AN EFFICIENT PSO BASED ENSEMBLE CLASSIFICATION MODEL ON HIGH DIMENSIONAL DATASETS

G. Lalitha Kumari[1] and N. Naga Malleswara Rao[2]

[1]Research Scholar, Acharya Nagarjuna University, Guntur, AP, India
[2]Professor, Dept. of IT, RVR & JC College of Engineering, Guntur, AP, India

Abstract

As the size of the biomedical databases are growing day by day, finding an essential features in the disease prediction have become more complex due to high dimensionality and sparsity problems. Also, due to the availability of a large number of micro-array datasets in the biomedical repositories, it is difficult to analyze, predict and interpret the feature information using the traditional feature selection based classification models. Most of the traditional feature selection based classification algorithms have computational issues such as dimension reduction, uncertainty and class imbalance on microarray datasets. 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. The main objective of the feature selection based ensemble learning models is to classify the high dimensional data with high computational efficiency and high true positive rate on high dimensional datasets. In this proposed model an optimized Particle swarm optimization (PSO) based Ensemble classification model was developed on high dimensional micro-array datasets. Experimental results proved that the proposed model has high computational efficiency compared to the traditional feature selection based classification models in terms of accuracy , true positive rate and error rate are concerned.

Keywords

PSO, Neural network, Ensemble classification, High dimension dataset.

1. INTRODUCTION

Feature selection is an important and essential technique used in data filtering for improving machine learning models on large databases. Different feature selection models have been implemented in the literature, each of them uses a feature evaluation measure to extract feature subsets on the same databases. Learning classification models with all the high dimensional features may result serious issues such as performance and scalability. Feature selection is the process of selecting subset of features so that the original feature space is reduced using the selection measures. Feature selection measures can be categorized into three types: wrappers, filters and embedded models.

Extreme learning has wide range of applications in different domains like face recognition, human action recognition, landmark recognition and protein sequence classification, medical disease prediction [1]. Extreme Learning Machine can be defined as a single-hidden layer feed-forward neural network (SLFN) with learning model [2-4]. 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. But, EL has two major issues, those are:-

1)This model has over fitting problem and the performance cannot be predicted for unknown datasets. 2) This model is not applicable to binary classification and uncertain datasets.

Feed-forward neural network can be considered as most commonly and widely implemented neural network. The method has one or more hidden layer(s) along with an output layer. The output layer is responsible for transmitting final response on the training dataset [6]. A large number of research works have been implemented in the field of Feed-Forward Neural Networks since years. This model has complex linear or nonlinear structure directly mapping from inputs. These structures are not appropriate for classical parametric constraints to manage large inputs in the traditional models. Another important feature of feed-forward neural network is the inter-dependency among the layers through parameter mapping. Single-Hidden-Layer Feed-Forward Networks (SLFNs) are treated as the most efficient and widely used feed-forward neural networks on small datasets.

In order to resolve the issues of traditional EL, weighted-EL approach is developed subsequently. The weights are increased gradually with respect to time in case of large sample size. In most of the feed forward ANN techniques, parameters of each and every layer is required to be tuned through several learning approaches. Gradient Descent-Based Approaches and Back-Propagation (BP) techniques are some important learning algorithms in feed forward neural networks [7]. The speed of learning models is very slow in case of feed forward neural networks compared to ANN. Because of better generalization capability and fast computational speed, EL approach is named as 'Extreme Learning Machine' (EL). A lot of problems are detected in case of conventional Gradient-Descent Algorithms like stopping criterion, learning rate, number of epochs and local minima [9].

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

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 EL 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 EL classifier is implemented in order to evaluate the classification accuracy by considering the outcomes of principle component analysis(PCA) and the outcomes are averaged later.

The training datasets used in this paper have a significant issue for any classification models as they have large number of feature space, ranging from 100 to 12000 features. The larger the feature space increases the search space and computational memory for disease prediction. Another crucial issue for handling the high dimensional features is the small sample size problem. The accuracy of the model employed will be reduced if the size of the training data is not sufficient relative to the feature space.

