A Hybrid Approach to Parkinson Disease Classification using speech signal: The combination of SMOTE and Random Forests

Kemal Polat
Department of Electrical and Electronics Engineering, Faculty of Engineering,
Bolu Abant Izzet Baysal University, Bolu, Turkey
kpolat@ibu.edu.tr

***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 (over-sampling), 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 resultshad been achieved in discrimination of the PD dataset with this hybrid method.

***Keywords*—** Parkinson disease, Hybrid method, imbalanced dataset, SMOTE, classification

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 [1,2,3].

Data resampling can be done by reducing data in the intensive class, or by increasing instances of the minority class. This strategy is primarily used in large data sets, where the loss of information by marginalization is marginal [1,2,3]. In this study, SMOTE is an over-sampling method developed by Chawla et al. in 2002 [4].

Over-sampling is carried out by simply replacing existing elements of the minority class in the educational setting. This method leads to overfitting [1,2,3]. To prevent this overfitting, new samples can be artificially produced by the distribution of the minority class. This approach is the Synthetic Minority Over-sampling Technique (SMOTE) [4].

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 [5]. In the classification of Parkinson's disease, Salama A.Mostafa et al. have proposed a new method called Multiple Feature Evaluation Approach (MFEA) and individually combined with classification algorithms. They achieved their best success with the SVM-MFEA combination [6]. 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 [7].

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.

II. MATERIAL AND METHOD

A. Parkinson Disease Data Set

Parkinson's disease dataset was taken from the UCI machine-learning database [8]. In this dataset, there are 756 samples and 753 features. Also, this dataset is a two-class problem: 192 samples belonging to healthy and 564 samples belonging to patients. Therefore, this dataset is an imbalanced dataset. Sakar and et al. have created the Parkinson disease dataset in 2018 [5]. 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 [5]. Figure 1 shows the class distribution

978-1-7281-1013-4/19/$31.00 ©2019 IEEE

Of Parkinson disease dataset. In this figure, there are two classes including green one (majority class) and red one (minority class). The discrimination between these classes is very challenging.

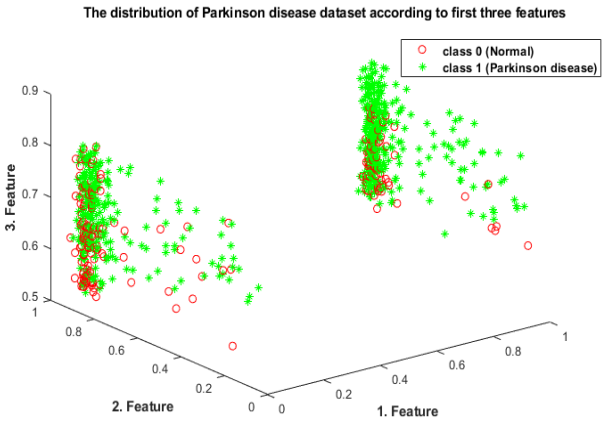

Fig.1. The class distribution of Parkinson disease dataset

B. *The Proposed Method*

In this study, a novel method combining SMOTE and Random Forest classifier is proposed to detect the PD dataset having an imbalanced class distribution. First, the SMOTE method has been applied to PD dataset to handle this problem in the dataset. Therefore, PD dataset has been converted to a balance class distribution using SMOTE. Figure 2 shows the block diagram of the proposed hybrid method.

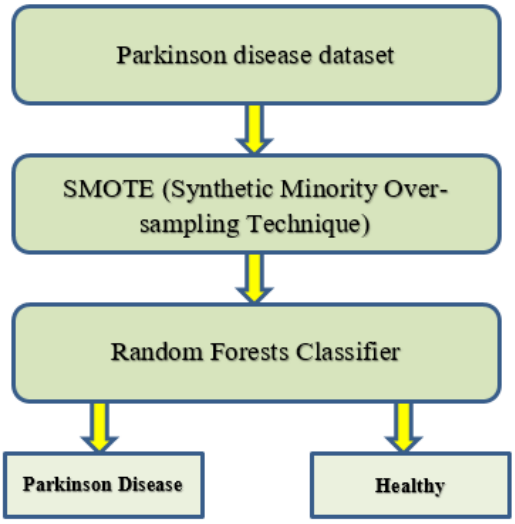

Fig.2. The flow chart of the proposed method to classify the PD dataset

C. *The SMOTE (Synthetic Minority Over-sampling Technique)*

SMOTE is a new method of over-sampling of the minority class by producing artificial specimens instead of over-sampling [4]. SMOTE is one of the new methods used to solve class-unbalance problems in machine learning problems. Figure 3 presents a schematic representation of the SMOTE algorithm.

