Disease Prediction Using Decision Tree Based Weighed Voting System in Data Mining

[1]S. Ramasamy, [2]Dr.K.Nirmala
[1]Research scholar, Quaid-E-MillathGovt college for women Chennai, India
[2]Research supervisor Department of computer science
Quaid-E-MillathGovt college for women Chennai, India

Abstract: The accessibility of the enormous size of medical dataset hints towards the requirement of a tool which analyses data to extract valuable information. Data scientists have attempted several methods in order to improvise the examination of large data sets. Previously, various data mining techniques have been implemented in the healthcare systems. The proposed approach can adjust each classifier's weight based on their ability and history of making correct predictions. A rule that mixes majority voting and weights of classifiers was proposed and applied for the final diagnosis decision.

Keywords: classification, disease prediction, voting system, accuracy.

1. Introduction

Classification in medical diagnostics is able to aid in disease diagnosis and predicts outcomes in response to the treatment. Many efforts have been made to improve the classification performance. For instance, in traditional methodology for classification, logistic regression–based trichotomous classification tree was applied in diagnosing breast cancers [1]. Non-parametric empirical bayes algorithm was developed for integrative genetic risk prediction of complex diseases with binary phenotypes [2]. Hierarchical support vector machine-based algorithm was employed in the EEG-based motor imagery classification task [3]. Bionic algorithms were also introduced in the classification of medical data. Self-adaptive niche genetic algorithm with random forest was proposed to build model for sepsis patient's stratification [4], Classification rules were extracted by ant-miner algorithm and thereby applied in diagnosing heart disease etc..,

However, in the practice of medical classification, data are usually class-imbalanced [5], which means the distributions of classes are not uniform [6]. In binary-classification cases, the class with larger distribution is named as the majority while the other is named as the minority [7]. Dealing with the class-imbalanced data, conventional algorithms are prone to consider tend to minority observation as noise or outliers and ignore them in the classifying [8], thereby tend to classify samples into the majority class. Consequently, the predictive accuracy for the minority class will be much lower than that for the majority class[9].

To diagnose particular disease, a physician has to explore patient's data and consider many factors (e.g. family history, age, body mass index, etc.). A physician's diagnosis can be subjective and is highly dependent on the experiences. Hence, many automated classification systems that use machine learning approaches have been developed to help physicians obtain an objective second opinion for diagnosis decision. A variety of classifiers have been utilized for diagnosis, such as artificial neural network [10], support vector machine , naïve bayes , decision tree , nearest-neighbor , etc. In addition, hybrid models that harness the power of different classifiers have also been proposed. However, the diagnosis decision based on the classification result of a single classifier or a hybrid model only might be weak. Different classifiers probably offer contradictory classification results while providing complementary information. Therefore, it is helpful to combine the decisions of multiple classifiers. If the decision making is based on a group of classifiers which takes individual opinion of each classifier into consideration, the misclassified data - especially the patients who are undiagnosed by a certain classifier might be correctly diagnosed due to the correct decisions of other classifiers. There are a number of methods for combining classifiers, including mixture of experts [11], voting , boosting [12], bagging , etc. A few of them have been adapted for diagnosis of diabetes [13]. Some of these methods do not consider the weight of classifiers or each classifier has equal weight. But in fact, the weights of classifiers should be different and should be counted in the final decision. It makes more sense to give larger weights to classifiers which often make correct decisions and smaller weights to classifiers which usually make wrong decisions. On the other hand, some other methods adjust the weights of classifiers based on their power of prediction. In the meanwhile, they iteratively adjust the weights of instances, meaning that hard-to-classify instances get higher weights, which again influence the predictions of classifiers. The iterative interference between classifiers and instances makes the decision-making procedure complicated and time-consuming.

In this paper diagnose of various diseases within less time using weighted based voting system is explained whereas the organization of paper is as follows: Section 1 gives the detailed explanation about disease prediction process in data mining. Section 2 shows the literature survey part which gives the explanation about existing works. In section 3 weighted based voting process is clearly explained and section 4 gives the results and its analysis. Finally the paper ends with conclusion in section 5 which shows the consolidated summary of our work.