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.

The main contribution to this paper includes:

- A novel PSO based ensemble model is usually designed and implemented to improve the overall classification true positive rate on high dimensional feature selection.
- 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 and chaos gauss based randomization measure.

2. RELATED WORKS

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 medical disease prediction. In [9], association rule mining approach is integrated with multi-layer perceptron(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].

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].A micro-array feature 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].

[10-12] Implemented a PSO 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 techniques Bayesian analysis is used to predict the original data distribution using the randomization operator and the randomization data. Feature selection is an important step in defect prediction process and has been extensively studied in the field of machine learning. [10-13] Investigated different feature selection models that represent ranked list of features and applied them to UCI repository datasets. They concluded that wrapper is best model for limited data and limited feature sets. Software defect prediction models are commonly classified in the literature as Decision tree models, Support vector machine models, artificial neural network models and Bayesian models are summarized in the table 1.

Table 1: Overview of Traditional Classification Models on High dimensional Datasets

Classification Model	**Imbalance and High dimensionality Property**	**Advantages**	**Disadvantages**
Decision Trees[5]	Affected	Robust to noisy data; and decision rules evaluation.	1)Prone to over-fitting. 2)Performance issue under imbalanced property.
Artificial neural networks [7]	Affected	Able to learn non-linear functions. Robust against errors.	1) Difficult to interpret results. 2) Slow training and prediction process.
Ensemble Models [9-10]	Affected	Best model for high dimensional datasets with complex feature selection models	1)Fast processing 2)Low performance under imbalanced data and missing values.
Feature selection + wrapper method[11]	Highly Affected	The wrapper feature selection approach is useful in identifying informative feature **subsets** from high-dimensional datasets. Used traditional algorithms such as C4.5, KNN, Random forest..	**Limitations:** Accuracy is computed on the subset of features. Considers only less number of features for decision making patterns. In the microarray datasets, subset of features failed to give affective decision making patterns due to insufficient gene patterns. In order to handle feature subsets (>100), single classifier failed to perform efficient results due to memory constraints.
Ensemble classifier+ Biomedicine Field	affected	Ensemble methods proved to be superior to individual classification method for high dimensional datasets.	The performance of the classification need to improve much.Features are transformed into new ones hence loss of originality. It leads to overfitting of data.
Enhancing Ensemble on Tweet Sentiment Data	Affected	Supports high dimensionality upto 200 features..	Implemented existing base classifiers for ensemble model such as C4.5, NB,SVM ,LR.Greatly affects the Accuracy.

3. PROPOSED MODEL

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. Traditionally, as the size of the training dataset is small, medical disease prediction rate could be dramatically reduced due to class imbalance and high dimension space. In this filtering method, each attribute is tested for missing values. Most of the traditional methods are capable of extracting missing values from training data that store numeric attributes. Filtering makes models more accurate and faster. 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 and chaos gauss based randomization measure. In our model, different base classifiers such as decision tree, Naïve Bayes, FFNN 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.

Load dataset $HD^1, HD^2 .... HD^n$

For each attribute A(i) in the $HD^1, HD^2 .... HD^n$

do

For each instance value I(j) in the A(i)

do

if(isNum(I(j)) && I(j)==null)

then

$$I(j) = \sum_{j=1/i \neq j}^{n} (|I(j) - \mu_{A(I(j))}|) / (Max_{A(I(j))} - Min_{A(I(j))}) \quad \text{----(1)}$$

end if

if(isNominal(A_i) && A_i(I)==null)

