International Journal of Engineering & Technology, 7 (2.7) (2018) 546-552

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research Paper



An Improved Particle Swarm Optimization based classification model for high dimensional medical disease prediction

P.R.Sudha Rani [1]*, Dr.K.Kiran Kumar [2]

[1] *Research Scholar,Department of CSE, KLEF*
[2] *Professor, Dept of Electronics and Computer Engg, KLEF*

Abstract

Recently, machine learning techniques have become popular and widely accepted for medical disease detection and classification on high dimensional datasets. Classification models is one of the essential model in machine learning models for medical disease prediction due to its fast processing speed, high efficiency and noisy datasets. Traditional machine learning models are failed to estimate the disease patterns with high true positive rate due to large number of features and data size. In this paper, a novel particle swarm optimization based hybrid classifier was implemented for medical disease prediction with high dimensions. The main objective of the feature selection based hybrid classifier is to classify the high dimensional data for large medical feature set. Proposed filtered based hybrid classifier is usually designed and implemented to improve the medical prediction rate on high dimensional data. In this work, we have used different ensemble learning models such ACO+NN, PSO+ELM, PSO+WELM to analyze the performance of proposed model(IPSO+WELM). Experimental results are evaluated on different types of medical datasets including lung cancer, diabetes, ovarian, and DLBCL-Stanford. Performance results show that proposed IPSO+WELM with ensemble model has high computational efficiency in terms of true positive rate, error rate and accuracy.

Keywords: *Machine learning, Neural network, Extreme learning, Disease prediction.*

1. Introduction

Classification can be defined as a special kind of learning model which is responsible for categorization of different gene-disease datasets. These datasets are classified into set of finite or infinite classes. Apart from supervised and unsupervised machine learning approaches, there are two other machine learning techniques generally used for classification are:- regression and clustering. A learning function generally maps original data into their real-value variable in the process of regression. This technique can estimate the predictive variable for every individual sample. Clustering is categorized under the category of unsupervised learning and here groups are formed according to the similarity of data items. The groups which are built in the process of clustering are known as clusters. Data items having high similarity are included in the same cluster, whereas data items having no similarity or least similarity are included in different clusters.

The classification scheme Bagging is categorized under a special kind of Bootstrap aggregation. Furthermore, the process of bagging also supports all characteristics of machine learning and meta-algorithm. Meta-algorithm can be defined as a specific algorithm which is developed for improvement of stabilization factor. Bagging has wide range of applications in the fields of statistical classification and regression. The process of bagging not only reduces variance, but also limits over fitting. Besides these, there exists another application of bagging classification i.e., decision trees. Some common factors are generally responsible for errors of machine learning algorithms, those are:- noise, bias and variance. Noise is generally defined as an error occurs by the target function. Biases are the targets which are not qualified to be learnt by the classification algorithms. Variance is the outcome of sampling process. The above mentioned classification approach reduces overall errors.

Boosting can be defined as a special kind of machine learning meta-algorithm. This algorithm has the prime objective of reducing bias significantly. Additionally, it is also responsible for decreasing variance. In other words, boosting is the process of transforming weak learners to strong learners. Weak learners are the learners which are very poorly correlated along with true classification. But, strong learners are strongly correlated with true classification.

Extreme Learning Machine can be defined as a single-hidden layer feed-forward neural network (SLFN) with learning model [4]. The traditional optimization approaches like gradient descent based back-propagation [5] evaluate weights and biases. The proposed technique is responsible for decreasing the training time effectively through random assignment of weights and biases. The above extended EL method results better efficiency and performance as compared to all traditional approaches. EL has wide range of applications in different domains like face recognition, human action recognition, landmark recognition and protein sequence classification, medical disease prediction [1, 3,4]. But, EL has two major issues, those are:- 1) This model has over fitting problem and the performance can't be predicted for unknown datasets. 2) This model is not applicable to binary classification and uncertain datasets.

Feature extraction based ELM:

The process of feature extraction has high significance in the field of classification. All features are divided into two groups, those are:-

