IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 14, Issue 5 (Sep. - Oct. 2013), PP 104-111
www.iosrjournals.org

Efficient Disease Classifier Using Data Mining Techniques: Refinement of Random Forest Termination Criteria

K. Kalaiselvi[1], K.Sangeetha[2], S.Mogana[3]
[1]*Department of Computer Science and Engineering, SNS College of Engineering,Tamilnadu, India.*
[2]*Department of Information Technology, SNS College of Technology, Tamilnadu, India.*
[3]*Department of Computer Science and Engineering, SNS College of Technology, Tamilnadu, India.*

Abstract: *In biomedical field, the classification of disease using data mining is the critical task. The prediction accuracy plays a vital role in disease data set. More data mining classification algorithms like decision trees, neural networks, Bayesian classifiers are used to diagnosis the diseases. In decision tree Random Forest, Initially a forest is constructed from ten tress. The accuracy is measured and compared with desired accuracy. If the selected best split of trees matched the desired accuracy the construction terminates. Otherwise a new tree is added with random forest and accuracy is measured. The fitting criteria of random forest are accuracy and correlation. The accuracy is based on the mean absolute percentage error (MAPE) and the mean absolute relative error (MARE).In proposed system to refine the termination criteria of Random Forest, Binomial distribution, multinomial distribution and sequential probability ratio test (SPRT) are used. The proposed method stops the random forest earlier compared with existing Random Forest algorithm. The supervised learning model like support vector machine takes a set of inputs and analyze the inputs and recognize the desired patterns. The disease data sets are supplied to SVM and prediction accuracy is measured. The comparison is made between Random Forest and SVM and best class labels are identified based on disease.*

Keywords: *SVM, Random Forest, Disease classification.*

I. Introduction

Initially machine learning (ML) systems were developed to analyze the medical data sets. The knowledge of the medical diagnosis is derived from the past history. The derived classifier can be used to diagnosis the new datasets with more reliability, speed and accuracy. The ML system is more useful to solve medical diagnosis problems because of its good performance, the ability to deal with missing data, the ability to explain the decision and transparency of knowledge [1].

In decision tree algorithm of Random Forest, the tree is constructed dynamically with online fitting procedure. A random forest is a substantial modification of bagging [3] – [6]. The generation of trees is based on two steps. First the tree is constructed on a bootstrap replicate of original dataset and second a random feature subset, of fixed predefined size, is considered for splitting the node of the tree. To select a best split Gini Index is used. In ensemble classifier like random forest the size of the ensemble depends on 1) the desired accuracy, 2) the computational cost, 3) the nature of the classification problem, and 4) the number of available processors. In existing methods the size of the ensemble is determined by one of the three ways. 1) the method that preselect the ensemble size, 2) the method that post select the ensemble size ,3) methods that select the ensemble size during training [17]. In pre selection method, the size of the ensemble is determined by the user. The second type of post selection method, over – produce and choose strategy is used to select the ensemble from the pool of classifier.

The method which selects the size of the ensemble in training phase is determined dynamically. Initially the Random forest is constructed from the bootstrap replicate and in every step, the new classifier is considered for the ensemble selection. If its contribution to the ensemble is significant then the classifier is retained. From Banfield *et al.* [12] method, it decides the ensemble, when a sufficient number of classification trees in random forest have been created. The method smoothes the out-of-bag error graph by using a sliding window of size five. After smoothing has been completed, the method examines windows of size 20 and determines the maximum accuracy within that window. It continues processing windows of the same size until the maximum accuracy within that window no longer increases. At this point, the stopping criterion has been reached and the algorithm returns the ensemble with the maximum accuracy from within that window.

The proposed method, the construction of tree based on classical Random Forest, Random forests with ReliefF, random forests with multiple estimators, RK Random Forests, and RK Random Forests with multiple estimators [2]. Random Forest with ReliefF evaluates partitioning power of attributes according to how well

their values distinguish between similar instances. An attribute is given a high score if its values separate similar observations with different class and do not separate similar instances with the same class values. ReliefF samples the instance space, computes the differences between predictions and values of the attributes and forms a statistical measure for the proximity of the probability densities of the attribute and the class. Its quality estimates can be explained as the proportion of the explained class values. Assigned quality evaluations are in the range [¡1;1]. The computational complexity for evaluation of *an* attributes is *O* (*m¢ n ¢ a*), where *m* is the number of iterations [8].

