Parkinson's disease,
medical diagnosis,
data classification

Wojciech FROELICH[1], Krzysztof WROBEL[1], Piotr PORWIK[1]

DIAGNOSING PARKINSON'S DISEASE USING THE CLASSIFICATION OF SPEECH SIGNALS

This paper addressees the problem of an early diagnosis of Parkinson's disease by the classification of characteristic features of person's voice. A new, two-step classification approach is proposed. In the first step, the voice samples are classified using standard state-of-the-art classifiers. In the second step, the classified samples are assigned to patients and the final classification process based on majority criterion is performed. The advantage of using our new approach is the resulting, reliable patient-oriented medical diagnose. The proposed two-step method of classification allows also to deal with the variable number of voice samples gathered for every patient. Preliminary experiments revealed quite satisfactory classification accuracy obtained during the performed leave-one-out cross validation.

1. INTRODUCTION

In this paper an important problem [5], [8] of the early diagnosis of Parkinson's disease is addressed. The name of the disease comes from James Parkinson who described it in work [11]. Parkinson's disease is a disorder of central nervous system and leads to many health issues, e.g., rigidity, imbalances, difficulty in talking, or slowness of movements. It is estimated that seven to ten million people currently suffer Parkinson's disease worldwide. First symptoms are noted for people over the age of fifty (the average age of the onset of disease is about fifty nine).

Parkinson's disease manifests also in form of disorders of person's speech. Therefore it is possible to diagnose Parkinson's disease using voice signals [12], [14]. On the basis of voice sample a vector of numerical values is calculated. The obtained values represent selected characteristic features of the recorded voice. Finally, the characteristic vector of the voice sample is classified indicating whether the corresponding person exhibits a symptom of the disease.

The problem of an early diagnosis of Parkinson's disease has raised an interest of numerous researchers [3], [7], [13]. In particular, the application of artificial neural networks to discriminate healthy people from those with Parkinson's disease using voice signals was proposed in [3]. Parkinson's disease classification using gait characteristics and wavelet-based feature extraction was proposed in [13]. Application of fuzzy k-nearest neighbor model to an efficient detection of Parkinson's disease was proposed in [7].

[1] Institute of Computer Science, University of Silesia, ul. Bedzinska 39, Sosnowiec

Recently, a hybrid system based on model-based clustering and classification using support vector machine and artificial neural networks has been proposed [6]. Comparative experiments with different classifiers applied to the diagnosis of Parkinson's disease using voice signals was made in [12]. A literature overview on signal processing algorithms for the classification of Parkinson's disease can be found in [15].

There are several limitations of the existing studies. To the best of our knowledge, all existing approaches deal with the problem of classifying individual voice samples and do not take into account the distribution of those samples and their classes with respect to persons. Moreover, the existing studies do not take into account that the voice samples for a given person are expected to be highly correlated with each other. For those reasons, even if the obtained classification accuracy of all voice samples in their entire population is high, the accuracy obtained for diagnosing individual patients remains unknown. In the case of unbalanced distribution of samples and their classes with respect to persons the final accuracy of diagnosis maybe much lower. Just this distribution is not taken into account in all of the existing studies raising a doubt, how many persons were in fact correctly diagnosed ?

Also in many of the existing studies 10-fold or 5-fold cross validation tests were performed using the datasets of hundreds of samples. However, after assigning those numerous samples to real persons, it may turn out that the number of persons is too low for using those types of validations with respect to people. This raises another doubt regarding the usability of the obtained results.

In this paper we address the above stated limitations. Our study proposes a new, two-step classification approach. In the first step, the voice samples are classified using standard state-of-the-art classifiers. In the second step, the classified samples are assigned to patients and the final classification process based on the proposed criterion of majority is performed. Leave-one-one cross validation (dedicated to data sets with low cardinality) is performed with respect to persons. Dependent on the type of classifiers used during the first step of the procedure, the resulting accuracy of the proposed approach differs. Comparative experiments revealed the highest accuracy of diagnosis when applying naive Bayes and fuzzy rule-based system during the first step of the proposed approach. The approach presented in this paper is investigated for the first time, therefore the obtained results are not directly comparable to those available in the existing literature.