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1) According to the first group, features extraction using noisy attributes and contextual information.
2) The second group contains correlated features.
Traditional feature extraction models discard noisy features in order to decrease the high dimensional features to a lower dimensional feature. Let us assume Nmin and Nmax are minimum and maximum numbers of hidden neurons respectively, where N denotes the present value of hidden neurons. For each and every N, the average accuracy rate of ELM through 10-fold cross-validation scheme is evaluated. At last, hidden neurons having maximum average accuracy is chosen as optimal. After selecting the optimal numbers of hidden neurons, the ELM classifier is implemented in order to evaluate the classification accuracy by considering the outcomes of PCA and the outcomes are averaged later.
Medical data prediction is the most important and complicated requirement in recent era. Many approaches are developed in the field of medical disease prediction such as cancer prediction. In [9], association rule mining approach is integrated with MLP and back propagation approaches in order to predict and detect the chances of breast cancer. Modular neural network is basically implemented to recognise and analyse cardiac diseases using Gravitational Search Algorithm (GSA) and fuzzy logic embedded algorithm for better performance [3]. With the exponential growth of information technologies, data mining approaches are implemented in various domains such as biomedical applications and disease prediction. Special emphasis can be given in not only detecting cancer, but also predicting the disease in an early stage [6].
A gene selection technique is generally applied to the pre-existing approaches for pattern classification. It not only decreases the influence of noise, but also reduces the error rate in medical datasets [4-6]. According to [1], gene selection technique is used to enhance the performance of the classifier through the detection of gene expression subsets. Among most widely implemented gene selection approaches, some are given below:- principal component analysis [4], singular value decomposition [5], independent component analysis [6], genetic algorithm (GA) [7], and recursive feature elimination method [8, 9].
Deep learning is a special type of back propagation scheme for multi-layer networks. It supports unsupervised initialization, rather than classical random initialization. On the other hand, this method supports multilayer based unsupervised initialization. After successful initialization, the whole network is trained through back propagation based neural networks technique. The major issue in this model is the hidden parameters of deep learning framework are fine-tuned many times. It makes the overall training phase lengthier and time consuming.
These approaches are very slow because of improper learning phases or coverage of local minimum. Additionally, a lot of iterative learning schemes are developed in order to achieve better learning efficiency and performance. This approach includes neurons in an increasing order and completely depends upon kernel based approaches. In case of Bayesian ELM, the hidden layer is not tuned unlike training of conventional feed-forward network. All the parameters of hidden layers are selected randomly. The input weights, output weights and biases are selected randomly in order to minimize the training error. The major disadvantages of the above schemes are:- slower learning rate of networks and reduced capability to manage complexities. The complexity issue of traditional neural networks are overcome by Modular Neural Network (MNN) .Modular Neural Network can be defined as a neural network where all modules are assigned with separate tasks. The outcomes of all modules are combined together in order to form the final outcome.
In the past, machine learning models used a single classification model to predict the test data using the training samples. However, multiple classifiers can be used to predict the same test data using the training samples. This process is known as ensemble learning. Ensemble classification has been successfully applied to different classification problems to improve the classification accuracy using the optimal feature selection measures.
Particle swarm optimization (PSO) is very popular optimization techniques in machine learning models. PSO is generally applied in the literature to adjust the initialization parameters of base classifiers in the ensemble learning models. The main objective of this paper is to optimize the traditional PSO parameters in the ensemble classification model in order to improve the accuracy and error rate. In the ensemble model, neural network is used as one of the base classifier and weights are initialized using the proposed PSO technique.
In the past, machine learning models used a single classification model to predict the test data using the training samples. However, multiple classifiers can be used to predict the same test data using the training samples. This process is known as ensemble learning. Ensemble classification has been successfully applied to different classification problems to improve the classification accuracy using the optimal feature selection measures.
During the last two decades, different feature selection methods and classification models have been used for prediction and diagnosis of different cancer diseases. Today, with the exponential growth of technology, it is merely impossible to use these conventional methods for cancer prediction due to the high dimensional features and imbalance properties. Machine learning models are used to classify different medical datasets such as microarray data, clinical data, and proteomic data as input. Most of the traditional approaches consider the features as independent and linear. Most of the biological systems are non-linear and its parameters are interdependent, thus machine learning has become better choice. Curse of dimensionality is another problem, where there exist more variables and fewer examples. Both machine learning and conventional methods suffer from this problem. This problem can be resolved by either decreasing the number of variables or increasing the number of training datasets. Sample-feature ratio must be more than 5:1 every time. Machine learning algorithms can be categorized into three broad types, they are:- Supervised machine learning, Unsupervised machine learning and Reinforcement machine learning. Supervised learning consists of a prescient provider which provides the labelled training dataset as input to the algorithm and produces output after mapping. But on the contrary, for unsupervised machine learning only training datasets are given as input without labels. Some of the examples of unsupervised learning are:- Self Organizing feature Maps, hierarchical clustering, k-means clustering, and so on. All machine learning algorithms used for cancer prediction come under supervised learning. Most widely applicable algorithms are:- Artificial Neural Networks, Decision Trees, Genetic algorithms, Linear Discriminant Analysis, k-Nearest Neighbour etc.
The problem of gene extraction or selection has become a main challenging factor in the field of microarray disease datasets. There are two types of feature selection models, the filter based feature selection and wrapper based feature selection. Filter based feature selection rank features according to certain statistical measure such as the information gain; gain ratio, t-statistic etc. The wrapper based feature selection models are more complex than filter based feature selection, because they require a training dataset to select subset of features. Wrapper based methods finds a subset of features and predicts its relevance using a machine learning classifier.