In RK –Random Forest the number K of features randomly selected at each node during the tree induction process. The new Forest-RK decision tree induction procedure can be summarized as below:
1) Let N be the size of the original training set. N instances are randomly drawn with replacement, to form the bootstrap sample, which is then used to build a tree. 2) Let M be the dimensionality of the original feature space. Randomly set a number K 2 [1; M] for each node of the tree, so that a subset of K features is randomly drawn without replacement, among which the best split is then selected. 3) The tree is thus built to reach its maximum size. No pruning is performed [18].

Support Vector Machines are based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships. Intuitively, a good separation is achieved by the boundaries that have the largest distance to the nearest training data point of any class called functional margin, since in general the larger the margin the lower the generalization error of the classifier.

II. Random Forest

Random Forest is a collection of decision trees. From the training data the Random forest is constructed. In each step the tree is constructed with other data which has been selected as a best split. The forest is constructed without pruning. Forest construction is based on three step process [2]. 1) Forest construction, 2) the polynomial fitting procedure, and 3) the termination criteria.

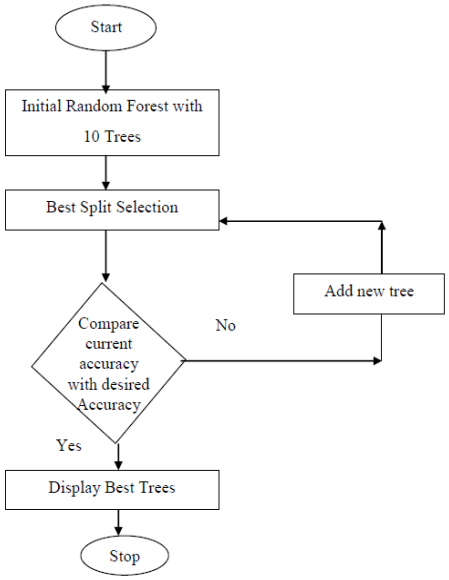


Fig 1.Random Forest construction method

a) Forest Construction: Initially a forest is constructed from ten trees. For that classical random forest is combined with performance evaluation criteria's like Relief and multiple estimators. More specifically random forests with ReliefF, random forests with multiple estimators, RK Random Forests, and RK Random Forests with multiple estimators along with Classical Random Forest are constructed.

The forest construction is shown below. Initially, the forest started with ten trees and select a best fit is selected from the remaining dataset and the construction is made. The same procedure is repeated up to the 100 tree [2].

b) Polynomial fitting procedure: Forest construction is an iterative process .Each time a new dataset is selected for the construction. The selection based on the accuracy of the predicted ensemble. The following polynomial equation is applied for selecting best fit [2].

$$fn-1\ (x) = pn\ xn + pn-1\ xn-1 + \cdot \quad + p0,\ n= 2,\ 9. \qquad (1)$$

c) The termination criteria: In the termination of the forest accuracy, correlation and the combination accuracy and correlation is used. The criterion accuracy is based on the consecutive fitted curve. In correlation, the comparison is made between the fitted curve and original. The polynomial of two to eight would be applied to select a best one. In the third criterion, the accuracy and correlation are combined to select a best curve.

III. Support Vector Machine

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

IV. Proposed Method

In the proposed method binomial distribution and multinomial distribution and sequential probability ratio test are used to determine the best case for tree construction. The binomial distribution is the discrete probability distribution of the number of success in a sequence of n independent experiments. From the definition, with the random variable X follows the binomial distribution with parameters n and p, we write $X \sim B(n, p)$. The probability of getting exactly k successes in n trials is given by the probability mass function:

$$f(k;n,p) = \Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

In probability theory, the multinomial distribution is a generalization of the binomial distribution. For *n* independent trials each of which leads to a success for exactly one of *k* categories, with each category having a given fixed success probability, the multinomial distribution gives the probability of any particular combination of numbers of successes for the various categories. The sequential probability ratio test (SPRT) is a specific sequential hypothesis test.

In general, the binomial, multinomial and SPRT are used to find the optimal result with given test cases. In proposed all these methods are applied with Random Forest algorithm and its various modifications before the construction of random forest. This will produce best split of test case selection from the data sets. It will reduce the time for testing each new test case from the disease data sets. In existing each test cases are tested for new tree construction. Proposed method avoids the time delay. It will produce ensemble with more accuracy.