The remainder of this paper is organized in the following way. In Section 2, the proposed approach to the classification of persons is proposed. Section 3 provides selected details on the implementation of the method. The description of the applied data and the results of experiments are presented in Section 4. Section 5 concludes the paper.

2. DIAGNOSING TECHNIQUE

In this section we propose a novel two-step classification technique enabling to diagnose persons regarding Parkinson's disease. The approach is based on the classification of vectors of characteristic features calculated on the basis of voice samples. A formal description of the proposed model is proposed in the following.

Let $P = \{P_1, P_2, ..., P_n\}$ be the set of persons, where $n = card(P)$ is its cardinality. For every person P_i, the classification $d_i \in K$ is made, where: K is the assumed set of classes. For our purpose $K = \{0, 1\}$, where: 0 denotes a healthy person and 1 indicates the occurrence of Parkinson's disease. The value of d_i plays in fact the role of medical diagnosis.

Let us assume that the medical diagnosis d_i can be calculated on the basis of voice samples collected for the i^{th} person. For every voice sample a vector of characteristic features is calculated. This way to every person P_i a set of characteristic vectors S_i is assigned. Let

$S = \{S_1, S_2, ..., S_n\}$ be the set of characteristic vectors collected for all persons, where: $card(S) = \sum_{i=1}^{n} card(S_i)$. The cardinality $card(S_i)$ can obviously differ with respect to every person.

Let us denote a single characteristic vector as $s_{ij} \in S_i$, where the subscripts i, j denote the index of the person and the index of the characteristic vector respectively. It is assumed that for every vector s_{ij} a binary classification reflecting the severity of Parkinson's disease is assigned. Let us denote as $d_{ij} \in K$ the classification of s_{ij}. In case of historical data, when the medical diagnosis d_i of the person P_i is known and verified by the doctor, every classification d_{ij} of the vector s_{ij} is also known, i.e., $\forall_{i=1,...,card(S)} \forall_{j=1,...,card(S_i)} (d_{ij} = d_i)$, where: $d_{ij}, d_i \in K$.

The situation becomes complex in the case of the newly diagnosed person whose medical diagnosis should be calculated on the basis of the corresponding set of characteristic vectors S_i. The classification of every characteristic vector $s_{ij} \in S_i$ is also unknown and should be estimated with the help of any predictive model. Let us denote $d'_{ij} = M(s_{ij})$ as the predicted classification, where M denotes a model (classifier) trained on the basis of previously classified historical data. It is assumed that the same classifier M is used for all persons and characteristic vectors.

After performing the classifications $d'_{ij} = M(s_{ij})$ of all characteristic vectors $s_{ij} \in S_i$ there is a need to predict the final medical diagnosis for the person P_i, i.e., to calculate the value of d'_i using the set of previously calculated $d'_{ij}, i = 1, 2, ..., card(S_i)$.

We will say that the characteristic vector s_{ij} supports the class $k \in K$, when its classification is equal to k, i.e., $d'_{ij} = k$. First, let us calculate the number $supp_i(k)$, of characteristic vectors supporting each of the possible classes $k \in K$ for a given i^{th} person. For the class of Parkinson's disease, i.e., for $k = 1$ we define its support as follows:

$$supp_i(1) = \sum_{j=1}^{card(S_i)} d'_{ij} \tag{1}$$

and for the vectors indicating a healthy person, i.e., for $k = 0$ the support is defined by means of the formula:

$$supp_i(0) = card(S_i) - supp_i(1) \tag{2}$$

Second, for the final medical diagnosis we propose to apply the classification criterion Eq. (3) that selects the class with the highest support:

$$d'_i = \arg\max_{k \in \{0,1\}} supp_i(k) \tag{3}$$

This way, the value of d'_i is assigned to the patient P_i as the predicted medical diagnosis of the Parkinson's disease. At this stage of research, we assume that in the case when: $supp_i(1) = supp_i(0)$, the value of k is assumed that is assigned to the characteristic vector that occurs as first in the source data file.