2. Related Work

Sarkar, et.al, developed an advanced classifier ensemble framework in order to classify huge amount of multimedia big data [1]. There exist a numbers of different classification approaches for different data types. A single classifier can't give optimized results in all types of datasets. Hence, an ensemble learning approaches have become very much popular. Generally accuracy of the ensemble learning approaches depend on base classifiers and data dimensionality. In this work, an advanced classifier ensemble framework is developed along with a group of predefined base classifiers. The main objective of this approach is to classify the

test instances based on the majority voting of multiple base classifiers. This ensemble system was executed on Spark environment in order to process large quantity of big data. Series of experiments are performed on multimedia big data and showed that this framework performs better than that of other traditional classification algorithms. Further future works can be carried out in order to modify and enhance the ensemble model with high dimensionality. Advanced techniques can be developed to resolve multi-class classification issues with big datasets.

Myoung, et.al, proposed a mobel to resolve the issues of evolutionary unsampling for imbalanced big data classification [2]. Most of the applications require appropriate classification approaches in order to classify big data. The dimensionality issue of microarray data has become more complex in case of imbalanced big data classification process. A proper solution for imbalanced classification is evolutionary ensemble approaches. In this paper, an optimized MapReduce approach is implemented to classify imbalanced data. The whole process can be broadly divided into two phase MapReduce. In the initial phase, a decision tree is learnt on every individual map after successful completion of pre-processing and prediction of test samples using trained dataset. A windowing approach is applied in order to enhance the EUS process. EUS is implemented on arbitary sized datasets in order to achieve parallelization. In future, this model will be extended to implement hybrid oversampling or undersampling techniques instead of EUS in order to enhance the system's performance.

I. Triguero, et.al, introduced a new approach for evolutionary under-sampling for extremely imbalanced big data classification [3]. Ensemble under-sampling technique is considered as a perfect solution for this skewed class distribution. Due to memory and time limitations, an evolutionary approach can't be implemented in high dimensional applications. In order to resolve such types of problems, divide-and-conquer techniques are implemented using MapReduce scheme. According to this technique, the whole dataset is divided into numbers of subsets of data. In the high dimensional data, divide-and-conquer methods are failed to find the majority and minority problems. [3] proposed a novel technique using the Apache Spark. The main objective of this model is to solve the problem of extremely imbalanced datasets. In the proposed technique, majority and minority class instances are handled independently in order to decrease the influence of small sample size. The process of evaluation is carried out along with different datasets up to one million instances. In future, several preprocessing schemes like hybrid oversampling or under-sampling techniques can be implemented to improve the performance of the true positive rate and accuracy.

B Sudha *et.al,* developed a new approach for efficient big data classification [4]. They implemented a novel model know as Random projections fuzzy k-nearest neighbor (RPFKNN). As the size of the data features increases, there is need to minimize the dimensionality of feature space for classification models. Let us consider an example of activity recognition process. Here, a large amount of high dimensionality data are required to process real time classification applications such as medical disease prediction and sensor applications. Some traditional dimension reduction techniques like PCA and feature selection models are appropriate for datasets of limited size and high dimensionality. In this paper, they integrated traditional classifier with fuzzy k-nearest neighbor and feature selection through random projection technique in order to carry out classification of big datasets. In PCA approach, the projection matrix is evaluated only once through a least square optimality. But in random projections, the projection matrix is evaluated more than once. A fusion scheme is required to combine these outcomes of random projection technique and it is named as RPFKNN. This technique usually depends upon class membership values generated by FKNN and classification accuracy.

Zhao et.al, developed a cloud-based framework for big data applications[5]. Classical classification models and feature reduction methods are incapable of processing these big data efficiently. The framework involves two phases i.e., feature extraction and machine learning phases evaluated in both training mode as well as testing mode. Proposed technique gives a better solution for big data classification with limited feature components. The automatic adjustment of classification makes this approach very much efficient as compared to other traditional classification approaches.

Lan, et.al, designed and developed a solution to imbalanced classification on small datasets with limited feature space[6]. The main problems of the traditional classification models are:- imbalanced data , high dimensionality and error cost estimation. Here, a new distributed evolutionary cost sensitive balancing approach is proposed in order to handle imbalanced data as well as error cost estimation. At first, a genetic algorithm is applied to obtain an optimal cost matrix along with the base classifier. In this paper, two distinguished alternatives are proposed which are given below:- : a computation-oriented technique and a data-oriented technique. These two techniques are implemented using the Apache Watchmaker framework. Apart from this, both the techniques are deployed in Hadoop-based environment. Series of experiments are carried out in order to evaluate both the approaches and demonstrated that, the computation-oriented technique results better performance than that of the other one. Besides, it is incapable of decreasing the overall execution time. In case of data-oriented techniques, they can easily decrease the overall execution time significantly compared to traditional SVM and kNN. The main limitations of these models are high runtime, high error rate . In future, this work can be extended to implement faster algorithms like ensemble decision trees.

