Journal of Critical Reviews

ISSN- 2394-5125 Vol 7, Issue 14, 2020

AN OPTIMAL CLASSIFICATION MODEL FOR MICROARRAY CANCER DISEASE PREDICTION

[1]P.Ramya, [2]Dr.T.Bhaskar Reddy

Research scholar, Dept. of Computer Science & Technology, Sri Krishnadevaraya University Ananthapuram,India
Professor, Dept. of Computer Science & Technology, Sri Krishnadevaraya University Ananthapuram,India

Received: 06.04.2020 ***Revised: 08.05.2020*** ***Accepted: 03.06.2020***

Abstract
As the size of the biomedical databases areincreasing day-by-day, finding anessential featureset for classification problem is complex due to large data size and sparsity problems. Microarray feature ranking and classification is one of the major challenges to scientific and medical researchers due to its high dimensional feature space and limited number of samples. Feature transformation, feature ranking and data classification are the essential components to improve the microarray cancer prediction on high dimensional datasets. In this work, a novel framework is designed and implemented to classify the high dimensional data with high true positive rate. In the proposed work, a hybrid feature transformation, hybrid feature selection and advance classification approach are implemented to improve the true positive rate and error rate of the disease prediction. A novel principal component ranking measure is integratedin order to find the subset of features for classification problem. Finally, a hybrid decision tree classifier is used to predict the classification accuracy on the selected features set. Experimental results proved that the present framework has better performance compared to the traditional models for variable microarray datasets.

Keyword: Microarray data, feature selection, data transformation and classification.

DOI: http://dx.doi.org/10.31838/jcr.07.14.02

INTRODUCTION

Micro array data generally contains a set of genes and its disease associations. Most of the traditional approaches are inappropriate and computationally infeasible to find patterns on high dimensional datasets. Hence, it is difficult to process all of the genes that are not required during the process of classification. Also, the overall computational overhead also increases significantly. Unwanted noise is resulted during the process of classification. Hence, it is very much required to select few numbers of genes those usually take part during the classification process. All of the traditional gene selection techniques involve a perfect combination of filter and wrapper schemes[1]. Filtering approaches have the responsibility to rank every individual feature according to their goodness. During the process of ranking, the relationship among every individual gene with respective class label is considered. Univariate scoring metric play a significant role in the above ranking process. The top ranked genes are selected prior to the execution of classification schemes. On the contrary, wrapper schemes require the gene selection approach in order to integrate with a classifier. The prime objective of this technique is to evaluate the classification performance of every individual gene subset. The optimal subset of genes is detected according to the ranking of performance. Traditional filtering schemes are incapable and inefficient to measure the relationship in between different genes. The gene expression data play significant role during the process of biomedical diagnosis. According to the latest research concepts, limited numbers of genes may result high prediction accuracy during the diagnosis process of cancer disease[2]. Large numbers of genes are not relevant to the disease of interest. Therefore, the gene selection procedure is very complicated task during the medical data processing. Feature selection is considered as the most powerful tool in order to decrease the size of available data.

Extreme classification model is an extension of traditional neural network model for data classification. It partitions the whole problem into numbers of sub-problems and merges them to find an optimal solution. The parameters of hidden layer contain training data samples are mapping to output layer. In the traditional SLFN approach, the adjustments of parameters are iterative in nature and results some issues. These issues are overcome by the suggested extreme classifier approach [3]. Hence, ELM approach achieves better generalization performance as well as good learning speed. Most of the traditional learning models for training SLFNs are comparatively slower than that of non-parametric approaches. This approach operates slowly because parameters are required to be tuned iteratively. Moreover, these models require high computational memory and also increase the overall computation time of the mapping process. An extended and slightly modified version of traditional SLFNs approach is developed as extreme classifier [4]. This method is used to enhance the efficiency and performance of conventional SLFNs. Also, most of the Neural network-based learning schemes perform manual tuning of control parameters (such as learning rate, learning epochs etc.) as well as local minima. But, extreme classifier is applied automatically and there is no need of manual iterative tuning. The classification boundary is not optimal in ELM and the boundary is constant throughout the training phase. Hence, there are chances of misclassification of samples closer to boundary. This approach requires a large number of hidden neurons as compared to other traditional tuning-based approaches .

Machine learning allows a classifier to learn a set of rules, or the decision criterion, from a set of labelled data that has been annotated by an expert. This approach allows for better scaling and a reduced cost in classifying medical data when compared to a system that relies on manual input only. Most of the research in the field of machine learning based medical data classification has been done on binary classifiers. That is, building a classifier from a set of positive and negative examples to then deduce a medical data's membership in a class. Classification can take many forms, from fully automated systems with limited human intervention [4] to semi-automatic systems that employ a hybrid human machine approach.

Most of the traditional ensemble classification models are processed with limited feature space and small data size. As the size of the feature space increases, traditional ensemble classifiers select a predefined number of features for classification. Learning classification models with all the high dimensional features may result serious issues such as performance and scalability. Feature selection measures can be categorized into three types: wrappers, filters and embedded models. Learning classification models with all the high dimensional features may result serious issues such as performance and scalability. The main problems in the existing models are:

1. Problem of feature selection on high dimensional datasets.
2. Problem of predicting disease with high mis-classification rate.
3. Problem of handling high dimensional and large datasets using the parallel processing model.