then

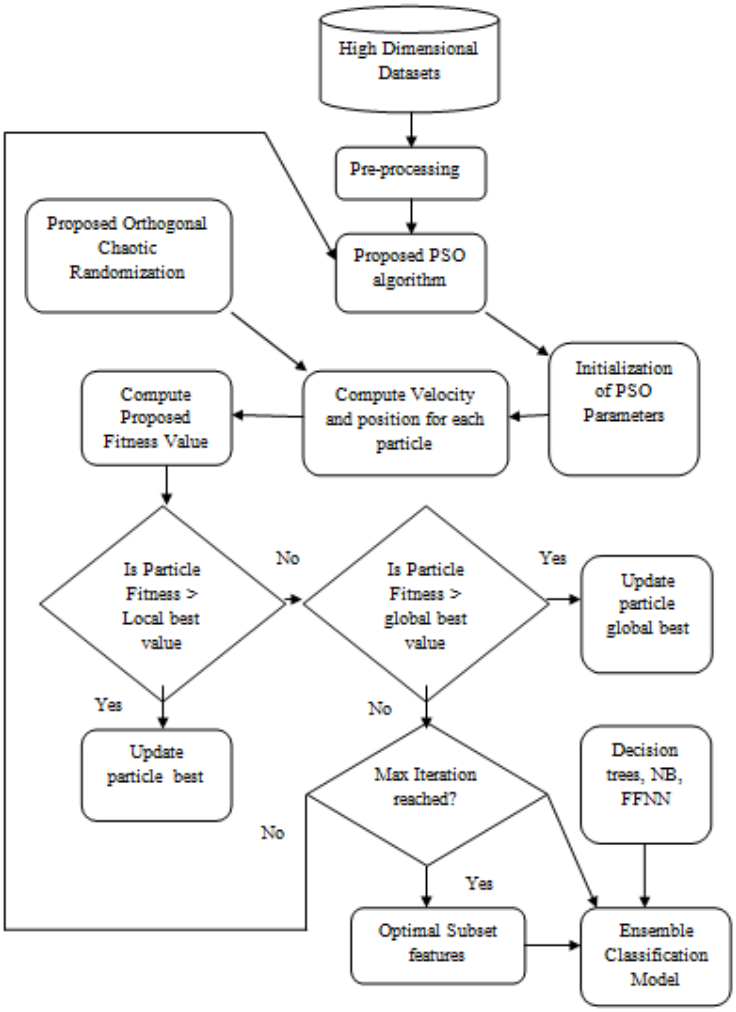

Figure 1: Proposed Model

$$I(j) = Prior\,Prob(A(i), class(m)); \quad ---(2)$$

Here, mth class of the missing value is used to find the prior probability in place of missing value

end if
End for

Step 2: // improved PSO for optimal feature selection

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

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

$$v(d+1,i) = \psi.[\omega(d,i).v(d,i).\theta_{chaos1}(pBest_i - X(d,i)) + \theta_{chaos2}(gBest_i - X(d,i))]$$

$$X(d+1,i) = X(d,i) + v(d+1,i)$$

ψ is the convergence factor computed as

$$\psi = \frac{2}{|2 - (\theta_{chaos1} + \theta_{chaos2}) - \sqrt{(\theta_{chaos1} + \theta_{chaos2})^2 - 4(\theta_{chaos1} + \theta_{chaos2})}|} \quad \text{------(3)}$$

where $\theta_{chaos1}, \theta_{chaos2} \in$ Ortho chaos gauss randomization

In this optimized model, inertia weight is computed as

$$\omega(d,i) = \omega_{max} - (I_{current} / I_{max}).(\omega_{max} - \omega_{min})$$

ω_{max} : max inertia

ω_{min} : min inertia

I_{max} : max iteration

Step 3: Computing fitness value using ortho chaotic gauss randomization measure.