Begum, et.al, introduced an adaptive classification scheme for microarray analysis using big data [7]. Microarray technology is used for study of biological operations within cells. Each and every cell contains some gene expression data those are represented by microarray. Microarray data is responsible for detecting required information out of numerous genes within a common cell at the same time. By studying these gene expressions thoroughly, detection of several diseases like viral infection and cancer can be identified at a very early stage of diagnosis. Let us consider a sample tissue where numbers of different genes are expressed. Among the total numbers of genes in a sample tissue, many are not related to medical diagnosis. Diseases can be properly classified when among vast numbers of genes a small portion of genes are chosen. [7] model emphasizes on some commonly implemented gene expression classification schemes based on MapReduce framework. They also worked for detecting differentially expressed genes and choose relevant genes out of gene expression. They cross validated their theory and performed evaluation process. The outcomes of evaluation process show better performance as compared to other conventional classification techniques.

Huang,et.al, proposed an advanced hybrid classification approach for microarray datasets [8]. In this research work, they developed a new hybrid technique in order to improvise the traditional microarray data classification approaches. Their proposed technique basically depends upon the base classifiers such as nearest neighbour, Naïve Bayes and SVM. Later the performance of given approach is compared with classical classification techniques (such as SVM, Naïve Bayes, and so on). The process of feature selection is successfully executed with the help of discrete wavelet transform and moving window approach. The hybrid classifier is developed by merging the features of NN, NB and SVM. During the classification phase, similar feature are retrieved and compared with the references generated in training phase. A three-fold cross validation process is carried out to measure the robustness of the system. Each and every fold is associated with a single classifier to test the system's accuracy. All generated classification accuracies from classifiers are assumed to have same weight in case of hybrid classifiers. This technique is evaluated over five gene expression datasets. Numbers of experiments are performed on both real as well as benchmark datasets. The main limitations of this model are feature space and feature subset selection.

Zong, et.al, proposed a new approach for big data classification by using RIPPER [9]. They termed their approach as distributed multiclass rule based classification scheme. There is necessity of more

computational resources with respect to space and time in order to retrieve the hidden patterns. Hence, distributed computing has gain popularity in the data mining process. The limitation of space and time can be overcome by presenting such data mining approach which will be capable of mining in a non-centralized manner. Here, they have implemented a new algorithm for distributed multiclass rule based classification (DiRUC). DiRUC is responsible for applying repeated evolutionary pruning strategy to generate error reduction at local level. These error reduction schemes are combined together at global level in a distributed manner by integrating all the local models, the global model is built and placed at all sites for next class label prediction. The outcomes of evaluation phase show better performance as compared to conventional RIPPER approach. The main limitation of the model is to high error rate and false positive rate as the number of dimensions is increasing.

Huang et.al, presented a new technique for gene selection and decision tree based classification [10]. The main objective of this paper is to predict cancerous severity class labels. Due to the high-dimensionality of microarray gene dataset and gene expression sequences, the process of data classification is getting more complicated in the traditional models. High dimensional problems such as gene selection and classifier construction can be minimized by advanced rule based classification techniques [12-16]. Rough set theory and decision tree approaches are integrated with the mathematical and statistical techniques to resolve these issues. In the process of evaluation, the overall accuracy of rule-based classifier J48 is computed for each and every dataset. In the classification model, the ranking of genes are computed by correlation and Jaccard coefficient to predict the cancer label.

3. Proposed Methodology

Proposed PSO based ensemble classifier is a multi-objective technique which finds the local and global optimum solutions by iteratively searching in a high dimensional space. If the attribute is numerical and the value is null then it is replaced with computed value of equation (1). Also, if the attribute is nominal and has missing values then it is replaced with computed value of equation (2). Finally, if the class is numeric then it is converted to nominal.

Proposed PSO based ensemble model is usually designed and implemented to improve the overall classification true positive rate on high dimensional feature selection. Generally, ensemble learning model is constructed from a group of base classifiers to predict the high dimensional class labels. Here, search space of the traditional PSO model is optimized using the proposed measures such as optimized fitness measure, candidate solutions, fitness measure and chaos gauss based randomization measure. In our model, different base classifiers such as ACO+NN,PSO+ELM,PSO+WELM, IPSO+WELM are used to test the performance of proposed model to the traditional models.

Proposed Improved PSO based Ensemble Classification Algorithm:

Step 1: Data pre-processing on Training high dimensional data.

Input : Input Data T_r.

Output: Preprocessed Data.