RELATED WORKS

A multi-scale filter bank is used in order to present the characteristics of local data texture and structure. Different efficient and effective classification schemes are implemented to train the system. In the subsequent time, another generalize system is developed which has the responsibility of regional lung classification.Feature subset selection can improve the classification accuracy by creating optimal subset features from the high dimensional feature space. A forward feature selection approach has been used for the selection of limited features from the original feature space[5]. Adaboost(Adaptive Boosting) is a meta learning based approach from the ensemble learning group. The main objective of the AdaBoost is to improve the strong classifier using the group of base weak classifiers. AdaoBoost approach is an iterative method, and in each iteration, a weak base classifier is selected to minimize the error rate of the model. High dimensionality poses a severe issue for machine learning models In order to optimize the precision of the classification algorithm, most of the classification approaches use feature selection measures such as mutual information, correlation coefficient , rough-set , chi-square test etc. to select subset of features from the high dimensional space. They implemented a PCA based spectral filtering model to high dimensional features of the original training data. Two reconstruction methods are used, one is the principle component analysis and the other is Maximum likelihood estimation. Several distribution algorithms were used in randomization models. In most of these approaches Bayesian analysis is used to predict the original data distribution using the randomization operator and the randomization data.

Most of the traditional approaches detect inappropriate and computationally infeasible patterns on high dimensional datasets. Hence, it is difficult to process all of the cancer patterns that are not required during the process of classification. Hence, the overall computational overhead also increases significantly. Unwanted noise is resulted during the process of classification. Hence, it is very much required to select essential cancer patches during the classification process. All of the traditional cancer selection techniques involve a perfect combination of filter and wrapper schemes. Filtering approaches have the responsibility to rank every individual feature according to their goodness. During the process of ranking, the relationship among every individual cancer with respective class label is considered. Univariate scoring metric play a significant role in the above ranking process. The top ranked cancer patches are selected prior to the execution of classification schemes. On the contrary, wrapper schemes require the cancer selection approach in order to integrate with a classifier. The prime objective of this technique is to evaluate the classification performance of every individual cancer subset. The optimal subset of disease patterns is detected according to the ranking of each feature. Traditional filtering schemes are incapable and inefficient to measure the relationship in between different genes. The cancer patches data play significant role during the process of biomedical diagnosis. A medical instance includes a large number of features or characteristics. Large numbers of cancer features are not relevant to the disease of interest. Therefore, the cancer selection procedure is very complicated task during the medical data processing. Feature selection is considered as the most powerful tool in order to decrease the size of available data. In the cancer classification process, feature selection models are generally divided into two types, wrapper approaches and filter approaches. Wrapper approach evaluate search feature or feature subset to optimize the classification accuracy. Filter method evaluates each feature independent from the classification algorithm, ranks the cancer features after evaluation and considers the superior one. This evaluation is performed using information, dependency, distance and consistency. In general, the speed of wrapper model is lower than the filter model because of cross validation and repeated iteration to evaluate the feature subsets. Traditional wrapper model is more efficient because classification technique affects the overall accuracy, although the subset selection is an NP-hard. However, if the number of features involved in complex data increases, finding new disease patterns can become difficult due to the complex relationships among features. Feature ranking methods compute the measure for each feature and rank them accordingly. These ranking methods select the top'k' features based on highest rank and eliminate those having lower feature ranks [6].Information gain is one of the attribute selection measures which are based on entropy value.

Good sub-sets of attributes contain class-related and non-related attributes. The method is used to determine how closely the characteristic vectors are linked. When the coefficient of correlation between the two vectors is above "0," the characteristics are said to have a highly positive correlation. Likewise, the functions are said to be negatively correlated if the correlation coefficient between these two characteristic vectors is less than "0" The features are said not to be correlated if the correspondence coefficient of the two characteristic vectors is equal to "0"[4]. Chi-Square is based on the analysis of statistics. The functional vector measures the independence. The strength of the relationship between two random variables is tested with observed and anticipated values. Feature descriptors are the pixels behind so that the classification performance is optimized and a sparse representation is given. The descriptors should ideally be invariant to operations like scale, rotation and illumination changes. This invariance enables descriptors to be matched across videos which have differences in these parameters.

Each medical data is scanned and transformed into normalized continuous data. The main issues of the medical datasets are high dimensionality and imbalance nature. Traditional machine learning classifiers consider subset of features for classification and disease prediction with high true negative rate and error rates. Attribute selection is used to compute the measure for each feature and rank them accordingly. These ranking methods select the top 'k' features based on highest rank and eliminate those having lower feature ranks. Information gain is one of the attribute selection measures which is based on entropy value. Information gain approach is the mutual information of a target random variable say *P* and independent random variable *Q*. The main limitation of this approach is, it chooses features having large distinct values over the features having less distinct values. Table summarizes the different attribute selection measures used in decision tree construction for pattern construction. Table 0, describes the comparison of different machine learning approaches for medical imbalance detection on different parameters[7].

Table 0 summarizes the various attribute selection measures used in decision tree construction for medical intrusion detection.