After further medical investigations and during medical therapy the doctor verifies the predicted diagnosis. The predicted diagnose d'_i is confronted with the real one d_i. The classification error for the i^{th} patient is calculated from the formula:

$$e_i = |d'_i - d_i|. \tag{4}$$

3. IMPLEMENTATION

The classification procedure described in Section 2 has been implemented using KNIME experimentation platform [9]. Due to the two steps of the proposed classification technique, the cross validation meta node available in KNIME could not be used for our purposes. The cross validation performed with respect to patients required to design a new workflow. The newly designed, annotated workflow is shown in Fig. 1. The workflow constructs two streams

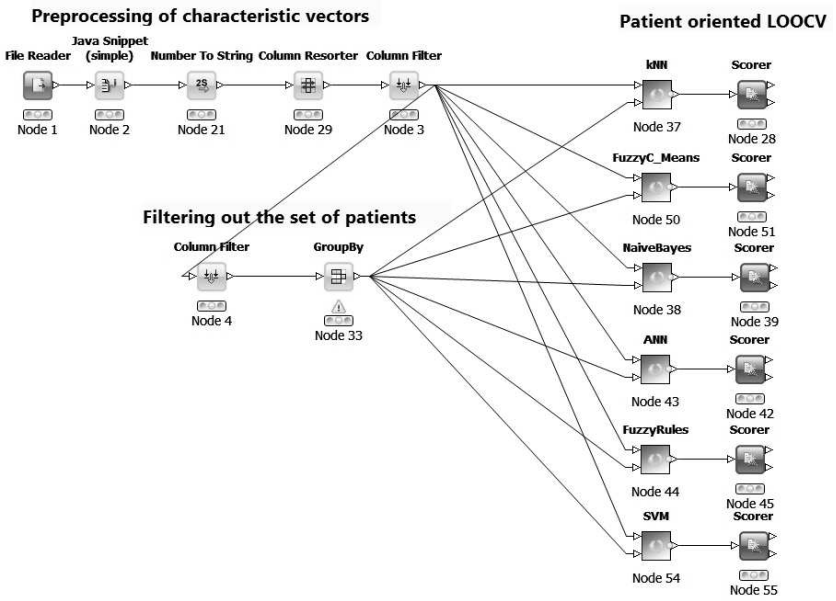

Fig. 1. Basic workflow of the classification procedure.

of data. The first contains the original set of characteristic vectors read from the text file. The second stream contains a set of labels identifying uniquely every person. Both of those data streams are an input to every of the meta nodes related to the cross validation performed with respect to persons. An exemplary meta node for the person-related cross validation and Naive Bayes classifier is shown in Fig. 2.

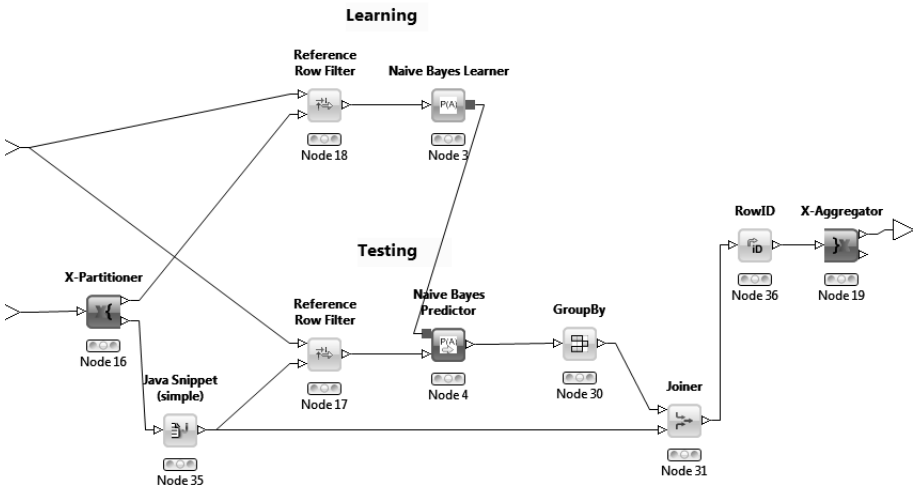

Fig. 2. Exemplary meta node for the patient-oriented cross validation.

