UTILIZING PROCESSED RECORDS OF PATIENT´S SPEECH IN DETERMINING THE STAGE OF PARKINSON´S DISEASE

Michal VADOVSKÝ, Ján PARALIČ
Department of Cybernetics and Artificial Intelligence, Faculty of Electrical Engineering and Informatics,
Technical University of Košice, Letná 9, 042 00 Košice, Slovak Republic,
e-mail: michal.vadovsky@tuke.sk, jan.paralic@tuke.sk

ABSTRACT

The medical procedures for disease diagnostics are significantly demanding and time-consuming. Data mining methods can accelerate this process and assist doctors in making decisions in complex situations. In case of Parkinson´s disease (PD), the diagnostics of the initial disease stage is the primary issue since the symptoms are not so unambiguous and easily observable. Therefore, this article is focused on determining the actual stage of PD based on the data recording signals of patient´s speech using decision trees (C4.5, C5.0 and CART). Methods such as RandomForest, Bagging and Boosting were also employed to improve the existing classification models. Estimation of model accuracy was achieved by using k-fold cross-validation and validation with omission of one record (Leave-one-out). In addition, experiments were also performed to remove collinearity in data by computing the Variance inflation factor (VIF) in order to increase the accuracy of the models.

Keywords: *speech, stage, Parkinson´s disease, correlation, data mining*

1. INTRODUCTION

Parkinson´s disease (PD) [1] is a very serious neurological disease which we cannot cure yet. The main cause of the disease is the dying of nerve cells that produce an important chemical in the brain called dopamine [2]. Scientists have found out that about 400 000 nerve cells (neurons) are producing dopamine at birth. Every year, about 6% of them die in a healthy adult, but in Parkinson´s this loss is accelerated. When dropping below 20% of the original level, the first clinical manifestations will appear. The primary symptoms in people suffering from PD include muscle stiffness, speech problems (dysphonia), movement or writing (dysgraphia).

For the disease progression was developed a uniform PD rating scale named UPDRS (Unified Parkinson´s Disease Rating Scale) [3]. It was created during the 1980s and since its inception has become the most widely used scale for clinical evaluation of PD. UPDRS is a form of questionnaire where the patient responds to the questions on the scale with 5 options and is divided into four parts: I. Part – Thinking, behaviour and mood; II. Part – Activities of ordinary life; III. Part – Investigation of momentum; IV. Part – Treatment complications in the last week. Based on UPDRS results, it is possible to divide (according to Hoehnon and Yahr) patients into the 8 stages of PD by means of a modified stage scale:

- **Stage 0** – free of disease symptoms,
- **Stage 1 –** one-sided one-sided symptoms,
- **Stage 1.5** – one-sided and axial handicap,
- **Stage 2** – two-sided disability without equilibrium,
- **Stage 2.5** – two-sided disability with mild disturbance of balance, ability to balance attitude,
- **Stage 3** – mild to moderate two-sided disability, self-sufficient,
- **Stage 4** – severe disability, able to walk or stand without help,
- **Stage 5** – reliant on a wheelchair or attached to the bed, standing up with help.

Over time, shortcomings of UPDRS have been identified. For example, the scale did not cover the whole spectrum of non-motor expressions, or several ambiguities were found in the text and unambiguous instructions for the investigator were not specified. Therefore, a new version of the UPDRS scale was built in 2008, preserving the strengths of the original scale and eliminating identified shortcomings. The new version is called MDS-UPDRS (Movement Disorder Society – UPDRS) [4] and consists of 4 parts that were partly internally reorganized: I. Part – non-motoric aspects of daily life; II. Part – motor aspects of daily life; III. Part – motor investigation; IV. Part – motor complications.