Measure Name	Formula
Entropy	$\text{Entropy} = \sum p_i \log_{10}(p_i)$
Information Gain	$\text{Gain} = \sum p_i \log_{10}(p_i) - \sum \frac{\lvert D_v \rvert}{\lvert D \rvert} \sum p_i \log_{10}(p_i)$ where D is dataset, D_v is subset of dataset, p_i is the probability of ith class
Gain Ratio	$\text{GainRatio} = \text{Gain} / (-\sum \frac{\lvert D_v \rvert}{\lvert D \rvert} \log \frac{\lvert D_v \rvert}{\lvert D \rvert})$ where D is dataset, D_v is subset of dataset, p_i is the probability of ith class
Pearson Correlation	$\text{CorrelationFS} = \frac{m.\bar{r}_{ic}}{\sqrt{m + m(m-1)\bar{r}_{ii}}}$
Improved Correlation measure	$\arg\max \frac{\lvert \text{cov}(\text{Vec}(A_i, c), \text{Vec}(D)) \rvert}{\sqrt{\text{var}(\text{Vec}(A_i, c) * \text{Vec}(D))}}$

Lu andYan et.al, proposed an advanced feature reduce based intrusion detection system [8][9]. In this model, a feature ranking measures such as information gain and correlation are used to filter the feature space. After successful completion of the feature ranking, feature reduction approach is implemented. The feature reduction technique is implemented through the integration of ranks generated from the process of information gain and correlation. The reduced features are given as input to feed forward neural medical to train and test the medical features in cancer dataset. In this method, the pre-processing is carried out manually which is a severe drawback of this model. Liu, et.al, proposed knowledge-maximized ensemble approach for various kinds of concept drift [10]. In this work, they presented an advanced data stream classifier which is known as knowledge maximized ensemble. Hence, it becomes hard and complicated to restrict the amount of training data. This technique can be influenced by different kinds of concept drift through integration of various imbalance detection approaches. Decision tree induction is a simple and powerful classification process that produces a tree and a set from a specific dataset[11] representing a model for different classes. The choice tree is a tree structure like a flow chart, in which each inner knot is a test on a single attribute, each branch is a test result and every node in the leaf is a class. The root node is the highest node of a tree. ID3 algorithm and Quinlan's successor C4.5 algorithm are commonly used for DT induction. In classifying unknown data records, the structure is widely applied. At every inner node of the tree, impurity measures are applied to the best split decision. The tree leaves consist of the class labels of the data items. Two phases are implemented in the decision tree classification technique: tree building and tree pruning. They introduced CART (Classification and Regression Trees). Different statistical measures are used to select the attribute from the list of feature space. For building a decision tree, CART utilizes both numerical and categorical attributes and includes features dealing with missing attributes[12]. In determining the best partition and data is sorted at all the nodes to determine the best fractional point, it uses a large number of single fractional criteria, such as a gini index, gini ratio etc, and one multi-variable (linear combination.

Class-imbalanced data are common in the domain of data categorization. It generally categorizes many irrelevant documents, but some articles are categorized under interesting category. BN approaches are mostly implemented as standard classifiers. These approaches give rise to exact results along with the capability for representing relationships in between variables. This approach is unable to resolve the traditional class-imbalanced problem [5].In order to overcome the class-imbalanced problem, many advanced approaches are developed to gain enhanced accuracy rate of standard classification approaches. These generalized classification approaches involve the following schemes:- sampling approaches, cost-sensitive approaches, recognition-based methods and active learning approaches. Sampling approaches have been developed in order to resolve the issues due to class imbalance through removing certain data from majority class. This approach is also known as under sampling approach. Sometimes few additional artificially produced data are included into the minority class and this phenomenon is also called as over-sampling. Cost sensitive learning method generally involves cost-matrix for all categories of errors or instances. It has main objective of facilitating learning mechanism out of imbalanced data sets. The above mechanism has equivalent influence on the process of oversampling the minority class. It can end with over specific rules or rules over-fitting training [16]. Random under sampling approach involves the process of randomly removing elements of the over-sized class. The above process continues executing till it matches the size of other class and cost-sensitive learning scheme. It includes the modifications of relative cost associated with misclassification of positive and negative class. The outcomes of both methods are analyzed and compared with performance achieved without balancing.

PROPOSED MODEL

The overall architecture of the proposed model is represented in fig 1. Initially, each microarray gene disease dataset is filtered tofill the sparsity problem and missing values of the gene featured. Here, a hybrid data transformation approach is used to transform the feature values using the gaussian transformation measure. Each value is normalized to improve the balancing property of each feature and its class. In the initial phase, each feature is transformed using the gaussian transformation process. In the second phase, essential features are extracted using the hybrid PCA approach. Finally, an optimized decision tree classifier is designed to find the essential cancer patterns for prediction process.