For the comparative experiments the following state-of-the art classifiers playing the role of the model M (see Section 2) have been selected: k-nearst neighbour (kNN) and Fuzzy c-Means

(FCM) lazy classifiers, naive Bayesian (NB) statistical classifier, artificial neural networks (ANN) blackbox classifier, fuzzy rule-based system (FRBS) and support vector machine (SVM) classifier . Learning algorithms and the parameters for the considered classifiers were selected on the trial-and-error basis.

4. EXPERIMENTS

For all experiments we used real-world data from the UCI repository [10]. The considered data are easily available and usually used in the literature for benchmarking. As mentioned in introduction, the approach presented in this paper is investigated for the first time, therefore the results presented in this section are not directly comparable to those obtained in the existing literature.

4.1. SOURCE DATA

The data file contains $card(S) = 195$ characteristic vectors, where 147 of them are assigned to sick persons and 48 to healthy persons. The a-priori distribution of classes among characteristic vectors is: $0.7538/0.2462$. The characteristic vectors are distributed not equally between persons. It is worth to mention here that the numbers hidden in person's identifiers does not map directly to the sequence of ordinal numbers that probably led to several mistakes that can be found in the relevant literature. For that reason the data had to be analyzed with care.

Among 32 persons for which the data were recorded, 24 are sick and 8 are healthy. This means that the a-priori distribution of classes among persons is: $0.75/0.25$. The difference between distributions of classes for characteristic vectors and for patients comes from the fact of varying number of characteristic vectors assigned to patients. To the most of the patients 6 characteristic vectors are assigned with the exception of patients identified as S21, S27 and S35 to whom 7 vectors are assigned.

There are 23 attributes describing the data. One attribute contains the string that is a person's identifier. One binary attribute determines the class to which the characteristic vector is classified. The dimension of search space in which the classification is made is equal to 21, that is the number of real valued attributes. The description of characteristic features used during the calculation of real valued attributes is given in Table 1.

Table 1. Characteristic features of voice samples.

Feature	Description
MDVP:Fo(Hz)	Average vocal fundamental frequency.
MDVP:Fhi(Hz)	Maximum vocal fundamental frequency.
MDVP:Flo(Hz)	Minimum vocal fundamental frequency.
MDVP:Jitter(%), MDVP:Jitter(Abs), MDVP:RAP, MDVP:PPQ, Jitter:DDP	Several measures of variation in fundamental frequency.
MDVP:Shimmer, MDVP:Shimmer(dB), Shimmer:APQ3, Shimmer:APQ5, MDVP:APQ, Shimmer:DDA	Several measures of variation in amplitude.
NHR, HNR	Two measures of ratio of noise to tonal components in the voice.
RPDE, D2	Two nonlinear dynamical complexity measures.
DFA	Signal fractal scaling exponent.
spread1, spread2, PPE	Three nonlinear measures of fundamental frequency variation.

4.2. ACCURACY OF CLASSIFICATION

Due to the low number of samples (32 persons) we decided to apply for learning and testing the Leave-One-Out Cross Validation (LOOCV) procedure [1], [2]. Table 2 shows the best results obtained for every classifier used during the first step of classification. All values of accuracies were calculated with respect to the population of persons (not individual characteristic vectors). As can be noted in Table 2, the same best accuracy rate was obtained for the naive Bayesian classifier and for the fuzzy rule-based system. The winning Bayesian classifier was trained using 'KD tree' algorithm applying Euclidean distance. In the case of FRBS, the 'Mixed fuzzy rule formation' learning algorithm with the min/max norm was used [4].

Table 2. Classification results.

Classifier	Accuracy
kNN	0.781
FCM	0.750
NB	0.812
ANN(1,5)	0.750
FRBS	0.812
SVM	0.781

Although NB and FRBS systems achieved the same classification accuracy, the distribution of errors in both cases differed. As can be seen in Tables 3 and 4, for both winning classifiers, 26 persons were correctly classified. However, the Bayesian classifier was slightly better in classifying healthy persons (5 for NB / 4 for FRBS) for which the sub-population was much lower. For the sick persons the FRBS classifier was better (21 for NB/ 22 for FRBS).

