyay and U. Maulik, "Multi-Class Clustering of Cancer Subtypes through SVM based Ensemble of Pareto-Optimal Solutions for Gene Marker Identification", *PLoS ONE*, Vol. 5, No. 11, pp. 1-8, 2010.

[7] U. Maulik and A. Mukhopadhyay, "Simulated Annealing based Automatic Fuzzy Clustering Combined with ANN Classification for Analysing Microarray Data", *Computers and Operations Research*, Vol. 37, No. 8, pp. 1369-1380, 2010.

[8] A. Mukhopadhyay and U. Maulik, "Towards Improving Fuzzy Clustering using Support Vector Machine: Application to Gene Expression Data", *Pattern Recognition*, Vol. 42, No. 11, pp. 2744-2763, 2009.

[9] U. Maulik, "Analysis of Gene Microarray Data in a Soft Computing Framework", *Applied Soft Computing*, Vol. 11, No. 6, pp. 4152-4160, 2011.

[10] Jia Lv, Qinke Peng, Xiao Chen and Zhi Sun, "A Multi-Objective Heuristic Algorithm for Gene Expression Microarray Data Classification", *Expert Systems with Applications*, Vol. 59, pp. 13-19, 2016.

[11] Shun Guo, Donghui Guo, Lifei Chen and Qingshan Jiang, "A Centroid-based Gene Selection Method for Microarray Data Classification", *Journal of Theoretical Biology*, Vol. 400, pp. 32-41, 2016.

[12] Hanaa Salem, Gamal Attiya and Nawal El-Fishawy, "Classification of Human Cancer Diseases by Gene Expression Profiles", *Applied Soft Computing*, Vol. 50, pp. 124-134, 2017.

[13] Sina Tabakhi, Ali Najafi, Reza Ranjbar and Parham Moradi, "Gene Selection for Microarray Data Classification using a Novel Ant Colony Optimization", *Neurocomputing*, Vol. 168, pp. 1024-1036, 2015.

[14] Ehsan Lotfi and Azita Keshavarz, "Gene Expression Microarray Classification using PCA-BEL", *Computers in Biology and Medicine*, Vol. 54, pp. 180-187, 2014.

[15] Nur Shazila Mohamed, Suhaila Zainudin and Zulaiha Ali Othman, "Metaheuristic Approach for an Enhanced MRMR Filter Method for classification using Drug Response Microarray Data", *Expert Systems with Applications*, Vol. 90, pp. 224-231, 2017.

[16] Vicente Garcia and J. Salvador Sanchez, "Mapping Microarray Gene Expression Data into Dissimilarity Spaces for Tumor Classification", *Information Sciences*, Vol. 294, pp. 362-375, 2015.

[17] Aiguo Wang, Ning An, Guilin Chen, Lian Li and Gil Alterovitz, "Improving PLS–RFE based Gene Selection for Microarray Data Classification", *Computers in Biology and Medicine*, Vol. 62, pp. 14-24, 2015.

[18] Huijuan Lu, Junying Chen, Ke Yan, Qun Jin and Yu Xue, Zhigang Gao, "A Hybrid Feature Selection Algorithm for Gene Expression Data Classification", *Neurocomputing*, Vol. 256, pp. 56-62, 2017.

[19] M. Dashtban and Mohammadali Balafar, "Gene Selection for Microarray Cancer Classification using a New Evolutionary Method Employing Artificial Intelligence Concepts", *Genomics*, Vol. 109, No. 2, pp. 91-107, 2017.

[20] Guang-Bin Huang, Hongming Zhou, Xiaojian Ding and Rui Zhang, "Extreme Learning Machine for Regression and Multiclass Classification', *IEEE Transactions on systems, Man and Cybernetics-Part B*, Vol. 42, No. 2, pp. 513-529, 2012.

[21] PMC-NCBI-NIH, Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC151171

[22] Gene Expression Project, Available at:http://microarray.princeton.edu/oncology

[23] Lt. Thomas Scaria and T. Christopher, "Microarray Gene Retrieval System based on LFDA and SVM", *International Journal of Intelligent Systems and Applications*, Vol. 1, pp. 9-15, 2018.

[24] Lt. Thomas Scaria and T. Christopher, "Supervised Microarray Gene Retrieval System based on KLFDA and ELM", *International Journal of Advanced Intelligent Paradigms*, 2018.