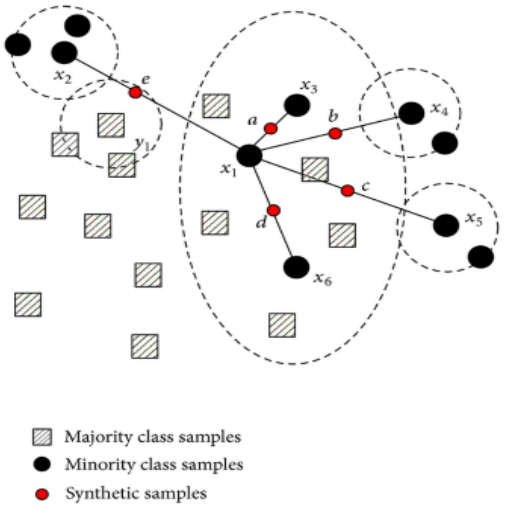

Fig.3. The schematic representation of the SMOTE algorithm [9]

After applying the SMOTE method to PD dataset, the new class distribution of PD dataset is given in Figure 4. As can be seen from the below figure, the number of samples of minority class (healthy) was increased by the production of new synthetic samples by the SMOTE method. So, the problem of class-imbalanced has been handled by using SMOTE algorithm in the classification of PD dataset.

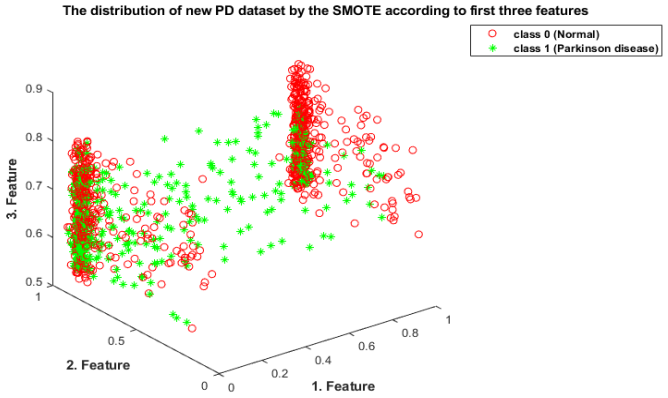

Fig.4. The class distribution of Parkinson disease dataset

Usually, in the PD dataset, 192 samples have a healthy group, and 564 samples are having Parkinson disease group. After applied the SMOTE method to PD dataset, the number of samples of the healthy group in the PD dataset has been increased and raised from 192 to 767 samples. The number of samples of Parkinson disease group has not been changed.

D. *Random Forests (RF) Classifier*

Random forest is a robust learning algorithm proposed by Breiman in 2001 [11]. Random Forests (RF) is a popular ensemble algorithm used to create forecasting models to solve

both classification and regression problems. Random Forests provide both high performance in the classification and high generalization of data. An RF classifier is a method of learning a consulted machine. As the name implies, it first creates a forest, then combines the results of decision trees trained by the "bagging "method [10, 11]. In this work, the RF classifier has been used to classify the new PD dataset having class-balanced distribution by the SMOTE algorithm.

III. Experimental Results

In the study, a novel hybrid classification model for classifying the Parkinson disease dataset having class-imbalance data distribution. In the train and test of the Random Forests classifier, the hold out method with 50-50% train-test partition, and 10-fold cross validation have been used. As the performance measures, the classification accuracy, Kappa value, Precision, Recall, F-measure, and AUC (area under the ROC curve) have been used to evaluate the proposed method. Table 1 shows the obtained results of the Random Forests classifier only in the classification of PD dataset with the holdout method of 50-50% train-test partition. Table 2 shows the obtained results of the Random Forests classifier only in the classification of PD dataset with the 10-fold cross-validation. Table 3 gives the obtained results of the combination of SMOTE and Random Forests classifier in the classification of PD dataset with the hold out the method of 50-50% train-test partition. Table 4 shows the obtained results of the combination of SMOTE and Random Forests classifier in the classification of PD dataset with the 10-fold cross-validation.