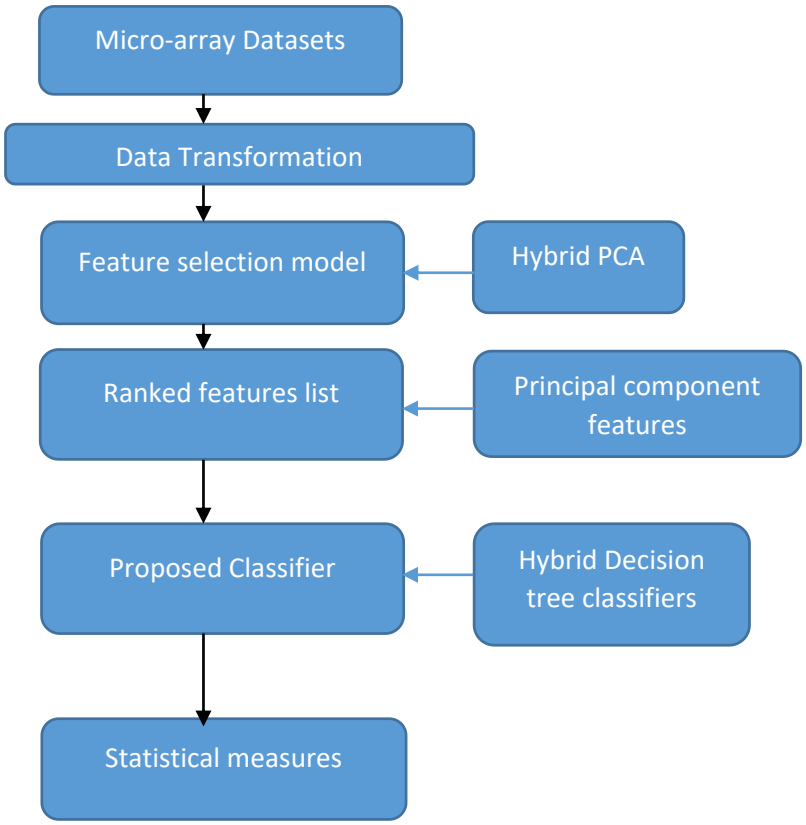

Figure 1: Proposed Gaussian filter based feature selection and classification model

Each microarray training dataset is pre-processed using the data transformation function to remove the variation among the data distribution. In the proposed work, a gaussian based data transformation function is used to normalize the input training data for clustering and wrapper feature ranking process in the mapper phase.

Gaussian based Data Pre-processing
Input : Training microarray dataset D, F(D): Feature space of D.
Output: Gaussian Filtering or Transformed data KD.
Procedure:

1. Read input data D.
2. For each feature F[i]in feature space F(D)
3. Do
4. Apply Gaussian transformation on each feature as
5. $$\text{GeneKernelTransform}(F[i]) = \phi = (\sum (F[i]) / \max\{F[i]\}).\frac{1}{\sqrt{2\Pi}} e^{-(F[i]-\mu(F[i]))/\sigma(F[i])}$$
6. If(ϕ >0)
7. Then
8. Normalize F[i] using Min-max normalization [0, ϕ]
9. Else
10. Normalize F[i] using Min-max normalization with [R1,R2]

$$x' = \frac{x - \min\_(x)}{\max\_(x) - \min\_(x)} * (R2 - R1) + R1$$

11. End if
12. Done

Here,gaussian normalization and min-max range normalization are used to improve the data distribution with uncertain values . R1 and R2 represents the minimum and maximum values of each feature. This approach is used to remove the sparsity problem and data normalization problem in high dimensional datasets.

Algorithm: HybridPCA (HPCA):
In the proposed feature ranking model, traditional PCA algorithm is enhanced to find the most essential features in the given feature space.

Input: Normalized training data.

Output: Ranked features.

Step 1: Input Normalized data ND.

Step 2: Mean of the normalized data is computed using eq.(1)

$$\mu_D = \sum ND[i] / N \quad \text{-----(1)}$$

Step 3: Let F={f[0],f[1]....f[m]} be the feature space with m features.

Find the candidate features pairs

CF={(f[0],f[1]),(f[0],f[2]),(f[0],f[3])……(f[m],f[0])....}.

For m feature space we will get $\frac{m!}{(m-2)!2!}$ candidate sets.

For each pair of candidate features CF

Do

Compute covariance between features as

$$\text{Cov(CF\{x,y\})} = \frac{\sum_{i=1}^{n}(CF[x_i]-\mu_{CF[x_i]})(CF[y_i]-\mu_{CF[y_i]})}{(n-1)} \quad \textbf{---(2)}$$

Done

Step 4: Compute the eigen vector and values using the eq.(3) and eq.(4)

$$\text{EigenValues[]} = \text{Det}(\lambda I - COV(CF)) = 0 \quad \text{---- (3)}$$

Here I is the identity matrix of same dimension as COV(CF). The corresponding Eigen vector is given as

$$(\lambda I - COV(CF))v=0 \quad \text{------(4)}$$

Here the optimal eigen sum is computed as

$$\text{OptimalEigenSum} = \frac{\sum EV[i])}{\text{MaxProb}\{\exp(F(EV[i])), \text{Math.log}(C[m]))\}}; \quad \text{----(5)}$$

for m classes

Step 5: Finally, features with highest eigen sum values are taken as subset of features for classification problem.

