Machine learning and data mining methods in neurodegenerative diseases research

Artur Chudzik[1], Artur Szymański[1], Andrzej W. Przybyszewski,[1,2]
Kamila Bojakowska[3]

[1] Polish-Japanese Academy of Information Technology, Warsaw 02-008, Poland
{artur.chudzik, artur.szymanski, przy}@pjwstk.edu.pl

[2] University of Massachusetts Medical School, Dept. Neurology, Worcester, MA 01655, USA

[3] Department of Neurology, Central Clinical Hospital MSWiA, Warsaw 02-507, Poland
kama.435@wp.pl

Abstract

Neurodegenerative diseases, in which we could distinguish Parkinson Disease (PD), and Alzheimer Disease (AD) are still fields of medicine that are not fully discovered in the matter of causation. Currently, available tests for Alzheimer's and Parkinson's are expensive and invasive. Furthermore, the first stages of the disease progression show no symptoms. When the symptoms became noticeable, changes in the brain circuits are already on the very advanced level. Reliable, affordable, and accessible biomarkers have the potential to become a factor that would allow a better understanding of disease progression, more easily identify people for clinical trials, and more accurately monitor their response to treatments. The aim of the experiment will be the acquisition and analysis of potential biomarkers in the context of emotional states, enabling the prediction of the dynamics of the patient's disease, with the support of machine learning methods.

Keywords

Parkinson's disease, eye-tracking, FACS, data mining, machine learning

1 Introduction

Previous studies (Przybyszewski, et al., 2014) have shown that there is a possibility of prediction of PD progression based on tests conducted on patients that are under various methods of treatments and disease progression level. One of the research included the eye gaze movement test (saccadic movements measurements) on the three groups of patients divided by the treatment model: BMT patients (only medication), and patients on medication and with implanted electrodes in the subthalamic nucleus- during study: DBS group or before study: POP group. Data acquired from the eye movement registration has been combined with demographical details of patients and different neuropsychological tests (e.g. Epworth - sleep problems, or PDQ39 - quality of life). The goal was to predict disease progression according to the Unified Parkinson's Disease Rating Scale (UPDRS) which was successful and possible with the usage of Rough Sets and different supervised learning

models (Przybyszewski, et al., 2018). The outcome of that research was a proof of biomarkers relevance in the prediction of PD progression.

The research described in the following paper has a goal to extend previous findings by additional biomarkers that could be relevant in neurodegenerative disease progression. Along the eye tracking technique, that found to be a significant index, we introduced high-resolution camera and EEG combined with dedicated software which registers gaze movements, emotions recognition and brain activity under the different stimulus.

2 Methods

Central Clinical Hospital of the MSWiA in Warsaw, which participates in the described project, shared a laboratory in which research equipment has been installed. The installed workplace is de facto a high-performance desktop computer with a touch screen with high-speed USB camera aligned with Bluetooth EEG headband. Peripheral devices are combined with one ecosystem by dedicated software which synchronises raw data from sensors and displays a stimulus.

Data processing would be performed post factum due to its high capacity. The study requires only the controlled light environment. However, patients' position and research conditions should be as non-invasive as possible. The test should cover two separate groups of patients: those who suffer from PD and AD. Most of the patients would be in the pre-clinical stage, who do not have visible symptoms of illness.

The first part of the examination would be neuropsychological tests. Examined patients will be in pre-clinical (no symptoms) stages of ND disease, and our goal would be to describe their health condition in the form that could be measured and brought to the form understandable for the computers, which is not depended on doctors' subjective opinion.

To achieve that task, we will use a set of "core" tests that would build a general image of the patient and which results are countable. The original list includes below items. The selection was based on the examination results described in the Background section and has a direct reflection in progressive mental state degeneration revealed, inter alia, in depression and/or (depending on the disease) cognitive interference and motoric functions.

- Epworth sleepiness scale (ESS) questionnaire which is shown to provide a measurement of the subject's general level of daytime sleepiness (Johns, 1991),
- The Montreal Cognitive Assessment (MoCA), which is a brief measure of global cognitive function; originally developed to detect mild cognitive impairment (Nasreddine & et al., 2005),
- The Stroop Color and Word Test (SCWT) that assesses the ability to inhibit cognitive interference, which occurs when the processing of a stimulus feature affects the simultaneous processing of another attribute of the same stimulus (Stroop, 1935),
- Parkinson's Disease Questionnaire Short Form (PDQ-8), a disease-specific measurement of quality of life (Jenkinson, 1997),
- Beck Depression Inventory (BDI-II), one of the most widely used psychometric tests for measuring the severity of depression (Beck, et al., 1996),
- Trail Making Test (TMT), which provides information on the visual search, scanning, speed of processing, mental flexibility, and executive functions(Tombaugh, 2004).