Parkinson´s disease is still considered incurable. Nor is the cause of its origin known exactly. It is most likely a combination of genetic and environmental factors [5], but no conclusions are available at the end. One of the major causes of the disease is the dying of dopamine-producing nerve cells as already mentioned in the beginning. With its wasting in the brain, the disease of the patients increases. Methods for treating Parkinson´s disease differ in the early and late stages of the disease and should therefore be considered separately. The symptoms of PD are also specific for other diseases. The patient is diagnosed with PD only after excluding other causes of the disease and after detecting its symptoms. There is no cure in the treatment that can cure the disease. The most successful is pharmacological treatment combined with specific non-pharmacological procedures. The most basic and most effective drug is levodopa (L-dopa) [6], which is converted to dopamine in the brain. Levodopa has been considered the most effective antiparkinsonian agent to date. Although it is important in the treatment of PD, it may have side effects such as blood pressure fluctuations, nausea and vomiting.

In our article, we focused on determining the stage of PD using transformed data from voice recordings. Diagnosis of PD at an early stage is very challenging, because the symptoms are ambiguous and harder to recognize. That´s why we´ve tried to find out in this article, how much accuracy we can get using models that classify patient records into 8 different classes. In the case of high precision, these models could be implemented into systems to support PD diagnostics for physicians.

2. OVERVIEW OF CURRENT STATE

Since PD is a very common disease and there is still no medicine, several researchers are focused on diagnosing this disease directly from the initial symptoms, such as speech or writing. The main reason why the diagnosis of Parkinson´s disease due to the speech disorder is popular is that telediagnosis and telemonitoring systems based on speech signals are low cost and simple for their own use [7]. These systems reduce the inconvenience and cost of physical visits of patients to the clinic, enable early diagnosis of this disease and reduce the burden of healthcare personnel.

In the publication [8], G. Yadav et al. focused on the speech of patients and they used 3 data mining methods (Decision Trees, Logistic Regression and Support Vector Machine) to create models for classifying patients into two classes (1 – patient with PD, 0 – healthy patient). To calculate the accuracy, the authors used a 10-fold cross-validation to obtain 10 contingency tables for each of the methods used. For each of these tables, they calculated success, sensitivity and specificity, and ultimately averaged the results thus obtained. The best results were achieved using the Support Vector Machine with 76% accuracy, followed by the decision tree method (75%) and the logistic regression with 64% success rate.

A. Tsanas et al. [9] focused on prediction of the numerical value of UPDRS (0-176). The collected data also included patients´ speech signals and, in addition to the overall UPDRS value, they also focused on predicting the range of motor functions of the patient (0-108). All data contained 5 875 voice recordings and 22 attributes. The authors of this publication examined the potential of persistent vowel pronunciation to predict the motor and total UPDRS values using three linear and one nonlinear regression methods. They chose the optimally reduced subset of attributes that created a useful model where each attribute in the subset extracted the overlapping physiological properties of the speech signal. The UPDRS prediction error was measured by a mean absolute error (MAE) that was relatively low. This has shown that persistent vowel pronouncement provides enough information to predict UPDRS attributes. Based on the results, we can predict motor UPDRS values within approximately 6 points (full range reaches 108 points) and total UPDRS within the range of 7.5 points (full range is 176 points). These results reflect the best predictive error estimate for 1000 runs of 10-fold cross validation. Final predictions of values using models are very close to medical observations at the clinic.

3. UNDERSTANDING AND PREPARING DATA

The data we worked with are freely available at UCI Machine Learning Repository [10] and consist of a total of 31 biomedical voice measurements (of which 23 with PD). A total of 195 records (rows) were available because each patient had multiple records in the data that were taken independently of each other. The patient´s speech has been transformed into 22 attributes:

- **Basic voice frequency** – average average (MDVP:Fo(Hz)), maximum (MDVP:Fhi(Hz)) and minimum (MDVP:Flo(Hz)) basic voice frequency.
- **Jitter** – describes frequency instability and is often used as a parameter for measuring voice quality. We can define it as a short-term irregularity of the lengths of each speech signal period, with multiple variants and modifications of this parameter [11] – MDVP.Jitter.% (Jitter Percent), MDVP.Jitter.Abs (Absolute Jitter), MDVP.RAP (Relative Average Perturbation), MDVP.PPQ5 (Pitch Period Perturbation Quotient), Jitter.DDP.
- **Shimmer** – describes the amplitude instability of the analysed signal. Like by Jitter also with Shimmer, there are several variants that describes this phenomenon as well as the number of periods included in the analysis [12] – MDVP. Shimmer (Shimmer Percent), MDVP.Shimmer.db, Shimmer.APQ3 (Amplitude Perturbation Quotient), Shimmer.APQ5, Shimmer.DDA.
- **NHR** (Noise to Harmonic Ratio) – the ratio of noise to the harmonics of the signal. It is the total duration of the noise divided by the duration of the signal. For a healthy individual, NHR ranges around 0.005.
- **HNR** (Harmonic to Noise Ratio) – ratio of harmonic components to noise [db]. It determines the amount of noise in the signals, with a healthy person having a HNR value of about 20 [db] for a vowel “a”.
- **RPDE** (Method for determining the periodicity or repeatability of the signal), DFA (Signal fractal scaling), Status (Binary attribute and tells if the individual has PD (1) or not (0)), Spread1 / Spread2 (Nonlinear measurement of fundamental change of frequency), D2 (Nonlinear dynamic complexity of measurement) and PPE (Nonlinear measurements of fundamental change of frequency).

We have added additional attributes to that dataset which we found in the scientific article [13]. They contained additional information about Sex, Age, Stage of PD (8 levels). After adding the attribute Stage, we removed the attribute Status that provides information about whether a patient is suffering from PD (1) or not (0). Patient data with the worst 5th Stage were unavailable. After total data editing (merging datasets, removing missing values), we worked with 189 rows and 25 columns.

To better understand the available data and to work with the data, we first determined the individual dependencies on the target attribute. The data contain predominantly numeric values. As a target attribute we chose Stage (PD stage) and we calculated its dependence on numeric attributes (23 attributes). We used One-way ANOVA test [14] to calculate the dependency of numerical attributes against the nominal attribute. In this test, we mainly look at a p-value which, if less than the significance level (alpha = 0.05), rejects the hypothesis H0, accepts H1 and claims that there are differences in the average values of the selected numerical attributes divided by the nominal attribute (in our case Stage), e.g. that the factor Stage affects the average values of the selected attribute. The lowest p-value and hence acceptance of the H1 hypothesis with the highest confidence were obtained with the attributes MDVP.APQ (2.519407e-29), MDVP.Shimmer.db (4.454684e-23) and MDVP.PPQ (5.885382e-23). According to the obtained p-values we can see that these are really strong dependencies between the given numerical attributes and the target attribute Stage. In Fig.1 is presented a graph showing the strongest dependence between the target attribute and the numerical attribute MDVP.APQ. In this figure, we can notice that the increase in the PD stage usually increases also the average values of the attribute MDVP.APQ (rough line in the boxplot), the only exception is at stage 1.5 where a more pronounced increase in values compared to stage 1, even higher that at stage 2 and 2.5. Therefore, we can assume that when classifying patients into individual stages, the created models can be most mistaken in assigning patients to stage 1.5.

