Parkinson's Diseases Diagonstic System Using Machine Learing Classifiers

Aishwarya K[1], Kiruthika M [2], Kiruthika S[3],Sadasivam V[4]
[1, 2, 3] Dept of Information Technology
[4]Professor, Dept of Information Technology
[1, 2, 3, 4] K S.Rangasamy College of Technology, Tiruchengode, Tamil Nadu

Abstract- *In this study, a novel method is proposed for the detection of Parkinson's disease with the features obtained from the speech signals. Detection and early diagnosis of Parkinson's disease are essential in terms of disease progression and treatment process. Parkinson's disease dataset used in this study was obtained from the UCI machine learning repository. The proposed hybrid machine learning method consists of two stages: i) data pre-processing (oversampling), ii) classification. The Parkinson's disease dataset (PD dataset) is a two-class dataset. While 192 data belong to normal (healthy) individuals, 564 data belong to the diseased class (PD). The data set has an imbalanced class distribution. To transform this imbalanced dataset to balanced dataset, SMOTE (Synthetic Minority Over-Sampling Technique) method is used. Then, after converting to a balanced class distribution, Random Forests classification method was used for classification of Parkinson's disease dataset. The PD dataset consists of 753 attributes. Only the random forests classification were classified as 87.037% in the classification of PD dataset, while the proposed hybrid method (the combination of SMOTE and random forests) achieved 94.89% classification success. Obtained results showed that promising results had been achieved in discrimination of the PD dataset with this hybrid method.*

I. INTRODUCTION

It is difficult to classify and categorize unbalanced data sets. While there are any data in a class, there is very little data in the other class, so generalization on the data set is a difficult problem. There are several approaches to solve the class imbalance problem in the literature. Before the model is created, the problem of imbalance can be eliminated, regardless of the classifier by artificially balancing the training data set. This method is known as data re-sampling. Alternatively, classifiers can solve the problem of class-imbalance by classifying models for better estimation of the minority class. Alternatively, only one class of classes can be modeled, and this is called one-class learning.

II. EXISTING SYSTEM

Over-sampling is carried out by simply replacing existing elements of the minority class in the educational setting. This method leads to over fitting. To prevent this over fitting, new samples can be artificially produced by the distribution of the minority class. This approach is the Synthetic Minority Over-sampling Technique (SMOTE). There are many studies in the literature regarding the classification of Parkinson's disease dataset. Some of these studies are given below. Sakar et al. used several signal processing algorithms for Parkinson's disease from speech signals and formed the PD data set. The authors examined the effect of tunable Q-factor wavelet transform (TQWT) method and obtained good results.

III. PROPOSED SYSTEM

Some of these studies are given below used several signal processing algorithms for Parkinson's disease from speech signals and formed the PD data set. The authors examined the effect of tunable Q-factor wavelet transform (TQWT) method and obtained good results. In the classification of Parkinson's disease, have proposed a new method called Multiple Feature Evaluation Approach (MFEA) and individually combined with classification algorithms. They achieved their best success with the SVMMFEA combination. Deepak Joshi et al. proposed a new hybrid method on the detection of Parkinson's disease from walking signals. In this method, wavelet analysis methods are combined with SVM.

- The features have been obtained from the speech signal processing methods. The data in PD dataset have been taken from 188 patients with PD (107 men and 81 women) with ages ranging from 33 to 87.
- The performance of K-NN, Random Forest, and Ada-Boost was compared, and it was found that K-NN had the highest accuracy of 90.26%. In 2015, a study performed feature selection and classified the dataset using SVM.

Feature selection was then performed using Principal Component Analysis (PCA) and th3 accuracies of different classifiers were compared.

IV. SYSTEM MODULES

Classification capability of hybrid features:

Seen from the, over 80% of classification accuracy (ACC) can be obtained regardless of different evaluation criteria. Compared with the ACC from the provider of the dataset, the improvement is apparent. With the number of the selected feature increasing from zero to tens of features, the ACC rapidly reach 70% or so. After that, the improvement of the ACC becomes slow. When the number of selected features reaches 200 or so, the ACC is close to the optimal ACC. the curves of ACC based on different evaluation criteria. For corrcoef, the curve is fastest to reach close to the best ACC. For p_value and recorr2, the curves are similar and can obtain best ACC.