TABLE I
RANDOM FORESTS RESULT IN DETECTION OF PARKINSON DISEASE DATASET USING HOLD OUT METHOD (50-50% TRAIN –TEST PARTITION)

Classification Acc. (%)	Kappa	Precision	Recall	F-Measure	AUC
81.74	0.438	0.815	0.817	0.795	0.837

TABLE II
RANDOM FORESTS RESULT IN DETECTION OF PARKINSON DISEASE DATASET USING 10-FOLD CROSS VALIDATION

Classification Acc. (%)	Kappa	Precision	Recall	F-Measure	AUC
87.03	0.612	0.871	0.87	0.86	0.94

TABLE III
THE PROPOSED HYBRID METHOD RESULTS IN DETECTION OF PARKINSON DISEASE DATASET USING HOLD OUT METHOD (50-50% TRAIN –TEST PARTITION)

Classification Acc. (%)	Kappa	Precision	Recall	F-Measure	AUC
92.34	0.833	0.924	0.923	0.923	0.973

TABLE IV
THE PROPOSED HYBRID METHOD RESULTS IN DETECTION OF PARKINSON DISEASE DATASET USING 10-FOLD CROSS VALIDATION

Classification Acc. (%)	Kappa	Precision	Recall	F-Measure	AUC
94.89	0.894	0.951	0.949	0.949	0.991

It can be seen from the obtained results that the proposed hybrid model has achieved good results in the discrimination of Parkinson disease dataset having a class-imbalanced problem.

IV. Conclusions

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.

References

[1] Sebastián Maldonado, Julio López, Carla Vairetti, An alternative SMOTE oversampling strategy for high-dimensional datasets, Applied Soft Computing, Volume 76, 2019, 380-389.

[2] Chawla N.V., Japkowicz N., Kotcz A. Editorial: special issue on learning from imbalanced data sets SIGKDD Explor., 6 (2004), pp. 1-6.

[3] Sun Y.M., Kamel M.S., Wong A.K.C. Classification of imbalanced data: A Review Int. J. Pattern Recognit. Artif. Intell., 23 (2009), pp. 687-719.

[4] Chawla N.V., Hall L.O., Bowyer K.W., Kegelmeyer W.P. SMOTE: Synthetic minority oversampling technique J. Artificial Intelligence Res., 16 (2002), pp. 321-357.

[5] C. Okan Sakar, Gorkem Serbes, Aysegul Gunduz, Hunkar C. Tunc, Hatice Nizam, Betul Erdogdu Sakar, Melih Tutuncu, Tarkan Aydin, M. Erdem Isenkul, Hulya Apaydin, A comparative analysis of speech signal processing algorithms for Parkinson's disease classification and the use of the tunable Q-factor wavelet transform, Applied Soft Computing, 74, 2019, 255-263.

[6] Salama A. Mostafa, Aida Mustapha, Mazin Abed Mohammed, Raed Ibraheem Hamed, N. Arunkumar, Mohd Khanapi Abd Ghani, Mustafa Musa Jaber, Shihab Hamad Khaleefah, Examining multiple feature evaluation and classification methods for mproving the diagnosis of Parkinson's disease, Cognitive Systems Research, 54, 2019, 90-99.

[7] Deepak Joshi, Aayushi Khajuria, Pradeep Joshi, An automatic non-invasive method for Parkinson's disease classification, Computer Methods and Programs in Biomedicine, 145, 2017, 135-145.

[8] https://archive.ics.uci.edu/ml/datasets.html, (last accessed: April, 2019).

[9] https://medium.com/coinmonks/smote-and-adasyn-handling-imbalanced-data-set-34f5223e167, (last accessed: April, 2019).

[10] https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd, (last accessed: April, 2019).

[11] L. Breiman, Random forests, Machine Learning, 45 (1) (2001), pp. 5-32.