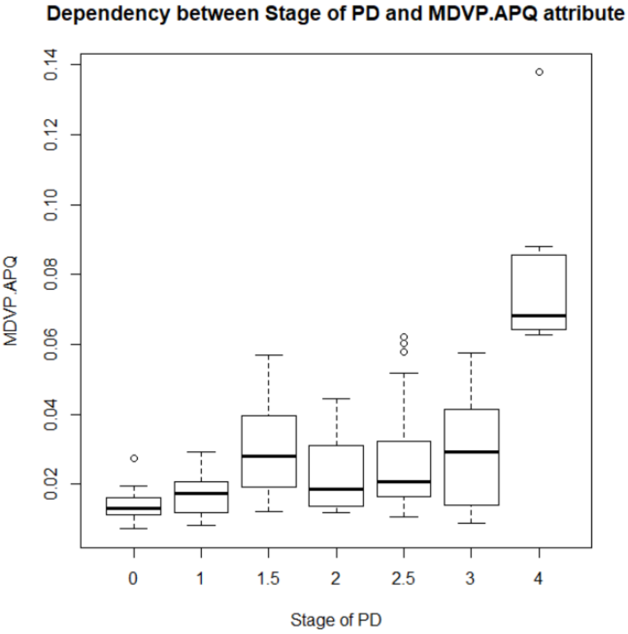


Fig. 1 Dependency between Stage of PD and MDVP.APQ

4. CLASSIFICATION MODEL

To create the classification models, we chose the decision tree methods C4.5, C5.0 and CART, which were obtained the highest precision from several algorithms. We used 10-fold cross validation (10-CV) and validation where one record is omitted (Leave One Out – LOO). In LOO is a similar way of evaluating than in the case of multiple cross validation, but the classifier is built on n-1 records in a dataset and is tested for only one record. This process is then repeated n times, so it is often too slow to calculate the load. In order to create the models, we first selected all attributes and later we cleaned data from collinearity (2 or more attributes are heavily dependent on each other), which may degrade the accuracy of the models [15]. The simplest way to detect collinearity is to create a correlation matrix and, if the absolute value of the selected element in this matrix is high, we are talking about highly correlated variables. Although no pair of variables has an exceptionally high absolute correlation value, there may still be collinearity between three and more variables. Therefore, for calculating the collinearity, it is better to calculate the variance inflation factor (VIF) for each attribute. The smallest possible VIF value is 1, which represent the total absence of collinearity. The rule is that if the VIF value exceeds 5 or 10, we are talking about the problematic amount of collinearity. Therefore, we have removed attributes with a VIF value greater than 5 and left the attributes shown in Table 1, which are arranged according to the variance inflation factor (VIF) from the highest value to the smallest.

Table 1 Assigned attributes by VIF

Rank	Attribute	VIF
1.	MDVP.APQ	**3.716**
2.	Spread1	**3.685**
3.	MDVP.Fo.Hz	**3.649**
4.	RPDE	**2.537**
5.	NHR	**2.426**
6.	DFA	**2.363**
7.	Spread2	**2.351**
8.	D2	**2.248**
9.	MDVP.Flo.Hz	**2.026**
10.	Age	**1.811**
11.	MDVP.Fhi.Hz	**1.372**

In addition, we have also retained the nominal attributes Gender and the Stage of PD (Stage). In such a way out of the total of 25 attributes, their number has been reduced to 13. The following table compares the accuracy with standard deviation of the created models using the algorithms CART, C4.5, C5.0 using all attributes and choosing only 13 attributes (11 with VIF < 5, Gender and Stage). We also compared the methods for selecting records to the training and test set using 10-CV and LOO-CV.

Table 2 Results of the 10-CV

Selection of attributes	10 - CV		
	CART	*C4.5*	*C5.0*
All attributes	69.71% ±13.49%	79.88% ±8.55%	83.54% ±7.02%
VIF < 5	67.66% ±12.56	85.15% ±6.95%	**86.2% ±6.34%**

Table 3 Results of the LOO-CV

Selection of attributes	10 - CV		
	CART	*C4.5*	*C5.0*
All attributes	71.96% ±45.04%	80.42% ±39.78%	81.48% ±38.95%
VIF < 5	72.47% ±44.78%	80.95% ±39.37%	82.01% ±38.51%