In this proposed PSO model, a random value between 0 to 1 is selected using the following equation as

$$R_i = \max\{\mu.\beta_j^k(1-\beta_j^k), \frac{1}{\sqrt{2\pi\sigma_X}} e^{-\frac{(X-\mu_X)^2}{\sigma_X^2}}\} \quad \text{-------(4)}$$

$K = 1, 2 ... \text{iterations}$

$\beta_j^k \in (0,1)$

Proposed feature selection fitness measure is given as

$$Fitness_i = w_1.acc_i + w_2.(1 - \frac{\sum_{i=1}^{|F|} f_i}{N_f})$$

where $w_1, w_2 \in R_i$

f_i is the flag value 1 or 0 . '1' represents selectd feature , '0' non-selected feature.

N_f represents number of features.

acc_i : accuracy of the selected features in the ith iteration

Step 4: Apply ensemble classification model with accuracy acc_i on the selected features in the ith iteration. Here, Decision trees, NB, FFNN are used as base classifiers for high dimensional classification.

Step 5: For each particle compute its fitness value and compute classification accuracy in the previous step.

Step 6: Update particle velocity, position, global best and particle best according to the fitness value conditions as shown in figure 1.

Step 7: This process is continuous until max iteration is reached. Otherwise go to step 2.

4. EXPERIMENTAL RESULTS

In this section, we performed experiments using proposed model on well known microarray datasets and compared the statistical performance with traditional models. Dataset [15] used for experimental evaluation is listed in Table 2. Proposed model was implemented in the java environment with Netbeans IDE. From these experimental results, 10% of the training instances are used as test data for cross validation and performance analysis.

From the experimental results, it is clear that optimizing PSO in the base classifiers improves classification rate along with true positive and true negative rate. Also, the main advantage of using proposed model is to reduce the error rate on high dimensional features.

Table 2: Datasets and its characteristics

Micro array Datasets	**Attributes Size**	**Data-Type**
lung-Michigan	7000	Continuous/Numeric
lungCancer_train	12000	Continuous/Numeric
DLBCL-Stanford	4000	Continuous/Numeric

Proposed 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.

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.

Gene based cancer disease patterns using proposed feature selection based ensemble learning model on ovarian cancer dataset.

Table 3, describes the optimization of improved PSO of individual weak classifiers using the ensemble classifier. Here, the performance evaluation are performed in terms of statistical analysis, true positive rate and error rate on ovarian dataset. From the table, it is clearly observed that the proposed optimization model has high computational efficiency compared to traditional classification models.

Table 3: Performance analyses of classification accuracy on ovarian dataset on 10 cross validation.

10 % test data Ovarian Dataset			
Model	**True positive**	**Error Rate**	**Run time(ms)**
PSO+NaiveBayes	0.87	0.2492	4526
PSO+RandomForest	0.8923	0.2293	4297
PSO+Neuralnetwork	0.9055	0.1987	3975
IPSO-Ensemble	0.9575	0.1314	3297

Figure 3, describes the optimization of improved PSO of individual weak classifiers using the ensemble classifier in graphical chart. Here, the performance evaluations are performed in terms of statistical analysis, true positive rate and error rate on ovarian dataset. From the table, it is clearly observed that the proposed optimization model has high computational efficiency compared to traditional classification models.

