ACUMULATED RANDOM FOREST CLASSIFIER WITH GENETIC ALGORITHM FOR DIABETES MELLITUS

Mr. Shaik Mastan [1]
3rd Year Student,
Department of Computer Science,
SV U CM & CS, Tirupati.

Dr. E. Kesavulu Reddy[2],
Assistant Professor,
Department of Computer Science,
SV U CM & CS,, Tirupati.

Abstract: Diabetes mellitus (DM) is defined as a group of metabolic disorders exerting significant pressure on human health worldwide. Various computerized information systems were outlined utilizing diverse classifiers for anticipating and diagnosing diabetes mellitus. Machine learning based classification algorithms helps in diagnosing the symptoms at early stages by prior diagnosing of symptoms and taking according to medications to it. Combining of genetic algorithm with the random forest classifier can optimize the results obtained only by the random forest classifier. In this proposed system, genetically optimized random forest classifier is used for the classification of Diabetes Mellitus. Here a Genetic algorithm is used in the first stage for optimizing random forest, and the optimized outputs are fed into the fine-grained random forest to diagnose the symptoms of Diabetes Mellitus. In this analysis, the proposal of hybrid optimized random forest classifier (GA-ORF) with a genetic algorithm is made. In this evaluation was on the various performance metrics of classifiers, GA-ORF has achieved accuracy higher than of the previously proposed classifiers for diabetes mellitus.

Keywords: *Classifier, Genetic Algorithm, Performance Metrics.*

I. INTRODUCTION:

Class decomposition is defined as the process of breaking down the huge dataset into a number of subsets by applying clustering to the attributes present in the dataset that belong from time to time. Let D={D_1, D_2, D_3,…..D_n) be a dataset which contains records, that contain various attribute values. The dataset is decomposed based upon the condition 'C' by clustering; the dataset D is decomposed as D_c, containing multiple classes of attribute values. Class decomposition [1] can be seen back in 2003 where the decomposition takes by clustering technique employed in classes of attributes. In order to apply the clustering to a medical database, first the data has to be preprocessed for supervised learning, it takes two stages for data preprocessing. First stage [2] applies to only the positive classes of the datasets and the second stage [3] comprises of generalizing the decomposition to negative and positive classes respectively. Even the diversification of the dataset came from the processes, can increase the performance further, however, class decomposition must have proper parameters set. The random forest comes with its own parameter setting such as a number of trees and number of features. Realizing that the parameter setting of the random forest has a greater influence on optimizing the random forest is an optimization problem. Genetic Algorithm is superior to the random forest classifiers containing parameter like local optima. The motive of the genetic algorithm is to detect the original subclasses of the random forest classifier. Ensemble on the subclasses of the random forest classifier provides an optimal and separability of classes.

II. RELATED WORK:

Random forest is considered as a superior of all classifiers, considered as some state-of-the art enable methods [4] [5], various comparisons has identified that random forest classifier is superlative [6] including gradient boosting trees. Random forest adopts models such as data replicas and bootstrap sampling called bagging [7]. The random forest has two main parameters such as a number of trees and number of features as discussed earlier. By default, the number of trees is set between 100 to 500 and number of features as log2 (n). Random forest extension has been proposed in [8]. Various problems of random forest haven been identified in [9] such as over sampling, under sampling and sensitivity. Most recently random forest has vision on machine learning for classification tasks [10] [11] [12]. Genetic algorithms envisioned in recent times for hard optimization problems [13] [14] [15]. It starts with providing the solution for each individual by a chromosome, and then each chromosome is evaluated on the fitness. The fitness is used to make the chromosome to survive in the entire population. Two basic adoptions are applied, crossover and mutation are used to generate the randomness in the solution area. Many varieties of these are proposed in [16]. Class decomposition was first

introduced in [1] with high bias and less variance. The goal of the clustering process is not only clustering but also the cluster separation, the class decomposition is applied to a medical diagnosis in [2], the clustering is done by the separation of positive and negative classes respectively. Random forest's high bias classifiers perform well when the clusters are remerged [1], class decomposition method using the very fast neural network is presented in [17]. A genetic algorithm has been widely used in optimizing random forest classifier in [18] where each chromosome is considered as a variety of trees and also the variable length chromosome in the solution space. Recently [19] have a thoroughly experimented a number of support vector machines and concluded that LPSVM is superior in diagnoses.