2. Literature survey

In 2016, S. Rajathi and G. Radhamani proposed an integrated framework k-nearest neighbor (kNN) with Ant Colony Optimization (ACO) technique [14]. The outcomes are compared with four dissimilar algorithms and the integrated framework shows accuracy, i.e., 70.26%. Few authors proposed an ensemble framework using hierarchical majority voting and multi-layer classification for the classification of disease and analysis using data mining approach [15].

The proposed framework named HMV overwhelms the limits of traditional routine blocks by using seven heterogeneous approaches and HMV is now based on three modules and gains an accuracy of 97%. Another ensemble combination method which is helpful in seeking up the best combination for heart disease analysis was proposed [16].

The method used in assembling is majority vote based and it is designed for every data set that belongs to the heart disease domain. The experiment prediction of data sets from different resources has two benchmarks. The accuracy of the ensemble model is 90%. Experimental observation shows that the best combination is when one of its classifier is a Naïve Bayes with an accuracy of 92%. In 2015, the researchers improved bagging technique and integrated it with the weighted voting scheme, they presented a novel classifier ensemble for the analysis and examination of heart disease [17].

The approach used 5 heterogeneous classifiers named as Naïve Bayes, SVM, linear regression, instance based learner, QDA (quadratic discriminant analysis) and obtained an accuracy of 84.16%. Extreme Learning Machine (ELM) is used to perfect attributes like age, sex, blood sugar, cholesterol, etc. The technique can substitute expensive medical checkups with a cautionary message for the patient which shows the probability of heart disease. This technique is applied on real world data where approximately 300 patients' data have been collected by Cleveland clinic foundation [18]. The accuracy shown by this model is 80%.

In 2014, the researchers developed intelligent, disease prediction classifier for heart disease prediction and analysis [19]. By combining five different machine learning classifiers, an ensemble model results that produce the prediction information for heart disease. Five different sets of attributes were used from five different data sets. They were assembled by a majority voting method for training and testing. The experimental result showed that MV5 predicts with a high accuracy, i.e., 88.52% as compared to other techniques.

SyedUmar et al. proposed a hybrid model in which major risk factors are used for the analysis of heart disease [20]. The hybrid model involves two data mining tools, one is neural network and the other one is genetic algorithm. Using global optimization, genetic algorithm initializes the weight of a neural network. Adapting power of this model is fast and as compared to other models and the prediction accuracy is 89%.

3. Research methodology

A disease diagnosis system is created in order to predict different diseases such as diabetes, kidney disease liver disease, heart disease. At decision support system, dataset of different diseases are loaded and apply data mining algorithms to train dataset. Requested user inputs are collected and processed on server to predict the diagnosis result.

3.1 Heterogeneous classifier system:

Multiple Classifier System is a set of classifiers whose individual predictions are combined in some way to classify new examples. That is, Classifier ensemble methods use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms. Combining identical classifiers is useless. A necessary condition for the approach to be useful is that member classifiers should have a substantial level of disagreement. That is, they should make error independently with respect to one another. Member classifiers should make uncorrelated errors with respect to one another; each classifier should perform better than a random guess. Using MCS can reduce the local different behaviors of individual classifiers by averaging the results of each member classifier.

3.2 Voting system

We make an assumption that there are M classes C1,C2,C3….Cn in the database, with each of Ci denoting the ith class. There are K classifiers E1,E2,E3…..Ek, where each Ek denotes the k th classifier. The confusion matrix PTk can be obtained by using Ek to classify a testing sample set. For classifier Ek, with its knowledge of the confusion matrix PTk , the probabilities that propositions Ci = 1,2…Mare true under the occurrence of the event.