Proposed Classification Algorithm
Step 1: Partition the given training dataset into 'm' classes.
Step 2: For each partition
Step 3: do
Step 4: Apply the decision tree classification feature selection measures one each partition.
Proposed hybrid measure is used to minimize the runtime (ms) and to improve the true positivity and false negative rate on large datasets. Since, network training data have nominal attributes and numerical attributes, proposed measure is used to find the nominal association between the numerical and nominal attributes using the following measures for node selection.
Hybrid Entropy Measure for Hoeffding Tree Construction
Let NDbe the normalized selected feature, cbrt is the cube root, chiVal is the chisquare value. The hybrid entropy of the Hoeffding tree is given by

$$n = \sum ND[i].\log(ND[i])$$

$$\text{Ent}(D_p) = \frac{(n + \log(\sum ND[i])}{\sqrt{(\sum(ND[i] * (ND[i] - \mu_{ND})^2}}$$

$$\text{HCondEntropy}(D_p) = \frac{-\text{Math. cbrt}(\text{entropyConditional}(ND[i]) * \text{total}) * n}{(\text{CramersV}(ND) + \text{chiVal}(ND))}$$

$$\text{CramersV}(D_p) = \text{Math. sqrt}(\text{chiVal}(ND)/(\sum ND[i]. * \min\{\text{nrow}, \text{ncolumns}\})$$

$$\text{chiVal(ND)} = \text{yates chisquare value for ND.}$$

$$\text{entropyConditionalent(ND)} = -(\sum \text{ND[i].log(ND[i])}/(\text{logm} * \sum \text{ND[i]})$$

where m represents m classes.

Proposed Random forest attribute selection measure (RFASM)

$$\text{Modified Gain} = e^{-n/(\log 2 * \sum \text{ND[i]})} + \text{Gain(D)}$$

$$\text{HRTASM} = \frac{-n * \sqrt[3]{\text{Ent(ND)}}}{(\text{ND[i].chiVal(ND)})^3}$$

Step 5: Construct the decisiontree using the enhanced attribute selection measure to improve the true positive rate of the classification.
Step 6: Apply ensemble learning model on the proposed decision tree and base classifiers.

EXPERIMENTAL RESULTS

To evaluate the performance of the proposed model to the existing models, different microarray datasets were selected from the biomedical repository. Different dataset used for experimental evaluation are summarized in Table 1. In the experimental results, 10% of the training data are used as testing data for performance evaluation. Proposed feature selection-based ensemble methods increase the performance of true positive rate and accuracy on entire high dimensional datasets. Proposed model uses the entire training data set for construction of decision patterns; therefore, the prediction accuracy of each cross validation tends to be more accurate than the traditional ensemble classification models. From the experimental results, it is clear that proposed ensemble classification improves the overall true positive and false negative rate. Also, the main advantage of using proposed model is to reduce the error rate on high dimensional features.

From the experimental results, it is clear that proposed ensemble classification improves the overall true positive and false negative rate. Also, the main advantage of using proposed model is to reduce the error rate on high dimensional features.

Table. 1 Datasets and Its Characteristics

Micro array Datasets	Gene sets	Data-Type
Prostate	2136	Continuous/Numeric
Lymphoma	5000	Continuous/Numeric
DLBCL-Stanford	4000	Continuous/Numeric
Breast cancer	24481	Continuous/Numeric
Leukemia	7129	Continuous/Numeric

Proposed model increase the performance of accuracy and error rate on entire high dimensional microarray datasets. Proposed classification model uses high dimensional data set to generate decision patterns; therefore, the prediction accuracy of each cross validation tends to be more accurate than the traditional ensemble classification models.

Leukaemia Data Results

```
PROPOSED PATTERNS
------------------

AFFX-CreX-5_at <= -202
|  AFFX-BioC-5_at <= 88: ALL (3.21)
|  AFFX-BioC-5_at > 88
|  |  AFFX-BioC-5_at <= 241: AML (13.5)
|  |  AFFX-BioC-5_at > 241
|  |  |  AFFXBioB_5_st <= -144: ALL (2.89)
|  |  |  AFFXBioB_5_st > -144: AML (9.96/1.29)
AFFX-CreX-5_at > -202
|  AFFX-BioC-5_at <= 38: AML (8.04/1.29)
|  AFFX-BioC-5_at > 38
|  |  AFFXBioB_5_st <= -56: AML (11.25/4.5)
|  |  AFFXBioB_5_st > -56
|  |  |  AFFXBioB_5_st <= 506: ALL (20.25/0.64)
|  |  |  AFFXBioB_5_st > 506: AML (2.89/0.64)

Number of Leaves  :  8

Size of the tree :       15


Weight: 2.03

PROPOSED PATTERNS
------------------

AFFX-BioC-5_at <= 328
|  AFFX-BioDn-5_at <= -246: ALL (58.5/13.27)
|  AFFX-BioDn-5_at > -246
|  |  AFFX-BioC-5_at <= 141: AML (5.09)
```

```
| | AFFX-BioC-5_at > 141: ALL (3.82/1.27)
AFFX-BioC-5_at > 328: AML (4.59)
```