III. PROPOSED SYSTEM

The main objective of this system is to optimize the inputs of Random forest classifier to produce an efficient result by genetic algorithm. The first step involves in handling missing values and normalizing the values in accordance with [20]. The detailed scenario is depicted in Fig 1. The workflow starts from the preprocessed data, where the data are subjected to be normalized, the preprocessed dataset is divided into a training set and a test/ validation set, the splitting is based upon the percentage split. The training set is given to the genetic algorithm where it will decompose into two classes such as positive and negative classes, the important parameters of random forest is Kvalues, ntrees and mtry, the random forest classifier produces number of trees according to the experimenter, these ntrees and mtry values are decomposed by class decomposition, by the way, the test/validation set is given to the fit random forest classifier and finally the optimized random forest is obtained, the obtained classifier is used in the prediction of Diabetes Mellitus.

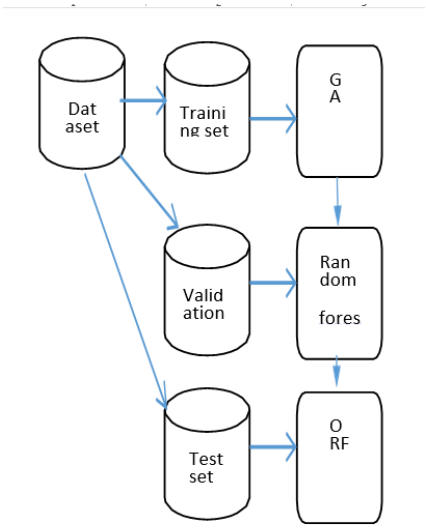

Fig. Optimizing RF framework

IV MULTI CLASS DECOMPOSITION ALGORITHM

Supervised multiclass classification algorithms aim at assigning a class label for each input example. Given a training data set of the form (x_i, y_i), where x_i Rn is the ith example and y_i 1, . . ., K is the ith class label, we aim at finding a learning model H such that $H(x_i) = y_i$ for new unseen examples. The problem is simply formulated in the two-class case, where the labels y_i are just +1 or -1 for the two classes involved.

Several algorithms have been proposed to solve this problem in the two-class case, some of which can be naturally extended to the multiclass case, and some that need special formulations to be able to solve the latter case. The first category of algorithms include decision trees [5, 16], neural networks [3], *k*-Nearest Neighbor [2], Naive Bayes classifiers [19], and Support Vector Machines [8]. The second category include approaches for converting the multiclass classification problem into a set of binary classification problems that are efficiently solved using binary classifiers e.g. Support Vector Machines [8, 6]. Another approach tries to pose a hierarchy on the output space, the available classes, and performs a series of tests to detect the class label of new patterns.

EXTENSIBLE ALGORITHMS

The multiclass classification problem can be solved by naturally extending the binary classification technique for some algorithms. These include neural net- works, decision trees, *k*-Nearest Neighbor, Naive Bayes, and Support Vector Machines.

Table 1: One-per-class Coding

Class 1	1000
Class 2	0100
Class 3	0010
Class 4	0001

Table 2: Distributed coding

Class 1	00000
Class 2	00111
Class 3	11001
Class 4	11110

DECOMPOSING INTO BINARY CLASSIFICATION

The multiclass classification problem can be decomposed into several binary classification tasks that can be solved

efficiently using binary classifiers. The most successful and widely used binary classifiers are the Support Vector Machines [8, 6]. The idea is similar to that of using codewords for each class and then using a number binary classifier in solving several binary classification problems, whose results can determine the class label for new data. Several methods have been proposed for such a decomposition [1, 10, 11, 13].

ONE-VERSUS-ALL (OVA)

The simplest approach is to reduce the problem of classifying among K classes into K binary problems, where each problem discriminates a given class from the other K 1 classes [18]. For this approach, we requīre $N = K$ binary classifiers, where the k^{th} classifier is trained with positive examples belonging to class k and negative examples belonging to the other classes.

When testing an unknown example, the classifier producing the maximum output is considered the winner, and this class label is assigned to that example. Rifkin and Klautau [18] state that this approach, although simple, provides performance that is comparable to other more complicated approaches when the binary classifier is tuned well.

ALL-VERSUS-ALL (AVA)

In this approach, each class is compared to each other class [11, 12]. A binary classifier is built to discriminate between each pair of classes, while discarding the rest of the classes. This requires building $\frac{K(K-1)}{}$ binary classifiers. When testing a new example, a voting is performed among the classifiers and the class with the maximum number of votes wins. Results [1, 13] show that this approach is in general better than the one-versus-all approach.