V. Results And Discussion

The Random Forest algorithm and modifications are verified with five different datasets PIMA Indians, SPECT, BCW, BT, and Ecoli. The classification is based on disease dataset. Some classifications are two class classifications. In case of Breast Cancer the classification is for the person is affected or not. Where as in Ecoli the classification is based on survival of patient based on surgery. The experiment was conducted to measure the performance. Random Forest is constructed up to 100 trees. The best one is selected with more accuracy.

Initially proposed method was evaluated with five disease datasets with binomial distribution. In existing, the termination criteria were based on accuracy and correlation. It was produced best result with combined criteria of accuracy and correlation. The termination of trees based on accuracy. It falls in the range of 100 trees. To refine the termination criteria binomial distribution is applied with Random Forest algorithm. The binomial distribution is frequently used to model the number of successes in a sample of size n drawn with replacement from a population of size N. If the sampling is carried out without replacement, the draws are not independent and so the resulting distribution is a hyper geometric distribution, not a binomial one. However, for N much larger than n, the binomial distribution is a good approximation, and widely used.

The classical RF used 68 trees to find the optimal solution whereas the classical RF with binomial distribution used 49 to 55 in MARE and MAPE to terminate. Like wise multinomial distribution also used with Random Forest algorithms. Multinomial distribution is a generalization of the binomial distribution. For n independent trials each of which leads to a success for exactly one of k categories, with each category having a given fixed success probability, the multinomial distribution gives the probability of any particular combination of numbers of successes for the various categories. Here also 54 and 55 no of trees are constructed to find the optimal solution of disease data set.

Finally, the sequential probability ratio test along with classical Random Forest is used. SPRT is a specific sequential hypothesis test offers a rule of thumb for when all the data is collected. While originally developed for use in quality control studies in the realm of manufacturing, SPRT has been formulated for use in the computerized testing of human examinees as a termination criterion. Here 67 and 59 no of trees are constructed for MARE and MAPE. It exceeds the classical Random Forest method.

The comparison is made between the existing method and the proposed one.

TABLE I

Datasets	Method				
	Classical RF	RF with ReliefF	RF with ME	RK- RF	RK - RF with me
BCW	68/68	57/57	85/88	100/100	100/100
ECOLI	100/100	100/100	80/82	62/97	97/97
SPECT	65/61	100/100	99/99	60/59	92/92
PIMA	100/52	73/73	80/80	67/74	91/92
BT	67/74	62/59	90/91	48/81	57/100

*mare/mape

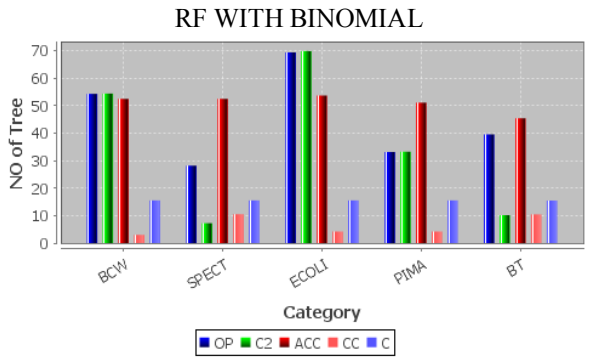
The above mentioned results are taken from existing system. Total no of trees generated by MARE and MAPE are given. Like wise the results are taken for the binomial RF and multinomial distribution along with SPRT. Same datasets are classified by support vector machine also. The results of binomial and multinomial are compared with classical RF. The following Table II shows the result of proposed system.

TABLE II

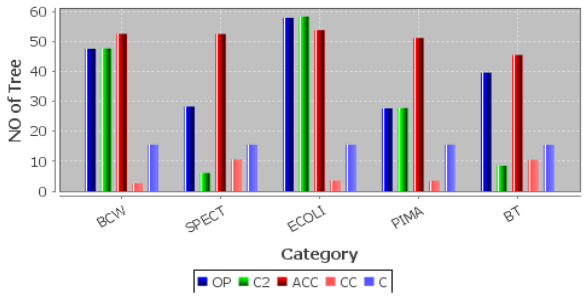
Datasets	Method			
	Classical RF	RF with Binomial	RF with Multinomial	RF with SPRT
BCW	68/68	49/55	55/54	67/59
ECOLI	100/100	50/50	65/63	58/63
SPECT	65/61	60/58	53/50	65/68
PIMA	100/52	56/60	60/55	53/56
BT	67/74	70/73	78/78	70/66