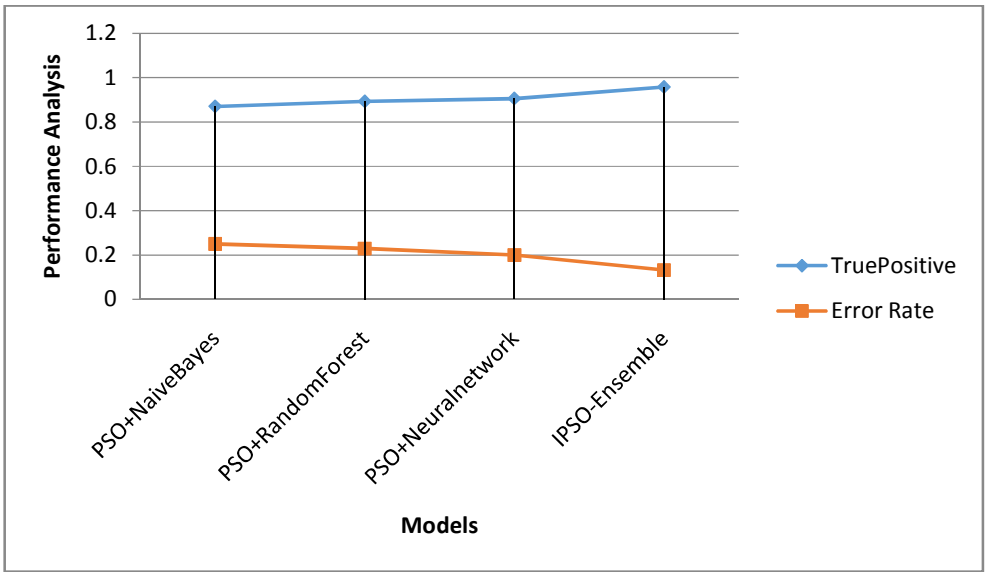

Figure 3: Performance of ovarian cancer dataset using IPSO ensemble classification model with existing models.

Table 4 and Figure 4 describe the performance of the improved PSO algorithm with ensemble classification to the traditional models on microarray cancer datasets. From the table, it is clearly observed that the proposed model has high accuracy and less error rate compare to traditional models in terms of error rate, accuracy and runtime.

Table 4: Performance analysis of proposed model to the traditional models on three microarray datasets.

Datasets	PSO+NaiveBayes	PSO+C4.5	PSO+FFNN	IPSO+Ensemble
Lung Cancer	0.697	0.8084	0.8525	0.9374
Lung Michigan	0.7935	0.8274	0.845	0.9153
Lymphoma	0.8153	0.8253	0.8574	0.9646
ErrorRate	0.364	0.304	0.294	0.225
Runtime(ms)	7351	6253	6364	5254

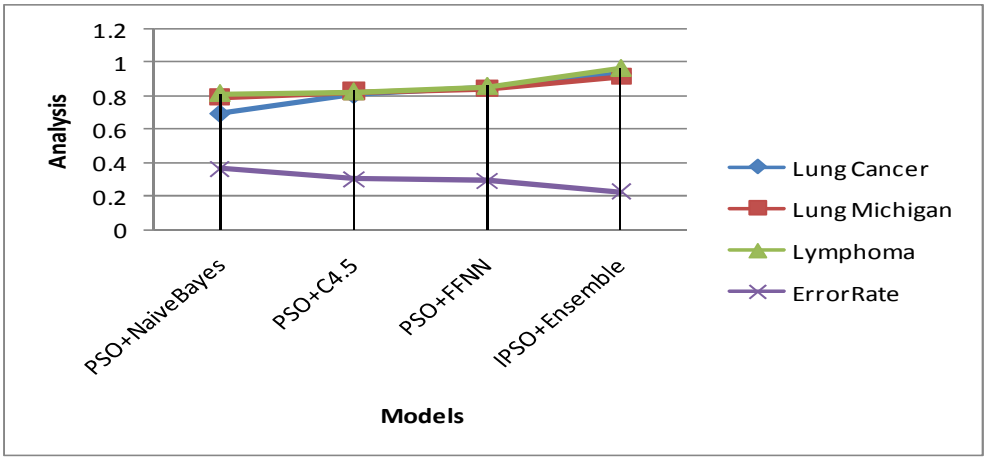

Figure 4 : Performance of proposed model to traditional models in terms of accuracy, error rate and run-time.

5. CONCLUSION

In this paper, an optimized PSO based feature selection method is integrated with the ensemble model 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. 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. Experimental results proved that the proposed model has high computational efficiency compared to the traditional feature selection based classification models in terms of accuracy, true positive rate and error rate are concerned.