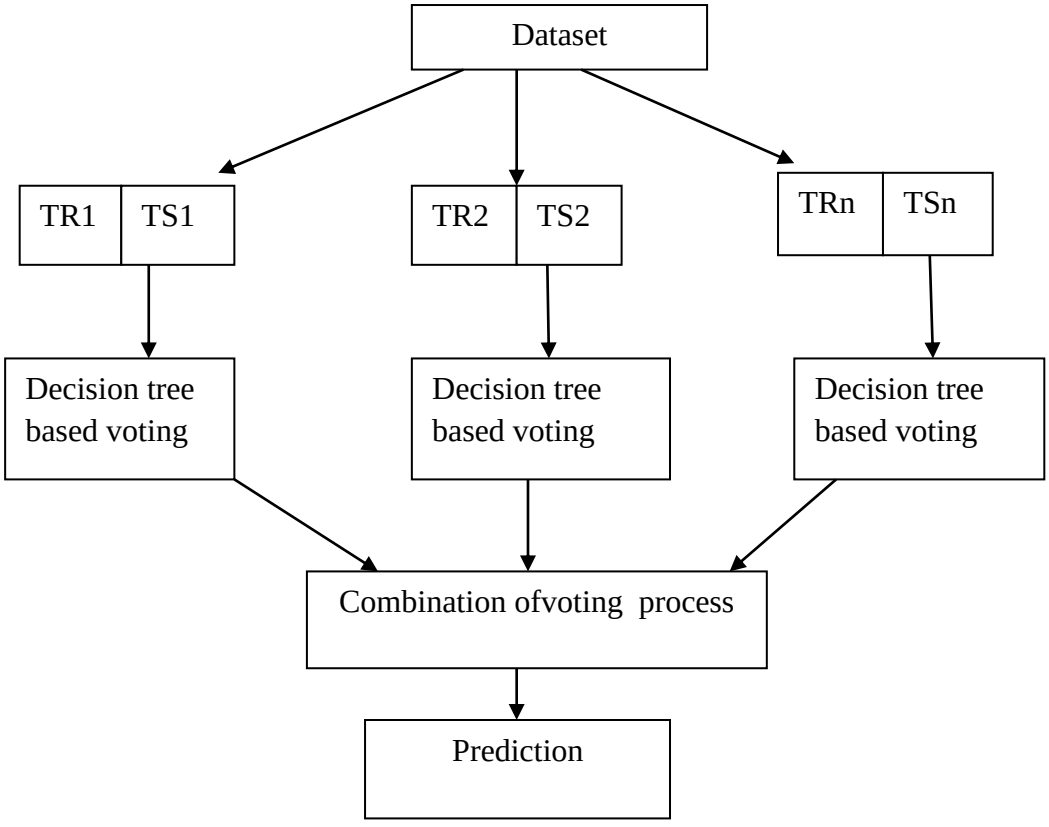

Figure-1 Disease prediction using voting system

The further computations are then performed on preprocessed data by classifier training module. Training set is a labeled dataset used for ensemble training. After training the classifiers of each model group, weneed to aggregate independently trained classifiers of eachgroup into an appropriate combination

method.The Weighted Majority Voting (WMV) ensemble mechanism sorts out an unlabeled instance into a class, which gets most common votes or the highest number of voting. The WMV ensemble mechanism is generally denoted as Plurality Vote (PV) approach.Most often, the WMV mechanism is applied for equating the performance of various models. Mathematically

$$\text{class(x)} = \arg max_{c_i \in dom\,(y)} \left(\sum_k g(y_k(x), c_i) \right)$$

Where the classification of the K[th] classifier is denoted as $y_k(x)$ and $g(y,c)$ represents about the index function which can demonstrated as follows

$$g(y,c) = \begin{cases} 1 & y = c \\ 0 & y \neq c \end{cases}$$

If the probabilistic classifier is utilized, the crisp classification $y_k(x)$ is got from the following equations

$$\text{class(x)} = \arg max_{c_i \in dom\,(y)} PM_k(y = c_i|x)$$

Where M_k is applied to demonstrate the classifier k and $PM_k(y = c_i|x)$ represents about the probability of class c for an instance x.

Each voting process specifies a different weight for each base classifier. This weight depends on the accuracy of the classifier in predicting this learner's severity level. For every bug report, each of the n base classifiers predicts a severity category.

4. Performance analysis

Accuracy is termed as ratio of the number of correctly classified instances to the total number of instances.
Accuracy = (TP+TN)/(TP+TN+FP+FN)

Precision is the ration of actually true predicted instances out of the total true instances. Precision = TP/(TP+FP) Recall is the ratio of actual true instances out of all the items which are true.

Recall = TP/(TP+FN) F-measure is the harmonic mean of both precision and recall.