*mare/mape

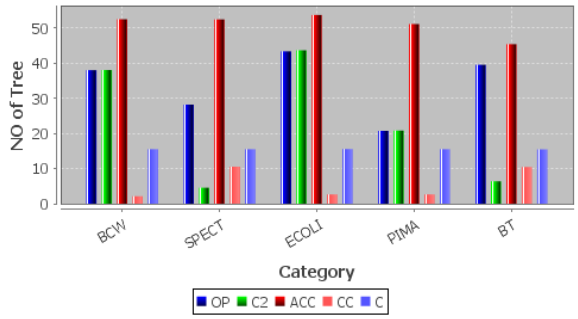
Fig 1.2 shows the result of classical RF with binomial along with RK Random Forest, RF with me and finally RK-RF with me. Like wise multinomial distribution method was combined with Random Forest algorithms. Third, SPRT method was utilized with Random Forest algorithms.



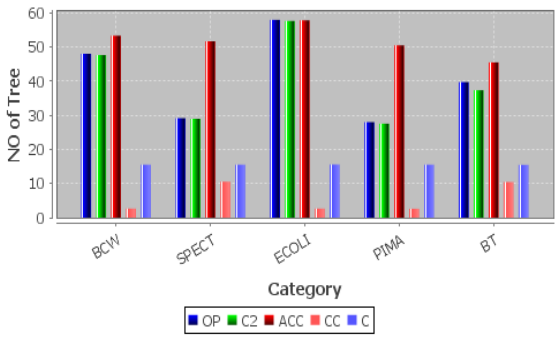
RF WITH MULTINOMIAL



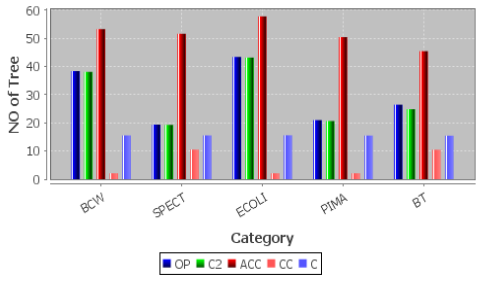
RF WITH SPRT



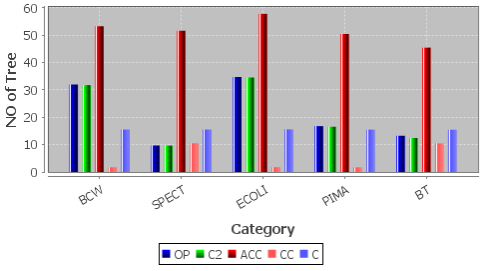
RF_ME WITH BINOMIAL



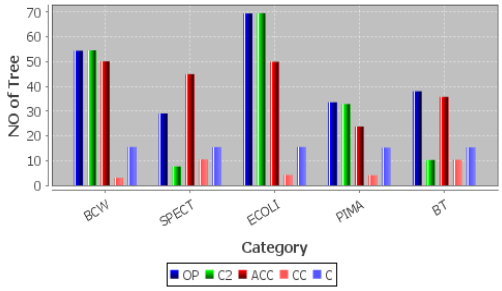
RF_ME WITH MULTINOMIAL



RF_ME WITH SPRT

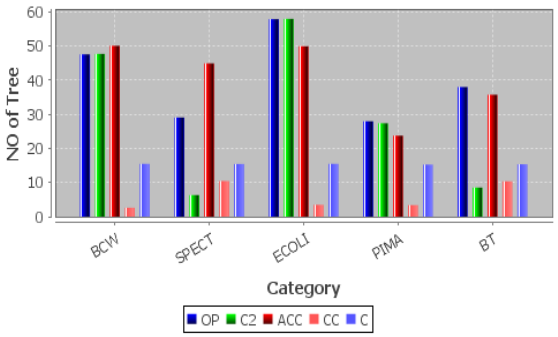


RK RF WITH BINOMIAL

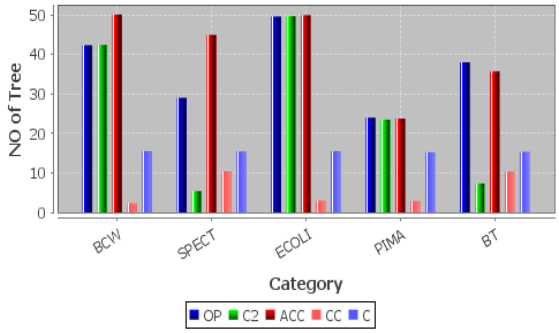


The results are generated with accuracy, correlation and optimal no of trees needed to find the optimal result in disease dataset. In Each case five disease datasets were tested with four different algorithms of Random Forest algorithm. Like wise the datasets are classified with support vector machine. The proposed method refines the termination criteria of existing with probability distributions.

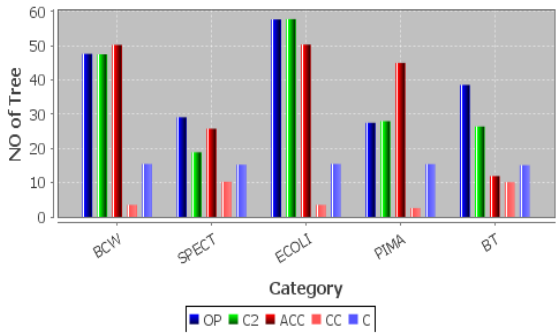
RK RF WITH MULTINOMIAL



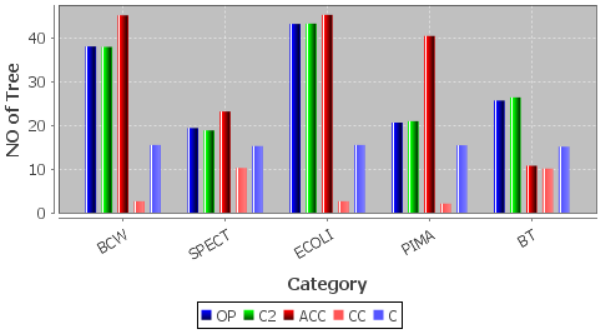
RK RF WITH SPRT



RK RF_ME WITH BINOMIAL



RK RF_ME WITH MULTINOMIAL



RK RF_ME WITH SPRT

In some cases the SPRT may produce similar results of classical Random Forest. Otherwise the binomial and multinomial are produced better result compared with classical RF. In Breast Tissue dataset the probability distributions took more no of trees to find the solution.

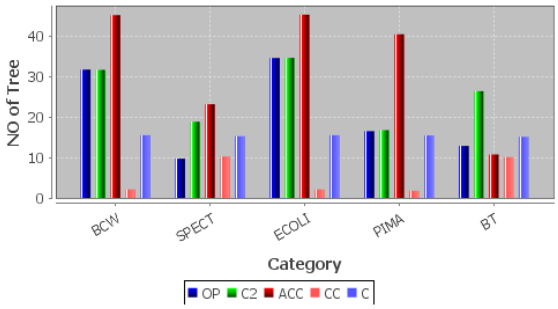