Procedure:

Initialize training dataset T_r.
For each attribute A(i=1,,n) in T_r
Do
If(A[i] is numeric)
Then
For each attribute value in A[i]
do
If(A[j]!=null)

$$\text{SNormalize}(A[j]) = | A[j] - \mu_{A[j]} | / (\text{Max}\{\mu_{A[j],S[m]}\} - \text{Min}\{\mu_{A[j],S[m]}\})$$

Else

$$\text{SNormalize}(A[j]) = | A[k] - SD_{A[k]} | / (\text{Max}\{\mu_{A[k],S[m]}\} - \text{Min}\{\mu_{A[k],S[m]}\}); \text{ wh}$$

End if
If(A[j] is nominal)
Then

$$\text{SNormalize}(A[j]) = \max\{\text{Prob}(A[i]/A)\}; i! = j$$

Where S_m represents the m –classes

$\mu_{A[j],S[m]}$ Represents the mean of the each mth class.

End if
Done

Step 2: Proposed PSO algorithm

Step 2: //Optimal feature selection for feed-forward neural network and Decision tree techniques

Initializing particles with feature space, number of iterations, velocity, number of particles etc.

Compute hybrid velocity and position for each particle in 'd' dimensions using the following equations

$$V_p(d(i+1),i) = \xi_p.[\omega_p(d(i),i).V_p(d(i),i).\phi_{m1}(pPBst_i - X_p(d(i),i)) + \phi_{m2}(gPBst_i - X_p(d(i),i))]$$

$$X_p(d+1,i) = X_p(d(i+1),i) + V_p(d(i+1),i)$$

ξ_p is the particle convergence parameter formulated as

$$\xi_p = \frac{1}{|2\pi - (\theta_{m1}+\theta_{m2}) - \sqrt{(\theta_{m1}+\theta_{m2})^2 - 4(\theta_{m1}+\theta_{m2})}|}$$

where $\theta_{m1}, \theta_{m2} \in$ Orthogonal gaussian functions

In this optimized model, particle inertia weights are calculated using the following equation as

$$\omega_p(d(i+1),i) = \omega_{p_max} - (\text{Iter}_{current} / \text{Iter}_{max}).(\omega_{p_max} - \omega_{p_min})$$

ω_{p_max} : maximum inertia

ω_{p_min} : minimum inertia

Iter_{max} : maximum iteration

$\text{Iter}_{current}$: current iteration

Step 3: Computation of probabilistic correlated fitness value for orthogonal Gaussian function.

In the hybrid particle swarm optimization model, fitness value and initial logistic chaotic parameters are computed using the following equation as

$$\text{Rand}_i = \max\{\psi_j^r(1-\psi_j^r), \frac{1}{\psi_X\sqrt{2\pi}} e^{-\frac{(X-\mu_X)}{\sqrt{\sigma_X}}}\}$$

$r = 1, 2 \ldots n$

$\psi_j^k \in \text{logisticchaostic}(0,1)$

Proposed fitness method is given as

$$\text{Fit}_i = k_1.TP_i + k_2.(1 - \frac{\sum_{i=1}^{|F|} fs_i}{N_A})$$

where $k_1, k_2 \in R_i$

fs_i is the feature selected labels 1 or 0 .

'1' indicates selected attribute ,

'0' indicates non-selected attribute.

$|N_A|$ indicates total attributes size.

TP_i : Truepositive rate of the selected attributes in the current iteration.

Step 4: Apply KNN classification model with highest true positive rate acc_i on the selected features in the ith iteration.

Step 5: To each particle compute fitness value and true positive rate of the attribute in the previous step.
Step 6: Update the particle position, velocity, local best, global best, fitness and logistic chaos randomization functions.
Step 7: This procedure is repeated until minimum error rate or high true positive rate.

Algorithm2: Proposed Weighted PSO- ELM algorithm based Hybrid Classifier

Basically, extreme learning machine is a single hidden layer feed forward network, where the instance weights are initialized on the normalized data. After weight initialization, instance biases are computed on the correlated objects. Finally, weights and biases are updates based on the error rate. For extreme learning machine, the number of neurons, error rate has a great influence on the evaluation of the model. If the number of neurons and error rate are high, then the proposed model will minimizes the error rate on the training dataset.

Input:
Normalized data ND, Max iterations Max, bias λ, Weighted vector W, c is class label ,I is instance , I[c] is instance class, I[t]: t Attribute instance value.
Output : class label prediction.
Procedure:
Initialize weights to all instances.
For each instance in ND
Do
$W_d \leftarrow 0$
Done
Initialize bias to 0.
$\lambda \leftarrow 0$
For each iteration i=1,2,…Max
do
For each instance I in D do
$R \leftarrow \sum_{d=1}^{\#Attributes} \log(W_d).I_t + \lambda$
Done
Done
if R(log(W)I+bias)<0
Then
For each attribute t do
$W_d \leftarrow \log(W_d) + I_c.I_t$
$\lambda = \lambda + R$
Done
// update previous cached weights
$W_{p,t} \leftarrow |\log(W_{p,t})| + R.C_t.I_t$
// update cached bias values
$\lambda_p = \lambda_p + I[c].C_t$
End if
$C_t = C_t + 1$
For each object O_i in ND[]
Do
For each object O_j in ND[] // where i!=j
Compute distance Dist=Correlation(O_i, O_j).
Done
N_m^k []=Find Top K- nearest objects from Sorted list of Dist;
Assign a class to p' based on majority vote: $c'=argmax_y \sum(xi,ci)$ belonging to S, $I(y=ci)$
Done
return $(W-\frac{W_{p,t}}{\sqrt{C_t}}, \lambda - \frac{\lambda_p}{\sqrt{C_t}})$
Where, $R^{'}$ always greater than 1.