Table 3. Confusion matrix for NB.

X	1	0
1	21	3
0	3	5

Table 4. Confusion matrix for FRBS.

X	1	0
1	22	2
0	4	4

The obtained results are encouraging providing evidence for the usefulness of the proposed method, however further research is required towards enhancing the method and improving its accuracy.

5. FINAL REMARKS

In this paper a new approach for the diagnosis of Parkinson's disease has been proposed. Instead of classifying only voice samples, a new two-step classification procedure has been proposed. In the first step of the proposed approach every individual voice sample was classified, in the second step on the basis of already classified voice samples assigned to a person, a final medical diagnosis was produced. This way the proposed procedure took into account the distribution of voice samples with respect to persons. This way the proposed procedure

is goal oriented aiming at the effective diagnosing of every person. Preliminary experiments demonstrated quite high, encouraging classification accuracy, especially when the Naive Bayes and the Fuzzy Rule-Based System was applied during the first step of the method. Further research is required to improve the obtained accuracy. Also more experiments are necessary to investigate the practical usefulness of the method.

BIBLIOGRAPHY

[1] DIETTERICH T., Approximate statistical tests for comparing supervised classification learning algorithms, Neural Computation, 1998, Vol. 10, pp. 1895–1923.

[2] EFRON B., Estimating the error rate of a prediction rule: improvement on cross-validation, Journal of American Statistical Association, 1983, Vol. 78, pp. 316–33.

[3] ENE MARIUS, Neural network-based approach to discriminate healthy people from those with Parkinson's disease, Annals of the University of Craiova, Math. Comp. Sci. Ser., 2008, Vol. 35, pp. 112–116.

[4] GABRIEL T. R., BERTHOLD M. R, Influence of fuzzy norms and other heuristics on mixed fuzzy rule formation, International Journal of Approximate Reasoning, 2004, Vol. 35, Issue 2, pp. 195202.

[5] GOLBE L. I., MARK M. H., SAGE J. I., Parkinson's disease handbook, 2010.

[6] HARIHARAN M., POLAT KEMAL, SINDHU R., A new hybrid intelligent system for accurate detection of Parkinson's disease, Computer Methods and Programs in Biomedicine, 2014, Vol. 113 (3), pp. 904–913.

[7] HUI-LING CHEN, CHANG-CHENG HUANG, XIN-GANG YU, XIN XU, XIN SUN, GANG WANG, SUJING WANG, An efficient diagnosis system for detection of Parkinson's disease using fuzzy k-nearest neighbor approach, Expert Systems with Applications., 2013, Vol. 40 (1), pp. 263–271.

[8] HUSE D. M., SCHULMAN K., ORSINI L., CASTELLI-HALEY J., KENNEDY S., LENHART G.. Burden of illness in parkinson's disease, Mov Disord., 2005, pp. 1449–1454.

[9] KNIME, `http://www.knime.org`.

[10] LITTLE MAX A, MCSHARRY PATRICK E, ROBERTS STEPHEN J, COSTELLO DECLAN AE, IRENE M, COSTELLO DECLAN AE, MOROZ IRENE M, Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. 2007.

[11] PARKINSON J., An essay on the shaking palsy, 1817.

[12] ROUZBAHANI HAMID KARIMI, DALIRI MOHAMMAD REZA, Diagnosis of Parkinsons disease in human using voice signals. Basic and Clinical Neuroscience, 2011, Vol. 2 (3), pp. 12–20.

[13] SANG-HONG LEE, JOON S. LIM, Parkinson's disease classification using gait characteristics and wavelet-based feature extraction, Expert Systems with Applications, 2012, Vol. 39 (8), pp. 7338–7344.

[14] TSANAS A., LITTLE M. A., FOX C., RAMIG L.O., Objective automatic assessment of rehabilitative speech treatment in Parkinson's disease, IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2014, Vol. 22 (1), pp. 181–190.

[15] TSANAS A., LITTLE M. A., MCSHARRY P. E., SPIELMAN J. L., RAMIG L. O., Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease, IEEE Trans. Biomed. Engineering, 2012, Vol. 59 (5), pp. 1264–1271.