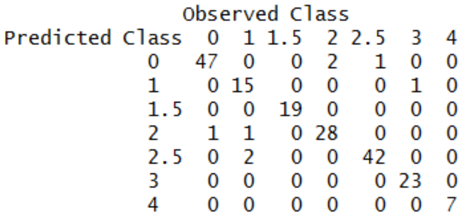
From the results in Table 2 and Table 3, we can notice that the removing high collinear attributes has ensured higher accuracy of models in almost all cases. The exception was the algorithm CART, where accuracy was reduced from 69.71% to 67.66%, but these accuracies are still significantly lower compared to the algorithms C4.5 and C5.0. The highest accuracy in all cases was achieved by the algorithm C5.0, while at the removal of the high collinearity attributes and the use of 10-CV we achieved an accuracy of 86.2% for the classification patients in 7 levels.

In order to improve the accuracy of the created models, we decided to use RandomForest, Bagging and Boosting methods that use trees as building blocks to create stronger prediction models. RandomForest creates multiple decision trees, where in each tree when choosing a test attribute, we take into account m randomly selected attributes of the total number of p. The resulting classification in the class is voted by all the generated trees. If all p attributes are taken into account at that note, then we are talking about bagging. The boosting method works in a similar way, but each decision tree also takes into account the information from the previous tree [16]. Records that have been incorrectly classified in the previous tree are assigned a greater weight in the next iteration, which will place greater emphasis on subsequent iterations. In the publication [17], the authors report that with the increasing number of generated trees, only the computational burden is increased and the differences in accuracy are already very small. Their analysis of 29 datasets showed that 128 trees were no longer a significant difference in accuracy than 256, 512, 1024, 2048 and 4096 trees. Therefore, we have set the number of decision trees to 50, 100 and 150 and for the accuracy calculation, we have only selected 10-fold cross validation.

Table 4 Results of RandomForest, Bagging and Boosting

Number of trees	RandomForest	Bagging	Boosting
m = 50	87.25% ±8.37%	77.78% ±9.25%	**93.65% ±6.51%**
m = 100	86.73% ±8.52%	77.78% ±9.25%	**95.24% ±5.01%**
m = 150	**86.73% ±8.52%**	**78.31% ±9.12%**	**95.77% ±4.95%**

The results in Table 4 clearly show that the highest accuracy achieved the Boosting method when 150 trees were generated (95.77%). With the growing number of trees, the accuracy of classification has grown with both Bagging and Boosting methods, but on the other hand the accuracy of RandomForest model decreased slightly. The algorithm C5.0 at 10-CV achieved better accuracy than Bagging in this case and comparable with RandomForest.



```
                  Observed Class
Predicted Class   0  1 1.5  2 2.5  3  4
            0    47  0   0  2   1  0  0
            1     0 15   0  0   0  1  0
            1.5   0  0  19  0   0  0  0
            2     1  1   0 28   0  0  0
            2.5   0  2   0  0  42  0  0
            3     0  0   0  0   0 23  0
            4     0  0   0  0   0  0  7
```

Fig. 2 Contingency table of Boosting method

For the Boosting method, which achieved the highest accuracy at 95.77%, we have shown in Figure 2 also a contingency table that compares the predicted value of attribute Stage with those identified and given in the test set.

Because 10-fold cross validation was used for model evaluation, testing was performed on 10 different test sets. Each element in the contingency table in Fig. 1 is calculated as the sum of all the contingency tables obtained with which it is clear to see what prediction was the most common error. The main goal is to maximize the values on the main diagonal in the matrix, which is the correct prediction of the given stage of Parkinson´s disease. Out of the total of 189 records, the model accurately predicted in 181 cases, representing 95.77% accuracy. For each stage we have achieved the following accuracy (in brackets is the ratio of correctly classified records to all records for a given PD stage):

- Stage 0 (47/48) – 97.92%
- Stage 1 (15/18) = 83.33%
- Stage 1.5 (19/19) = 100%
- Stage 2 (28/30) = 93.33%
- Stage 2.5 (42/43) = 97.67%
- Stage 3 (23/24) = 95.83%
- Stage 4 (7/7) = 100%