Parkinson’s disease:

Parkinson's disease (PD) is a common neurological degeneration disease; its clinical manifestations include static tremor, slow movement, muscle rigidity and postural gait disorder. According to statistics, the current total number of patients with Parkinson's disease in China more than 2 million, accounting for about half of the global number of such patients. Recent studies showed that speech signal (data) is helpful for recognizing PWP (people with Parkinson’s) from healthy people, because most of the PWP has vocal disorder in daily communication to some Extent.

Synthetic Minority over Sampling Technique:

Over-sampling is carried out by simply replacing existing elements of the minority class in the educational setting. This method leads to over fitting. To prevent this over fitting, new samples can be artificially produced by the distribution of the minority class. This approach is the Synthetic Minority Over-sampling Technique (SMOTE). There are many studies in the literature regarding the classification of Parkinson's disease dataset. Some of these studies are given below. Sakar et al. used several signal processing algorithms for Parkinson's disease from speech signals and formed the PD data set.

Tunable Q-factor wavelets transform:

The authors examined the effect of tunable Q-factor wavelet transform (TQWT) method and obtained good results. In the classification of Parkinson's disease, have proposed a new method called Multiple Feature Evaluation Approach (MFEA) and individually combined with classification algorithms. They achieved their best success with the SVMMFEA combination. Deepak Joshi et al. proposed a new hybrid method on the detection of Parkinson's disease from walking signals. In this method, wavelet analysis methods are combined with SVM. Apart from the literature, a new hybrid method based on SMOTE and Random Forests classification was proposed, and promising results were obtained by applying the Parkinson's disease dataset.

4.2 MACHINE LEARNING ALGORITHM

Acute respiratory distress syndrome (ARDS) is a serious respiratory condition In which Supervised machine learning predictions may help to predict patients accuracy level

Support vector machine(SVM)

“Support Vector Machine” (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. Support Vectors are simply the co-ordinates of individual observation.We proposed this formulation for account label uncertainty in the classification model in the following manner.

Naive Bayes classifier

It this classification technique based on Bayes' Theorem with an assumption of independence among a predictors. we proposed a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

Random forest classifier

We proposed in this Random forest is a supervised learning algorithm which is used for both classification and regression.we Similarly, random forest algorithm creates decision trees on data samples and then gets the prediction from each of them values and finally selects the solution.

4.3 PREDICTION

We can predict the uncertainty and also we can predict the accuracy level of Acute Respiratory Distress Syndrome (ARDS) datasets using the algorithms and methods mentioned in the above modules.

V. CONCLUSION

The solving of the class-imbalance problem is very hard to handle in machine learning. There are some approaches to handle this problem in the literature. One of the best solutions is the SMOTE (Synthetic Minority Over-sampling Technique). In this paper, the SMOTE and Random Forests classifier have been combined to classify the Parkinson disease dataset. In the SMOTE approach, the number of samples for minority class in the PD dataset has been synthetically increased to balance the dataset. Only the random forests classification were classified as 87.037% in the classification of PD dataset, while the proposed hybrid method (the combination of SMOTE and random forests) achieved 94.89% classification success. The proposed hybrid model could be used in other medical real world class-imbalanced classification problems.

VI. ACKNOWLEDGEMENT

We are very proudly rendering our thanks to our Principal Dr. R. GOPALAKRISHNAN, M.E., Ph.D., for the facilities and the encouragement given by him to the progress and completion of our project.

We proudly render our immense gratitude to the Head of the Department Dr.A.SABARI M.Tech.,M.B.A., Ph.D., for his effective leadership, encouragement and guidance in the project.

We are highly indebted to provide our heart full thanks to our guide Professor DrV.Sadasivam,B.E.,M.E.,Ph.D., for her valuable ideas, encouragement and supportive guidance throughout the project.

We wish to extend our sincere thanks to all faculty members of our Information Technology Department for their valuable suggestions, kind co-operation and constant encouragement for successful completion of this project.