Fig 1.2 Random Forest with binomial distribution, multinomial distribution and SPRT

VI. Conclusion

Random Forest algorithm is used to predict the disease with good performance with the termination criteria of accuracy and correlation along with binomial and multinomial distribution. Probability distributions produced better result compared with classical RF along with the termination criteria of accuracy and correlation. In some cases distributions are also produced the same result as like Random Forest. In support vector machine recursively the test cases are tested to predict the best ensemble. The comparison will be made to predict the best accuracy. According to the accuracy it will produce the better result. The proposed system finds 36 out of 40 cases with better termination criteria. In future the same method may also be applied to other diseases to predict the disease with high performance.

References

[1] I. Kononenko, "Machine learning for medical diagnosis: History, state of the art and perspective," *Artif. Intell. Med.*, vol. 23, no. 1, pp. 89–109, 2001.

[2] Evanthia E. Tripoliti et al, "Automated Diagnosis of Diseases Based on Classification: Dynamic Determination of the Number of Trees in Random Forests Algorithm" IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE, VOL. 16, NO. 4, JULY 2012.

[3] L. Breiman, "Bagging predictors," *Mach. Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[4] Y. Freud and R. E. Schapire, "A decision-theoretic generalization of online learning and an application to boosting," *J.Comput. Syst. Sci.*, vol. 55,no. 1, pp. 119–139, 1997.