Lung-cancer Michigan

```
PROPOSED PATTERNS
------------------

AB000114_at <= 91
| AB000220_at <= 616.6: Tumor (24.6)
| AB000220_at > 616.6: Normal (4.09/0.88)
AB000114_at > 91
| AB000114_at <= 118: Normal (48.33/2.19)
| AB000114_at > 118: Tumor (18.99/3.36)

Number of Leaves  :  4

Size of the tree :  7


Weight: 2.64

PROPOSED PATTERNS
------------------

AB000114_at <= 91: Tumor (21.44/1.72)
AB000114_at > 91
| AB000460_at <= 942.6
| | AB000449_at <= 99: Tumor (3.9/0.86)
| | AB000449_at > 99
| | | AB000449_at <= 130.9: Normal (35.18/0.58)
| | | AB000449_at > 130.9
| | | | AB000449_at <= 164.4: Tumor (14.62)
| | | | AB000449_at > 164.4: Normal (13.54/0.08)
| AB000460_at > 942.6: Tumor (7.32/0.9)

Number of Leaves  :  6

Size of the tree :  11


Weight: 3.1

PROPOSED PATTERNS
------------------

AB000409_at <= 502.8
| AB000220_at <= 702.8
| | AB000409_at <= 236.6: Tumor (3.83)
| | AB000409_at > 236.6
| | | AB000449_at <= 151.9: Normal (64.06/5.67)
| | | AB000449_at > 151.9: Tumor (3.43/0.48)
| AB000220_at > 702.8
| | AB000114_at <= 228.3: Tumor (9.61)
| | AB000114_at > 228.3: Normal (6.56)
AB000409_at > 502.8: Tumor (8.51)

Number of Leaves  :  6

Size of the tree :  11


Weight: 2.68

PROPOSED PATTERNS
------------------

AB000449_at <= 151.5
| AB000409_at <= 248.7: Normal (12.32/1.35)
| AB000409_at > 248.7
| | AB000220_at <= 523.9: Tumor (38.9/0.51)
| | AB000220_at > 523.9
| | | AB000220_at <= 702.8
| | | | AB000409_at <= 371.9: Tumor (3.99)
```

```
| | | | AB000409_at > 371.9
| | | | | AB000114_at <= 54.1: Tumor (2.22)
| | | | | AB000114_at > 54.1: Normal (14.71/0.32)
| | | AB000220_at > 702.8: Tumor (8.43)
AB000449_at > 151.5: Normal (15.43/2.86)

Number of Leaves  :  7

Size of the tree :        13

Weight: 2.89

PROPOSED PATTERNS
------------------

AB000409_at <= 236.6: Tumor (19.46)
AB000409_at > 236.6
|  AB000460_at <= 578.8
|  |  AB000114_at <= 53.4: Tumor (2.06)
|  |  AB000114_at > 53.4: Normal (10.72/0.22)
|  AB000460_at > 578.8
|  |  AB000449_at <= 118: Tumor (18.67)
|  |  AB000449_at > 118
|  |  |  AB000449_at <= 132.2: Normal (9.76/1.99)
|  |  |  AB000449_at > 132.2
|  |  |  |  AB000460_at <= 599.3: Normal (2.81)
|  |  |  |  AB000460_at > 599.3
|  |  |  |  |  AB000449_at <= 178.3: Tumor (17.65)
|  |  |  |  |  AB000449_at > 178.3
|  |  |  |  |  |  AB000449_at <= 188.1: Normal (5.27/1.45)
|  |  |  |  |  |  AB000449_at > 188.1: Tumor (9.6)

Number of Leaves  :  9

Size of the tree :        17
Average Classificatin Accuracy :0.972
Average TP Rate :0.969
Average Recall :0.972
Mean Absolute Error :0.047
Average Runtime of Each partition :2932.13
```

Table 2: Comparative analysis of present approach to the traditional approaches by using accuracy on different Microarray dataset.

Model	DLBCL	Prostate	Lymphoma	BreastCancer
PCA+Ensemble	87.45	91.43	90.54	85.35
ACO+Ensemble	86.35	89.13	91.33	82.54
Fuzzy PCA+Ensemble	83.24	87.34	89.23	84.67
KM+SNR+Ensemble	94.65	91.43	92.53	86.75
KM+t-test+Ensemble	95.74	94.64	95.09	87.14
KM+SAM+Ensemble	91.53	89.13	87.45	91.43
Gaussian based Deep neural network	98.93	96.35	97.14	94.92
Proposed Model	99.23	98.45	98.71	97.93

Table 2, describes the performance of the proposed model on all cancer datasets. Here, all the cancer datasets are evaluated using the proposed model to find the average true positive rate and precision rate on the high dimensional datasets. From the table, it is visualized that the present approach has better true positive rate and precision over the existing models.