Algorithm 3: Hybrid classifier Steps

$D_0 \leftarrow <1/D_1, 1/D_2, 1/D_3, ...1/D_N>$ N uniform data partitions
$C_1 \leftarrow PSO - ELM$
$C_2 \leftarrow IPSO - WELM$
$C_3 \leftarrow ACO - NN$
$C_4 \leftarrow Neural network$
for each classifier k in C_i do
$C_i = classifier(D, D_{k-1})$ // Training phase
$\hat{P}_n \leftarrow C_i(x_n), \forall n$ prediction on training instances
$\hat{E}_k \leftarrow \sum_n D_{k-1}[n] \ / P_n \neq \hat{P}_n$
$\phi_k = \log(\frac{1-\hat{E}_k}{\hat{E}_k})$
end for
return $f(\hat{m}) = sgn[\sum_k \phi_k . C_k(\hat{m})]$

4.Experimental Results

In this section, experimental results are performed on cancer medical datasets using proposed feature selection based classification model. Different datasets [11] used for experimental evaluation are summarized in Table 1. In the experimental results, 10% of the training data are used as testing data for cross-fold validation. In order to fully validate the analysis of the proposed model, different types of datasets are used in the experimental study, including continuous, nominal and mixed.

True negative measure evaluates the ratio of number of instances that are not affected with the cancer disease, which are identified correctly.

True positive measure defines the ratio of number of cancer disease instances that have been predicted as positive rate.

Mean Absolute error Error: Average misclassification rate of each test data in the cross validation.

Precision measure computes the ratio of correctly predicted cancer instances among the entire disease affected instances.

Table 1: Datasets and its properties

Dataset Name	Features	Type
lungCancer	12533	Continuous
DLBCL-ovarian	4219	Continuous
lung-Michigan	7230	Continuous

Sample Ovarian Dataset

```
@attribute MZ19969.128 numeric
@attribute MZ19971.766 numeric
@attribute MZ19974.404 numeric
@attribute MZ19977.042 numeric
@attribute MZ19979.68 numeric
@attribute MZ19982.319 numeric
@attribute MZ19984.957 numeric
@attribute MZ19987.596 numeric
@attribute MZ19990.235 numeric
@attribute MZ19992.874 numeric
@attribute MZ19995.513 numeric
@attribute Class {Cancer,Normal}

@data

0.494626,0.263735,0.321841,0.220934,0.297622,0.316458,0.154763,0.223685,0.304346,0.241757,0.277106,0.253731,0.371796,0.270002,0.490385,0.373736,0.3871,0.402174,0.3111
0.258063,0.406593,0.321841,0.069771,0.333335,0.354432,0.321431,0.14474,0.260869,0.142853,0.518075,0.328356,0.346156,0.270002,0.355769,0.414139,0.290324,0.586957,0.288
0.537636,0.032966,0.321841,0.209307,0.404762,0.113927,0.369049,0.223685,0.536231,0.131865,0.469877,0.044775,0.333335,0.33,0.365384,0.373736,0.354838,0.391305,0.233332
0,0.395605,0.310347,0.197673,0.404762,0.455701,0.416666,0.210527,0.420292,0.274723,0.54217,0.313431,0.179488,0.249998,0.240387,0.595962,0.215058,0.467389,0.177778,0.2
0.526884,0.395605,0.367817,0.383719,0.488099,0.392405,0.238094,0.5,0.362316,0.274723,0.28916,0.462688,0.410256,0.450001,0.346153,0.535352,0.268819,0.413043,0.222222,0
0.39785,0.395605,0.298853,0.372092,0.333335,0.1519,0.428571,0.565794,0.275362,0.395605,0.493973,0.298506,0.346156,0.379999,0.365384,0.454547,0.451614,0.521741,0.38888
0.64516,0.307689,0.241378,0.372092,0.392857,0.316458,0.249999,0.434212,0.550723,0.263735,0.445782,0.686562,0.294875,0.430002,0.346153,0.414139,0.462367,0.510872,0.077
0.720432,0.351644,0.344829,0.465117,0.452385,0.240511,0.285718,0.5,0.405793,0.186814,0.228915,0.313431,0.346156,0.869999,0,0.747472,0.408605,0.445651,0.6,0.241382,0.5
0.537636,0.307689,0.425287,0.348837,0.333335,0.354432,0.380953,0.289473,0.376808,0.35165,0.108436,0.522387,0.397436,0.470001,0.336538,0.313132,0.43011,0.641303,0.3555
0.526884,0.494503,0.333335,0.279072,0.523812,0.455701,0.34524,0.526315,0.449277,0.296701,0.325303,0.850743,0.474357,0.51,0.182692,0.515151,0.569896,0.391305,0.722224,
0.720432,0.351644,0.459768,0.244189,0.273813,0.468359,0.333335,0.605267,0.536231,0.373627,0.518075,0.447763,0.243588,0.68,0.567311,0.404038,0.580648,0.434782,0.777778
```