F_Measure = 2*(Precision*Recall)/(Precision + Recall)

Type of disease	Precision	accuracy	F-measure	Computational time
Diabetes Disease Detection	0.9821	0.9935	0.991	2.4sec
Kidney Disease Detection	0.9875	0.9937	0.9731	1.5sec
Liver Disease	0.9667	0.9914	0.9831	1.8sec

Detection				
Heart disease detection	0.9783	0.9846	0.961	1.9sec

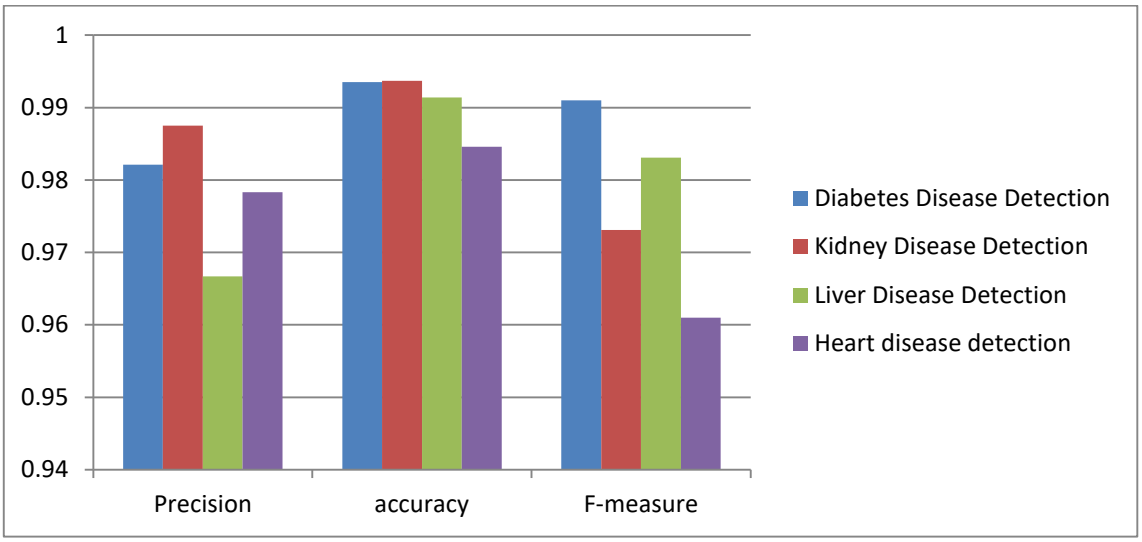

Figure-2 Result analysis of various disease detection

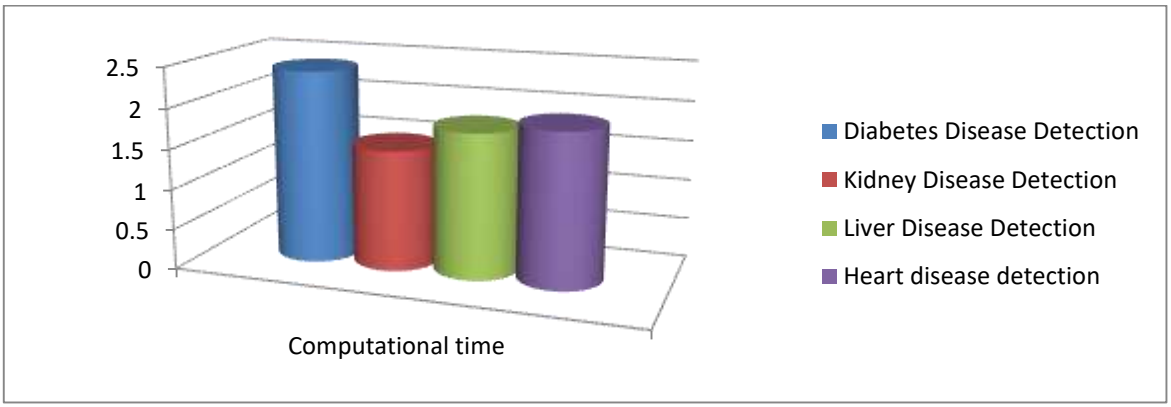

Figure-3Computational time analysis of various disease detection

5. **Conclusion**

This paper is mainly focused to predict disease possibility using data mining or machine learning approach in order to enhance the accuracy or precision of the disease detection expert system. Here, we propose a weight-adjusted voting approach to automatically diagnose diabetes based on the decisions of an ensemble classifier. Our approach beats each single classifier in the ensemble from the perspective of sensitivity while maintaining reasonably high specificity and accuracy.