Table 3: Gene-Disease features extraction using the proposed PCA model on 2000 test instances

GeneSamples	Chisquare	MI	IG	GA	PCA	ProposedPCA
GeneDis-50	63	65	63	62	66	45
GeneDis-100	63	67	62	67	67	52
GeneDis-150	65	65	60	64	60	45
GeneDis-200	65	61	60	63	64	53
GeneDis-250	61	63	65	68	65	38
GeneDis-300	59	61	62	66	65	47
GeneDis-350	68	65	60	60	60	48
GeneDis-400	64	59	66	60	58	52
GeneDis-450	59	65	64	60	59	49
GeneDis-500	60	64	58	64	64	40
GeneDis-550	61	60	62	64	63	37
GeneDis-600	62	67	66	58	62	43
GeneDis-650	59	65	63	63	62	41
GeneDis-700	60	64	61	59	64	50
GeneDis-750	64	67	65	64	66	39
GeneDis-800	67	61	67	60	59	37
GeneDis-850	59	63	63	65	61	48
GeneDis-900	63	64	64	58	60	42
GeneDis-950	60	62	61	64	59	50
GeneDis-1000	62	65	68	64	58	47

Table 3, illustrates the performance of gene-disease feature extraction using the proposed PCA approach on large datasets. From the table1, it is clearly shown that the present feature extraction procedure has high filtering rate as compared to the existing approaches.

Table 4: Performance analysis of computational runtime(ms) with different traditional feature selection models

GeneSamples	Chisquare	MI	IG	GA	PCA	ProposedPCA
GeneDis-50	6658	5180	5153	6639	7122	6511
GeneDis-100	5249	6220	7146	7035	5392	7279
GeneDis-150	5739	6731	6140	6861	6886	6564
GeneDis-200	6869	6729	6010	6750	5545	5973
GeneDis-250	5377	6886	6989	7088	5154	6146
GeneDis-300	6710	5935	7350	6345	7307	6179
GeneDis-350	6648	6163	6805	7383	5310	5904
GeneDis-400	6377	5462	5879	6262	6900	5597
GeneDis-450	7119	6227	6161	5699	6386	5406
GeneDis-500	6151	6522	5643	6736	6942	5219
GeneDis-550	7071	6287	5701	5734	5872	5138
GeneDis-600	6996	6607	6688	6952	5325	5915
GeneDis-650	5344	6132	7086	7151	5502	5559
GeneDis-700	5185	5369	6178	5468	6126	6410
GeneDis-750	6148	6317	6482	5613	6011	7137
GeneDis-800	6020	5046	6739	5660	5406	5294
GeneDis-850	6654	5173	6042	5532	6210	5355
GeneDis-900	6508	5717	5706	5775	5970	6828
GeneDis-950	6568	6470	5600	6005	6357	5680
GeneDis-1000	6700	7251	6893	5744	5407	7199

Table 4, describes the performance of computational runtime(ms) of gene-disease feature extraction using the proposed PCA approach on large datasets. From the table3, it is clearly shown that the present feature extraction procedure has low computation runtime as compared to the existing approaches.

Recall

Table 5: Performance analysis of recall using different traditional classification frameworks.

Gene Cancer Samples	SVM	RF	NN	CNN	Proposed Classifier
GeneDis50	0.8	0.82	0.87	0.93	0.98
GeneDis100	0.78	0.83	0.87	0.94	0.98
GeneDis150	0.81	0.82	0.88	0.94	0.98
GeneDis200	0.79	0.83	0.86	0.93	0.97
GeneDis250	0.82	0.84	0.86	0.93	0.97
GeneDis300	0.78	0.84	0.87	0.93	0.98
GeneDis350	0.82	0.84	0.87	0.93	0.98
GeneDis400	0.82	0.83	0.86	0.92	0.98
GeneDis450	0.82	0.83	0.87	0.93	0.98
GeneDis500	0.81	0.84	0.87	0.94	0.97
GeneDis550	0.81	0.82	0.87	0.94	0.98
GeneDis600	0.8	0.82	0.86	0.94	0.98
GeneDis650	0.79	0.83	0.87	0.92	0.98
GeneDis700	0.8	0.83	0.85	0.94	0.98
GeneDis750	0.8	0.84	0.86	0.92	0.98
GeneDis800	0.81	0.83	0.87	0.95	0.97
GeneDis850	0.78	0.82	0.88	0.93	0.98
GeneDis900	0.8	0.84	0.85	0.93	0.98
GeneDis950	0.8	0.83	0.87	0.93	0.98
GeneDis1000	0.81	0.83	0.87	0.93	0.98

Table4, describes the performance of recall ofgene-disease classification using the proposed classification framework on large datasets. From the table4, it is clearly shown that the present framework has high computational recall as compared to the existing frameworks.

Accuracy

Table 5: Performance analysis of accuracy using different traditional deep learning frameworks.

Gene Cancer Samples	SVM	RF	NN	CNN	Proposed Classifier
GeneDis50	0.81	0.84	0.87	0.95	0.98
GeneDis100	0.78	0.83	0.87	0.93	0.98
GeneDis150	0.8	0.84	0.86	0.93	0.98
GeneDis200	0.78	0.83	0.87	0.92	0.98
GeneDis250	0.8	0.82	0.86	0.94	0.97
GeneDis300	0.78	0.83	0.87	0.93	0.97
GeneDis350	0.81	0.82	0.86	0.92	0.98
GeneDis400	0.81	0.83	0.87	0.94	0.98
GeneDis450	0.8	0.83	0.88	0.92	0.99
GeneDis500	0.81	0.83	0.88	0.92	0.98
GeneDis550	0.8	0.84	0.87	0.94	0.97
GeneDis600	0.79	0.82	0.85	0.93	0.99
GeneDis650	0.82	0.83	0.88	0.95	0.98
GeneDis700	0.82	0.83	0.86	0.93	0.98
GeneDis750	0.82	0.83	0.87	0.93	0.98
GeneDis800	0.81	0.83	0.86	0.93	0.98

GeneDis850	0.82	0.83	0.85	0.94	0.98
GeneDis900	0.81	0.84	0.86	0.95	0.98
GeneDis950	0.8	0.82	0.88	0.92	0.98
GeneDis1000	0.81	0.83	0.87	0.92	0.98

Table 5, describes the performance of accuracy of gene-disease classification using the proposed classification frameworks on large datasets. From the table5, it is clearly shown that the present framework has high computational accuracy as compared to the existing frameworks.