REFERENCES

[1] Y. Yan, Q. Zhu, M. Shyu, and S. Chen, "A Classifier Ensemble Framework for Multimedia Big Data Classification", "IEEE 17th International Conference on Information Reuse and Integration", pp. 615-622, 2016.

[2] I. Triguero, M. Galar, S. Vluymans, C. Cornelis, H. Bustince, F. Herrera and Y. Saeys, "Evolutionary Undersampling for Imbalanced Big Data Classification ",pp.715-722, 2015.

[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] M. Popescu and J. M. Keller, "Random projections fuzzy k-nearest neighbor(RPFKNN) for big data classification ", "IEEE International Conference on Fuzzy Systems (FUZZ)",pp.1813-1817, 2016.

[5] R. Pakdel and J. Herbert, "Efficient Cloud-Based Framework for Big Data Classification", "IEEE Second International Conference on Big Data Computing Service and Applications ", pp.195-201, 2016.

[6] C. Lemnaru, M. Cuibus, A. Bona, A. Alic and R. Potolea, "A Distributed Methodology for Imbalanced Classification Problems", "11th International Symposium on Parallel and Distributed Computing ", pp. 164-171, 2012.

[7] X. R. Jenifer and R. Lawrance, "An Adaptive Classification Model for Microarray Analysis using Big Data", 2016.

[8] B. Jaison, A. Chilambuchelvan and K. A. Junaid, "Hybrid Classification Techniques for Microarray Data ","Springer Natl. Acad. Sci. Lett. ".

[9] N. C. Salvatore, A. Cerasa, I. Castiglioni, F. Gallivanone, A. Augimeri, M. Lopez, G. Arabia, M. Morelli, M. C. Gilardi and A. Quattrone, "Machine learning on brain MRI data for differential diagnosis of Parkinson's disease and Progressive Supranuclear Palsy", "Journal of Neuroscience Methods ", pp.230– 237, 2014.

[10] M. K. Shahsavari, H. R. Bakhsh and H. Rashidi, "Efficient Classification of Parkinson's Disease Using Extreme Learning Machine and Hybrid Particle Swarm Optimization", "4th International Conference on Control, Instrumentation, and Automation (ICCIA) 27-28 January 2016, Qazvin Islamic Azad University, Qazvin, Iran", pp.148-154, 2016.

[11] C. Lemnaru, M. Cuibus, A. Bona, A. Alic and R. Potolea, "A Distributed Methodology for Imbalanced Classification Problems", "11th International Symposium on Parallel and Distributed Computing ", pp. 164-171, 2012.

[12] X. R. Jenifer and R. Lawrance, "An Adaptive Classification Model for Microarray Analysis using Big Data", 2016.

[13] B. Jaison, A. Chilambuchelvan and K. A. Junaid, "Hybrid Classification Techniques for Microarray Data ","Springer Natl. Acad. Sci. Lett. ".

[14] A. Govada, V. S. Thomas, I. Samal and S. K. Sahay, "Distributed multi-class rule based classification using RIPPER", "IEEE International Conference on Computer and Information Technology", pp.303-309, 2016.

[15] http://datam.i2r.astar.edu.sg/datasets/krbd/index.html

AUTHORS

Lalitha Kumari Gaddala, B.Tech,M.Tech,has a work experience of 12 years. She is pursuing her Ph.D in Computer Science and Engineering under Acharya Nagarjuna University, Guntur and Andhra Pradesh. She is working as Senior Assistant Professor in Prasad V Potluri Siddhartha Institute of Technology, Vijayawada. Her research interests are Soft Computing-Optimization algorithms, Machine Learning. She has published few papers in International conferences and International Journals.

Dr. N. Naga MalleswaraRao,B.Tech, M.Tech and Ph.D, has a work experience of 25 years and 9 Months. He is working as Professor, Department of IT, RVR& JC College of Engineering, Guntur (Dt). His research interests are Computer Algorithms, Compilers, and Image Processing. He has published few papers in International conferences, National Conferences and International Journals.