References

1. Zhu, Y. and J. Fang, Logistic Regression-Based Trichotomous Classification Tree and Its Application in Medical Diagnosis. Medical Decision Making, 2016. 36(8): p. 973-989.
2. Zhao, S.D., Integrative genetic risk prediction using non-parametric empirical Bayes classification. Biometrics, 2017. 73(2): p. 582-592.
3. Dong, E., et al., Classification of multi-class motor imagery with a novel hierarchical SVM algorithm for brain– computer interfaces. Medical & Biological Engineering & Computing, 2017. 55(10): p. 1809-1818.
4. Zhu, M., et al., Dimensionality Reduction in Complex Medical Data: Improved Self-Adaptive Niche Genetic Algorithm. Computational and engineering
5. Bak, B.A. and J.L. Jensen, High Dimensional Classifiers in The Imbalanced Case. Computational Statistics & Data Analysis, 2016. 98: p. 46-59.
6. Zhang, Y., et al., Imbalanced Data Classification Based on Scaling Kernel-Based Support Vector Machine. Neural Computing and Applications, 2014. 25(3-4): p. 927-935.
7. Maurya, C.K., D. Toshniwal and G. VijendranVenkoparao, Online sparse class imbalance learning on big data. Neurocomputing, 2016. 216: p. 250-260.
8. Alstouhi, S. and C.K. Reddy, Transfer Learning for Class Imbalance Problems with Inadequate Data. Knowledge and Information Systems, 2016. 48(1): p. 201-228.
9. Elbanna, M., Modified Mahalanobis Taguchi System for Imbalance Data Classification. Computaional Intelligence and Neuroscience, 2017. 2017: p. 15
10. N. H. Barakat, A. P. Bradley, and M. N. H. Barakat, “Intelligible support vector machines for diagnosis of diabetes mellitus”, IEEE Transactions on Information Technology In Biomedicine, vol. 14 (4), pp. 1114 – 1120, 2010.
11. Y. Huang, P. McCullagh, N. Black, and R. Harper, “Feature selection and classification model construction on type 2 diabetic patients’ data”, Artificial Intelligence in Medicine, vol. 41, pp. 251 – 262, 2007.
12. M. A. Chikh, M. Saidi, and N. Settouti, “Diagnosis of diabetes diseases using an artificial immue recognition systems2 (AIRS2) with fuzzy knearest neighbor”, Journal of Medical Systems, vol. 36 (5) , 2721 – 2729, 2012.
13. E. D. Ubeyli, “Automatic diagnosis of diabetes using adaptive neurofuzzy inference systems”, Expert Systems, vol. 27 (4), pp. 259 – 266, 2010.
14. S. Rajathi, G. Radhamani, “Prediction and analysis of Rheumatic heart disease using KNN classification with ACO,” International Conference on Data Mining and Advanced Computing (SAPIENCE), IEEE, pp. 68- 73, 2016.

15. S. Bashir, U. Qamar, F.H. Khan, L. Naseem, " HMV: A medical decision support framework using multi-layer classifiers for disease prediction," Journal of Computational Science, Elsevier, vol. 13, pp. 10- 25, 2016
16. R. El Bialy, M.A. Salama, O. Karam, "An ensemble model for Heart disease data sets: a generalized model," Proceedings of the 10th International Conference on Informatics and Systems, ACM, pp. 191- 196, 2016.
17. S. Bashir, U. Qamar, F.H. Khan, "BagMOOV: A novel ensemble for heart disease prediction bootstrap aggregation with multi-objective optimized voting," Australasian Physical & Engineering Sciences in Medicine, Springer, vol. 38, pp. 305-323, 2015.
18. S. Ismaeel, A. Miri, D. Chourishi, "Using the extreme learning machine (elm) technique for heart disease diagnosis," Humanitarian Technology Conference (IHTC2015), IEEE Canada International, pp. 1-3, 2015.
19. S. Bashir, U. Qamar, M.Y. Javed, "An ensemble based decision support framework for intelligent heart disease diagnosis," In Information Society International Conference, IEEE, pp. 259-264, 2014.
20. S. Bashir, U. Qamar, F.H. Khan, M.Y. Javed, "MV5: a clinical decision support framework for heart disease prediction using majority vote based classifier ensemble," Arabian Journal for Science and Engineering, Springer, vol. 39, pp. 7771-7783, 2014