CONCLUSION

Ensemble classification algorithm with weighted function is used to find the essential feature sets from the large number of feature space. Since, the weights in the deep neural network is optimized using the weighted function and the logistic function, proposed model efficiently classifies the large data with high dimensionality. Most of the traditional feature transformation approaches such as log transformation, min-max normalization etc. are independent of data distribution and outliers. Traditional PSO based ensemble learning and ABC based ensemble learning are improved using the heuristic activation function and ensemble classification measures. To solve these problems effectively, it is therefore essential to develop an algorithm that can identify reliable candidates for disease using existing associations of gene-disease verified by the biological experiment. In the proposed work, a hybrid feature transformation, hybrid feature selection and advance classification approach are implemented to improve the true positive rate and error rate of the disease prediction. A novel principal component ranking measure is integrated in order to find the subset of features for classification problem. Finally, a hybrid decision tree classifier is used to predict the classification accuracy on the selected features set. Experimental results proved that the present framework has better performance compared to the traditional models for variable microarray datasets.

REFERENCES

1. M. Ghosh S. Begum R. Sarkar D. Chakraborty U. Maulik "Recursive memetic algorithm for gene selection in microarray data" Expert Systems with Applications vol. 116 pp. 172-185 2019.
2. Z. Rustam I. Primasari D. Widya "Classification of cancer data based on support vectors machines with feature selection using genetic algorithm and laplacian score" AIP Conference Proceedings vol. 2023 no. 1 pp. 020234 2018.
3. V.B. Canedo N.S. Marono "A Review of Microarray Datasets and Applied Feature Selection Methods" Information Sciences pp. 111-135 2014.
4. Q. Su "A Cancer Gene Selection Algorithm Based on the K-S Test and CFS" Biomed Research International pp. 1-6 2017.
5. M. Morovvat A. Osareh "An Ensemble of Filters and Wrappers for Microarray Data Classification" Machine Learning and Applications: An International Journal (MLAIJ) vol. 3 no. 2 June 2016.
6. N Matamala MT Vargas R González-Cámpora R Miñambres et al. "Tumor microRNA expression profiling identifies circulating microRNAs for early breast cancer detection" Clin Chem vol. 61 no. 8 pp. 1098-106 Aug 2015.
7. K. Yan L. Ma Y. Dai W. Shen Z. Ji D. Xie "Cost-sensitive and sequential feature selection for chiller fault detection and diagnosis" International Journal of Refrigeration vol. 86 pp. 401-409 2018.
8. H. Lu J. Chen K. Yan Q. Jin Y. Xue Z. Gao "A hybrid feature selection algorithm for gene expression data classification" Neurocomputing vol. 256 pp. 56-62 2017.
9. K. Yan Z. Ji H. Lu J. Huang W. Shen Y. Xue "Fast and accurate classification of time series data using extended ELM: Application in fault diagnosis of air handling units" IEEE Transactions on Systems Man and Cybernetics: Systems 2017.
10. Y. Liu H. Lu K. Yan H. Xia C. An "Applying cost-sensitive extreme learning machine and dissimilarity integration to gene expression data classification" Computational intelligence and neuroscience 2016.
11. C. Braicu D. Gulei B. De Melo Maia I. Berindan-Neagoe G. A. Calin "Mirna expression assays" in Genomic Applications in Pathology Springer pp. 65-92 2019.
12. T. Setoyama H. Ling S. Natsugoe G. A. Calin "Non-coding rnas for medical practice in oncology" The Keio journal of medicine vol. 60 no. 4 pp. 106-113 2011.
13. M. Ghosh S. Begum R. Sarkar D. Chakraborty U. Maulik "Recursive memetic algorithm for gene selection in microarray data" Expert Systems with Applications vol. 116 pp. 172-185 2019.
14. Dane, S., Welcome, M.O.A case study: Effects of wet cupping therapy in a male with primary infertility (2019) Journal of Complementary Medicine Research, 10, pp. 155-161.
15. H. Öztoprak M. Toycan Y.K. Alp et al. "Machine-based classification of ADHD and non-ADHD participants using time/frequency features of event-related neuroelectric activity" Clin. Neurophysiol. vol. 128 no. 12 pp. 2400-2410 2017.
16. P. Viday Sagar, Nageswara Rao Moparthi, Ch. Mukesh “Smart Meter Analytics for Optimizing the Utilization of Electricity using Arima, Navie & Holt Winter” International Journal of Innovative Technology and Exploring Engineering Vol 8, PP 585-590 (2019)