[5] T.K.Ho, "The random subspace method for constructing decision forests," *IEEE Trans. Pattern Anal.Mach. Intell.*, vol. 20, no. 8, pp. 832–844, 1998.

[6] L. Breiman, "Random forests," *Mach. Learning*, vol. 45, pp. 5–32, 2001.

[7] R. E. Banfield, L. O. Hall, K. W. Bowyer, and W. P. Kegelmeyer, "A comparison of decision tree ensemble creation techniques," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 173–180, Jan. 2007.

[8] Marko Robnik-˘ Sikonja and Igor Kononenko. "Theoretical and empirical analysis of ReliefF and RReliefF". *Machine Learning Journal*, 53:23–69, 2003.

[9] D. Byrne et al , "Breast cancer detection based on differential ultrawideband microwave radar " , *Progress In Electromagnetics Research M, Vol. 20, 231{242, 2011.*

[10] Hagness, S. C., A. Ta°ove, and J. E. Bridges, Two-dimensional FDTD analysis of a pulsed microwave confocal system for breast cancer detection: Fixed focus and antenna array sensors," *IEEE Transactions on Biomedical Engineering*, Vol. 45, 1470{1479, 1998.

[11] E. E. Tripoliti, D. I. Fotiadis, and G. Manis, "Modifications of random forests algorithm," Data Knowl. Eng., to be published.

[12] C. Orrite, M. Rodriquez, F. Martinez, and M. Fairhurst, "Classifier ensemble generation for the majority vote rule," in Lecture Notes on ComputerScience, J. Ruiz-Shulcloper et al., Eds. Berlin\Heidelberg, Germany: Springer-Verlag, pp. 340–347, 2008.

[13] P. Letinne, O. Bebeir, and C. Decaestecker, "Limiting the number of trees in random forests," in Lecture Notes on Computer Science. Berlin\Heidelberg, Germany: Springer-Verlag, 2001, pp. 178–187.
[14] S. Gunter and H. Bunke, "Optimization of weights in a multiple classifier handwritten word recognition system using a genetic algorithm," Electron. Letters Comput. Vision Image Anal., pp. 25–41, 2004.
[15] S. Bernard, L. Heutte, and S. Adam, "On the selection of decision trees in random forests," in Proc. IEEE-ENNS Int. Joint Conf. Neural Netw., 2009, pp. 302–307.
[16] G. D. Magoulas and A. Prentza, "Machine learning in medical applications," *Mach. Learning Appl. (Lecture Notes Comput. Sci.)*, Berlin/Heidelberg, Germany: Springer, vol. 2049, pp. 300–307, 2001.
[17] L. Rokach and O. Maimon, "Data Mining with Decision Trees Theory and Applications" (Machine Perception and Artificial Intelligence Series 69). H. Bunke and P. S. P. Wang, Eds. Singapore: World Scientific, 2008.
[18] S. Bernard, L. Heutte, and S. Adam, "Forest-RK: A new random forest induction method," in *Proc. Int. Conf. Intell. Comput. 2008*. Lecture Notes in Artificial Intelligence 5227, D.-S. Huang, *et al.*, Eds. Heidelberg, Germany: Springer, 2008a, pp. 430–437.
[19] G. Martinez-Munoz and A. Suarez, "Pruning in ordered bagging ensembles," in Proc. 23rd Int. Conf. Mach. Learning, 2006, pp. 609–616.
[20] J. Xiao and Ch. He, "Dynamic classifier ensemble selection based on GMDH," in Proc. Int. Joint Conf. Comput. Sci. Optimization, 2009, pp. 731–734.

AUTHORS PROFILE

K.Kalaiselvi received the Bachelors degree in Computer Science and Engineering from Anna University, Chennai in 2005 and the Master degree in Computer Science And Engineering from Anna University Chennai in 2013. Her research interest is Data Mining. Currently she is working as an Assistant Professor in SNS College of Engineering,Coimbatore.

K.Sangeetha received the Bachelors degree in Computer Science and Engineering from Anna University, Chennai in 2008 and the Master degree in Network Engineering from Anna University of Technology Coimbatore in 2011. Her research interests are Data Mining and Wireless Sensor Networks. Currently she is working as an Assistant Professor in SNS College of Technology, Coimbatore.

S.Mogana currently doing B.E Computer Science and Engineering in SNS College of Technology. Her research interest is Data Mining.