We can note that with 100% accuracy the model was able to predict stages 1.5 and 4. On the other hand, the lowest accuracy at 83.33% reached the first stage of PD, where the model knew in 15 records the correct stage and 3 errors occurred (in one case predicted Stage 2 and in two cases stage 2.5). In Fig. 1, we noticed that the created models may be most mistaken in predicting stage 1.5 because the attribute values with the highest dependence of MDVP.APQ were higher than in stage 2, 2.5 and 3. This unexpected increase could also have made errors in predicting these additional stages of Parkinson's disease. In the end, the model achieved almost all successes over 93%, with the only exception being stage 1, when the model reached only 83.33%.

5. CONCLUSION

In this article we focused on determining the stage of patients with PD from their speech using data mining methods. Already in the first experiment and the elimination of collinearity in data, we reached an accuracy of 86.2% using the decision tree and the algorithm C5.0 (before removing collinearity – 83.54%). By using the Boosting method, which creates multiple decision trees, we can increase our accuracy to 95.77% (at m = 150), which is a high number due to classification of records up to 7 levels. For example, in the publication [8] for binary classification (1 – patient with PD, 0 – healthy patient) with the same data, the authors achieved the highest accuracy only 76% using SVM. We also have a binary classification of patients in our previous publication [18] and the best result at 91.43% was achieved using the algorithm C4.5. Although it was a more complex classification of patients (7 levels) compared to binary classification (2 levels), we have achieved a higher accuracy of the model using Boosting method.

6. FUTURE WORK

In the future work, we would we would like to focus on processing spoken speech into the same attributes to create an application where people upload their speech and can tested themselves. We would also like to focus on the data we obtained from mPower: Mobile Parkinson´s Disease Study. They capture patient demographics, as well as data about their voice, walk, memory and tapping on the screen of the mobile. Thanks to this data, we could expand our research into multiple areas and symptoms of Parkinson´s disease.

We also work with MUDr. Škorvánek, who provided us with a sample of 3206 patients with PD described by 207 attributes. On this data, we aim to focus on analysing the remaining questionnaires (MDS-UPDRS, PDQ8, PDQ39, EQ_5D_3L), comparing them and searching for hidden relationships between individual items using hierarchical regression analysis and LASSO. We have already compared the MDS-UPDRS and PDQ39 questionnaires, but the results are not yet published.

ACKNOWLEDGMENTS

This work was supported by the Slovak Research and Development Agency under the contract No. APVV-16-0213 and by the Slovak Grant Agency of the Ministry of Education and Academy of Science of the Slovak Republic under grant No. 1/0493/16.

REFERENCIES

[1] DE LAU, L. M. – BRETLER, M. M.: Epidemiology of Parkinson´s disease, *The Lancet Neurology*, Vol. 5, No. 6, pp. 525-535, Jun. 2006.

[2] CNOCKAERT, L. et al.: Low-frequency vocal modulations in vowels produced by Parkinsonian subjects, *Speech Communication*, Vol. 50, No. 4, pp. 288-300, Apr. 2008.

[3] FANS, S. et al.: Members of the UPDRS Development Committee Unified Parkinson´s Disease Rating Scale, *Recent developments in Parkinson´s disease*, Vol. 2, pp. 153 – 163, 1987.

[4] GOETZ, CG. et al.: Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Process, format, and clinometric testing plan, *Movement Disorders,* Vol. 22, No. 1, pp. 41 – 47, Jan. 2007.

[5] ZVONČEKOVÁ, L.: Ľudí s s Parkinsonovou chorobou dokáže liečba vrátiť do života, 2015. Available at: < http://zdravie.pravda.sk/zdravie-a-prevencia/clanok/377609-ludi-s-parkinsonovou-chorobou-dokaze-liecba-vratit-do-zivota>.

[6] Parkinson.sk – Farmakologická liečba. Available at: < http://www.parkinson.sk/Liecba/i.folder.aspx>.