We wish to acknowledge the help received from various Departments and various individuals during the preparation and editing stages of the manuscript.

REFERENCES

[1] Accounting for Label Uncertainty in Machine Learning for Detection of Acute Respiratory Distress Syndrome,IEEE Journal Biomedical Health Information,Jan 2019.

[2] G.D. Rubenfeld et al, Incidence and outcomes of acute lung injury, N Engl J Med, vol. 353(16), pp. 1685-1693. Oct 2005.

[3] R.M. Sweeney, D.F. McAuley. Acute respiratory distress syndrome, Lancet, vol. 388(10058), pp. 2416-2430. Nov 2016.

[4] G. Bellani et al, Epidemiology, Patterns of Care, and Mortality for Patients With Acute Respiratory Distress Syndrome in Intensive Care Units in 50 Countries, JAMA, vol. 315(8), pp. 788-800. Feb 2016.

[5] B.J. Clark, M. Moss, the Acute Respiratory Distress Syndrome: Dialing in the Evidence? JAMA, vol. 315(8), pp. 759-761. Feb 2016.

[6] M.W. Sjoding, R.C. Hyzy, Recognition and Appropriate Treatment of the Acute Respiratory Distress Syndrome Remains Unacceptably Low, Crit Care Med, vol. 44(8), pp. 1611-1612. Aug 2016.

[7] M.W. Sjoding, Translating evidence into practice in acute respiratory distress syndrome: teamwork, clinical decision support, and behavioral economic interventions, Curr Opin Crit Care. Jul 2017.

[8] H.C. Koenig et al, Performance of an automated electronic acute lung injury screening system in intensive care unit patients, Crit Care Med,vol. 39(1), pp. 98-104. Jan 2001.

[9] G.D. Rubenfeld et al, Interobserver variability in applying a radiographic definition for ARDS, Chest, vol. 116(5), pp. 1347-53. Nov 1999.

[10] M.W. Sjoding et al, Acute Respiratory Distress Syndrome Measurement Error: Potential Effect on Clinical Study Results, Ann Am Thorac Soc,vol. 13(7), pp. 1123-8. Jul 2016.

[11] C.V. Shah et at, An alternative method of acute lung injury classification for use in observational studies, Chest; vol. 138(5), pp. 1054-1061. Nov 2010.

[12] D.F. Nettleton, A. Orriols-Puig, A. Fornells, A study of the effect of different types of noise on the precision of supervised learning techniques, Artificial Intelligence Reviews, Vol. 33(40), 2010.

[13] B. Frenay and M. Verleysen, ”Classification in the Presence of Label Noise: A Survey,” IEEE Transactions on Neural Networks and Learning Systems, vol. 25(5), pp. 845-869. May 2014.

[14] N. Natarajan et al, Learning with noisy labels,” in Neural Information Processing Systems, pp. 1196-1204. December 2013.

[15] Y. Duan and O. Wu, Learning With Auxiliary Less-Noisy Labels, IEEE Transactions on Neural Networks and Learning Systems, vol. 28(7), pp.1716-1721. May 2017.

[16] S. Vembu and S. Zilles, ”Interactive Learning from Multiple Noisy Labels,” in Joint European Conference on Machine Learning and Knowledge Discovery in

Databases, Springer International Publishing. 2016,pp. 493-508.

[17] X. Yang, Q. Song, Y. Want, A weighted support vector machine for data classification, International Journal of Pattern Recognition and Artificial Intelligence, vol 21, Nov 5 (2007).

[18] E. Osuna, R. Freund, F. Girosi. An improved training algorithm for support vector machines, in Neural Networks for Signal Processing[1997] VII. Proceedings of the 1997 IEEE Workshop (pp. 276-285).

[19] J. Shawe-Taylor et al, ”Structural risk minimization over data-dependent hierarchies,” IEEE Transactions on Information Theory, vol. 44(5), pp.1926-1940, Sep 1998.

[20] R. Bellazzi and A. Riva, Learning conditional probabilities with longitudinal data, in Working Notes of the IJCAI Workshop Building Probabilistic Networks: Where Do the Numbers Come From, 1995 (pp.7-15).