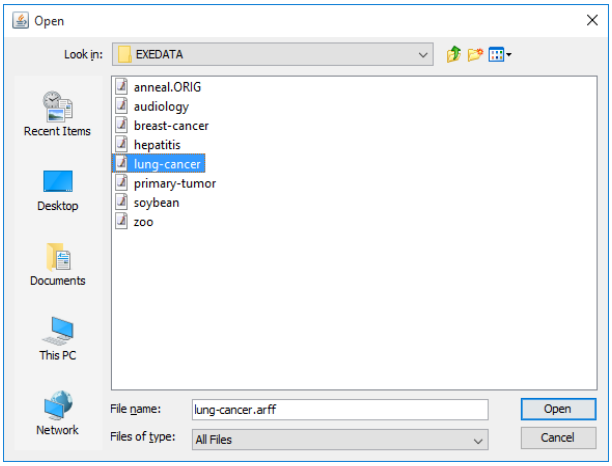


Figure 1: Input dataset

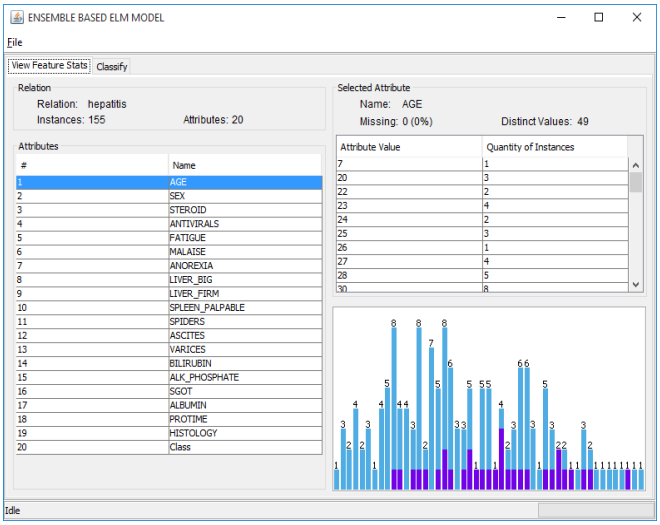


Figure 2: Statistical analysis of the input dataset.

Table 2: Performance analysis of classification accuracy on high dimensional medical datasets.

Algorithm	DLBCL	lungcancer	lung-Michigan
ACO+NN	87.43	90.65	85.97
PSO+NN	89.86	86.93	88.46
PSO+ELM	91.34	93.13	91.29
ACO+ELM	92.65	94.24	94.15
ACO+WELM	96.97	97.26	96.91
IPSO+WELM	98.42	97.98	97.81

Table 2, describes the performance analysis of proposed model to the existing models in terms of classification accuracy. Here the accuracy is computed in terms of true positive rate and true negative rate on medical dataset. From the table, it is clearly observed that the classification rate was optimized on medical dataset using our proposed feature selection based weighted ELM in ensemble classifier.

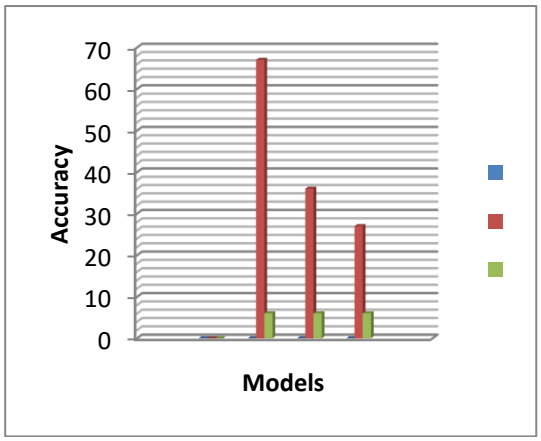


Figure 3: Performance analysis of proposed model to the existing models

Figure 3, describes the performance analysis of proposed model to the existing models in terms of classification accuracy. Here the accuracy is computed in terms of true positive rate and true negative rate on medical dataset. From the table, it is clearly observed that the classification rate was optimized on medical dataset using our proposed feature selection based weighted ELM in ensemble classifier.