[7] SINGH, N. – PILLAY, V. – CHOONARA, Y. E.: Advances in the treatment of Parkinson's disease, *Progress in Neurobiology*, Vol. 81, No. 1, pp. 29 – 44, Jan. 2007.

[8] YADAV, G. – KUMAR, Y. – SAHOO, G.: Predication of Parkinson's disease using data mining methods: a comparative analysis of tree, statistical, and support vector machine classifiers, *Indian Journal of Medical Sciences*, Vol. 65, No. 6, pp. 231 – 242, Jan. 2011.

[9] TSANAS, A. et al.: Accurate Telemonitoring of Parkinson's Disease Progression by Noninvasive Speech Tests, *IEEE Transactions on Biomedical Engineering*, Vol. 57, No. 4, pp. 884 – 893, Nov. 2009.

[10] UCI Machine Learning repository: Center for Machine Learning and Intelligent Systems – Parkinsons Data Set. Available at: <https://archive.ics.uci.edu/ml/datasets/Parkinsons>.

[11] PEŠTA, J. – KASL, Z. – VOHLÍDKOVÁ, M.: *Pooperační objektivní posouzení hlasu*. Plzeň: ZČU, FAV, 2004. Available at: <http://www.kiv.zcu.cz/~novyp/publ/foniatr04.pdf>.

[12] VYMLÁTIL, P.: Zjištění Parkinsonovy nemoci na základě analýzy řečového záznamu. Bakalářská práce, Brno: VUT FEKT, 2013. 53s.

[13] LITTLE, M. A. et al.: Suitability of Dysphonia Measurements for Telemonitoring of Parkinson's Disease, *IEEE Transactions on Biomedical Engineering*, Vol. 56, No. 4, pp. 1015 – 1022, Apr. 2009.

[14] SEBERA, M.: Vícerozměrné statistické metody: Analýza rozptylu. Available at: <http://www.fsps.muni.cz/~sebera/vicerozmerna_statistika/anova.html>.

[15] JAMES, G. et al.: An Introduction to Statistical Learning. Springer-Verlag, New York, 2013.

[16] FREUND, Y. – SCHAPIRE, R.: A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting, *Journal of Computer and System Sciences*, Vol. 55, No. 1, pp. 119 – 139, Aug. 1997.

[17] OSHIRO, T. M. et al.: How Many Trees in a Random Forest?, *Machine Learning and Data Mining in Patter Recognition,* Vol. 7376, pp. 154 – 168, Jul. 2012.

[18] VADOVSKÝ, M. – PARALIČ, J.: Predikcia Parkinsonovej choroby pomocou signálov reči použitím metód dolovania v dátach, *WIKT & DaZ 2016,* Bratislava: STU, 2016. Pp. 329 – 333. ISBN: 978-80-227-4619-9.

Received June 06, 2018, accepted September 05, 2018

BIOGRAPHIES

Michal Vadovský was born on 27. 06. 1992. In 2015 he graduated (MSc) with distinction at the Department of Cybernetics and Artificial Intelligence of the Faculty of Electrical Engineering and Informatics at Technical

University in Košice. Since September 2015 he works as PhD. student at the Department of Cybernetics and Artificial Intelligence. He has experience with data mining in the database using machine learning methods and different statistical methods. The dissertation for PhD study is focused on methods for the analysis of selected types in medical data. Currently, he is working on various analyses focusing on Parkinson´s disease. Additionally, he is member of the team working on evaluation of mild cognitive impairment based on the data gathered from patients. He also participates in various research projects.

Ján Paralič received his master's degree in 1992, his Ph.D. degree in 1998 and became associate professor in 2004 at the Technical University in Košice. Since 2012, he is a full professor and deputy head at the Department of Cybernetics and Informatics, Technical University in Košice. His research interests are in the areas of knowledge discovery, text mining, big data analytics, and knowledge-based approaches in information system