Table : ECOC example

	f_1	f_2	f_3	f_4	f_5	f_6	f_7
Class 1	0	0	0	0	0	0	0
Class 2	0	1	1	0	0	1	1
Class 3	0	1	1	1	1	0	0
Class 4	1	0	1	1	0	1	0
Class 5	1	1	0	1	0	0	1

V. CONCLUSION:

An optimized random forest classifier with the genetic algorithm is proposed in this paper. The main idea of the medical diagnosis is to extract the valuable information from the symptoms and providing an appropriate medication in a lesser amount of time. The proposed system is compared with the existing hybrid classifiers and the results are obtained. The proposed GA-ORF classifier outperformed with an accuracy of 0.923, specificity of 0.924, the sensitivity of 0.901, and kappa statistics of 0.879, which are higher than the existing classifier approaches for Diabetes Mellitus. The future investigation can take done in combining other classifier algorithms with hybrid genetic algorithms for greater accuracy.

VI REFERENCES

[1] R. Vilalta, M. K. Achari, C. F. Eick, Class decomposition via clustering: a new framework for low- variance classifiers, in Data Mining, 2003. ICD 2003. Third IEEE International Conference on, IEEE, 2003, pp. 673–676.

[2] Polaka, Clustering algorithm specifics in class decomposition, in: Applied Information and Communication Technology, 2013, Proceedings of the 6th International Scientific Conference, 2013, pp. 29–36.

[3] E. Elyan, M. M. Gaber, A fine-grained random forests using class decomposition: an application to medical diagnosis, Neural Computing and Applications (2015) 1–10.

[4] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.

[5] D. R. Cutler, T. C. Edwards Jr, K. H. Beard, A. Cutler, K. T. Hess, J. Gibson, J. J. Lawler, Random forests for classification in ecology, Ecology 88 (11) (2007) 2783–2792

[6] M. Ferna´ndez-Delgado, E. Cernadas, S. Barro, D. Amorim, Do we need hundreds of classifiers to solve real world classification problems?, Journal of Machine Learning Research 15 (2014) 3133–3181.

[7] L. Breiman, Bagging predictors, Machine learning 24 (2) (1996) 123–140.

[8] K. Fawagreh, M. M. Gaber, E. Elyan, Random forests: from early developments to recent advancements, Systems Science & Control Engineering: An Open Access Journal 2 (1) (2014) 602– 609

[9] S. del Ro, V. Lpez, J. M. Bentez, F. Herrera, On the use of mapreduce for imbalanced big data using random forest, Information Sciences 285 (2014) 112 – 137, processing and Mining Complex Data Streams.

[9] X. Liu, M. Song, D. Tao, Z. Liu, L. Zhang, C. Chen, J. Bu, Random forest construction with robust semi supervised node splitting, IEEE Transactions on Image Processing 24 (1) (2015) 471– 483.

[10] M. Ristin, M. Guillaumin, J. Gall, L. V. Gool, Incremental learning random forests for large- scale image classification, IEEE Transactions on Pattern Analysis and Machine Intelligence 38 (3) (2016) 490–503

[11] T. Li, B. Ni, X. Wu, Q. Gao, Q. Li, D. Sun, On random hyper-class random forest for visual classification, Neurocomputing (2016) 281 – 289.

[12] I. Boussa¨ıD, J. Lepagnot, P. Siarry, A survey on optimization metaheuristics, Information Sciences 237 (2013) 82–117.

[13] A. E. Eiben, J. E. Smith, Introduction to evolutionary computing, Springer Science & Business Media, 2003.

[14] D. Whitley, A genetic algorithm tutorial, Statistics and computing 4 (2) (1994) 65–85.

[15] L. D. Davis, K. De Jong, M. D. Vose, L. D. Whitley, Evolutionary algorithms, vol. 111, Springer Science & Business Media, 2012.

[16] S. Jaiyen, C. Lursinsap, S. Phimoltares, A very fast neural learning for classification using only new incoming datum, Neural Networks, IEEE Transactions on 21 (3) (2010) 381–392

[17] T. Azar, H. I. Elshazly, A. E. Hassanien, A. M. Elkorany, A random forest classifier for lymph diseases, Computer methods and programs in biomedicine 113 (2) (2014) 465-473

[18] A. T. Azar, S. A. El-Said, Performance analysis of support vector machines classifiers in breast cancer mammography recognition, Neural Computing and Applications 24 (5) (2014) 1163–1177.

[19] D. J. Stekhoven, missForest: Nonparametric Missing Value Imputation using Random Forest, r package version 1.4 (2013).