5. Conclusion

In this paper, an optimized PSO feature selection method is integrated with the weighted ELM model for ensemble learning on microarray datasets. Most of the traditional feature selection based classification algorithms have computational issues such as dimension reduction, uncertainty and class imbalance on microarray datasets. Feature selection based ensemble classifier is one of the scalable models for extreme learning machine due to its high efficiency, the fast processing speed for real-time applications. Proposed filtered based hybrid classifier is usually designed and implemented to improve the medical prediction rate on high dimen-

sional data. In this work, we have used different ensemble learning models such ACO+NN, PSO+ELM, PSO+WELM to analyze the performance of proposed model(IPSO+WELM). Experimental results are evaluated on different types of medical datasets including lung cancer, diabetes, ovarian, and DLBCL-Stanford. Performance results show that proposed IPSO+WELM with ensemble model has high computational efficiency in terms of true positive rate, error rate and accuracy.

References

[1] Sarkar, I., Planet, P., Bael, T., Stanley, S., Siddall, M., Desalle R., Characteristic Attributes In Cancer Microarrays, Computers And Biomedical Research, 35(2) (2002) 111-122.

[2] Myoung-Jong Kim,,Dae-Ki Kang , "Classifiers Selection In Ensembles Using Genetic Algorithms For Bankruptcy Prediction", Expert Systems With Application 39 Pp.9308-9314, 2012.

[3] I. Triguero, M. Galar, D. Merino, J. Maillo, H. Bustince, F. Herrera, "Evolutionary Undersampling For Extremely Imbalanced Big Data Classification Under Apache Spark", "Ieee Congress On Evolutionary Computation (Cec)", Pp.640-647, 2016.

[4] S.Sudha, "Analysis Of Random Projections Fuzzy And Map Reducing K-Nearest Neighbourhood Algorithm For Big Data Classification ", " Ijcta",Pp.1813-1817, 2016.

[5] Zhao, Y.O., Chen, Y.H., Zhang, X.Q., A Novel Ensemble Approach For Cancer Data Classification, Springer, Lncs 4492 (2007), Pp.1211-1220.

[6] Lan Y, Soh Yc, Huang Gb (2010) Constructive Hidden Nodes Selection Of Extreme Learning Machine For Regression. Neurocomputing 73(16–18):3191–3199

[7] S. Begum, D. Chakraborty And R. Sarkar, "Identifying Cancer Biomarkers From Leukemia Data Using Feature Selection And Supervised Learning", "Ieee First International Conference On Control, Measurement And Instrumentation", 2016.

[8] Huang G-B, Wang D (2013) Advances In Extreme Learning Machines (Elm2011). Neurocomputing 102:1–2

[9] Zong W, Huang G-B, Chen Y (2013) Weighted Extreme Learning Machine For Imbalance Learning. Neurocomputing 101:229–242

[10] Huang G-B, Lei C, Siew Ck (2006) Universal Approximation Using Incremental Constructive Feedforward Networks With Random Hidden Nodes. Ieee Trans Neural Netw 17(4):879–892

[11] Http://Datam.I2r.Astar.Edu.Sg/Datasets/Krbd/Index.Html

[12] Annabattula, J., Koteswara Rao, S., Sampath Dakshina Murthy, A., Srikanth, K.S. And Das, R.P., 2015. Underwater Passive Target Tracking In Constrained Environment. Indian Journal Of Science And Technology, 8(35), Pp. 1-4.

[13] Ramkiran, D.S., Madhav, B.T.P., Prasanth, A.M., Harsha, N.S., Vardhan, V., Avinash, K., Chaitanya, M.N. And Nagasai, U.S., 2015. Novel Compact Asymmetrical Fractal Aperture Notch Band Antenna. Leonardo Electronic Journal Of Practices And Technologies, 14(27), Pp. 1-12.

[14] Annabattula, J., Koteswara Rao, S., Sampath Dakshina Murthy, A., Srikanth, K.S. And Das, R.P., 2015. Multi-Sensor Submarine Surveillance System Using Mgbekf. Indian Journal Of Science And Technology, 8(35), Pp. 1-5.

[15] Jawahar, A. And Koteswara Rao, S., 2015. Recursive Multistage Estimator For Bearings Only Passive Target Tracking In Esm Ew Systems. Indian Journal Of Science And Technology, 8(26),.

[16] Madhav, B.T.P., Manikanta Prasanth, A., Prasanth, S., Krishna, B.M.S., Manikantha, D. And Nagasai, U.S., 2015. Analysis Of Defected Ground Structure Notched Monopole Antenna. Arpn Journal Of Engineering And Applied Sciences, 10(2), Pp. 747-752.