The next step would be the setup of the patient at the research workstation. His/her first task would be to perform a gaze-related test which consists of visually guided saccades and antisaccades, as well

as pursuit eye movements, measurement by the high-speed USB camera with dedicated software(Naruniec, et al., 2016). This method, called video-oculography (VOG) is a tool providing diagnostic information about the progress of the diseases that cause regression of the vergence eye movements. We measure fast eye movements in response to a light spot switched on and off, moving horizontally from the straight eye fixation position (0°) to 10° to the left or 10° to the right after arbitrary times ranging between 0.5–1.5 s. When the patient fixated eyes on the spot in the middle marker (0°) the spot then changed colour from green to red, indicating a signal for the performance of AS. In each test, the subject had to perform 20 RS and 20 AS in a row.

In the next step, we will record EEG (Muse) which is relatively simple device. We had chosen that model for the sake of simplicity of its application. Muse and several other portable EEG devices are way simpler to set up than medical-grade EEG. They connect well via Bluetooth to a smartphone, a computer, or a microcontroller, where data can be analysed directly. Dry electrodes used in this device do not require intensive preparation or clean-up, and these electrodes connect to the skin without the need for any liquid (in opposition to professional applications with conductive gels that have to be washed from the head after the research). Muse EEG system has electrodes located analogously to Fpz, AF7, AF8, TP9, and TP10 with electrode Fpz utilised as the reference electrode. It means that it uses two channels on the left and two on the right, so it is possible for exploring hemispheric asymmetries.

The whole setup is supported by the USB camera (the same that has been used in eye-tracking) connected to a custom software prepared for emotions recognition task, which is based on OpenFace(Baltrušaitis, et al., 2018). OpenFace, therefore, relies on the Facial Action Coding System (FACS) (Ekman & Rosenberg, 1997). FACS is a system to taxonomise human facial movements by their appearance on the face. FACS encodes movements of individual facial muscles from slight different instant changes in facial appearance. That system allows to code nearly any anatomically possible facial expression, deconstructing it into the specific Action Units (AU) that produced the expression. It is a common standard to describe facial expressions objectively. As AUs are independent of any interpretation, they can be used for any higher order decision making process including recognition of basic emotions. OpenFace works on two models trained independently. The first model provides binary information- *Presence* - which determines if AU is visible in the face. The second model is called *Intensity*, and it describes the magnitude of the AU (minimal to maximal) on a 5 point scale. The Intensity and the Presence predictors have been trained separately, and on slightly different datasets thus the predictions of both might not always be consistent. Our model gives priority to the Presence binary factor which is crucial for single emotion expression description. Emotions description had been based on (Ghayoumi & Bansal, 2016).

Table 1 Set of action units needed for basic emotions

e_i	Emotion	Action Units
e_0	Surprise	{ "AU01", "AU02", "AU05", "AU15", "AU20", "AU26" }
e_1	Fear	{ "AU01", "AU02", "AU04", "AU05", "AU15", "AU20", "AU26" }
e_2	Disgust	{ "AU02", "AU04", "AU09", "AU15", "AU17" }
e_3	Anger	{ "AU02", "AU04", "AU07", "AU09", "AU10", "AU20", "AU26" }
e_4	Happiness	{ "AU01", "AU06", "AU12", "AU14" }
e_5	Sadness	{ "AU01", "AU04", "AU15", "AU23" }

We compute the overall intensity for each emotion class e_i based on the AU intensity vector V of the length v and AU presence vector P (of the same length) for each frame. The results are normalised in the 0..5 scale.

Equation 1 Emotion intensity calculation formula

$$e_i = \prod_{j=0}^{v} P(j) * \frac{\sum_{j=0}^{v} V(j)}{v}$$

The whole setup would be prepared for the registration of the patient's behaviour and reactions (conscious and subconscious) on the stimulus. There would be a task in the form of a video presenting the faces of people expressing different types of facial expressions. The experiment is based on the so-called emotional contagion paradigm (Hatfield, et al., 1993), which is a process of transferring emotions or moods between persons. The video contains a set of recordings of persons which are expressing different emotions, both positives and negatives. The patient would get a request for video watching, without mentioning any specific tasks, because we would like to receive as many natural and subliminal reactions as possible.

During the experiment, we would collect data automatically by the software, which allows us to synchronise samples from every device with the specific kind of stimulus shown on the screen. Our goal would be to register patients' reaction on, basic emotions expressions: anger, disgust, fear, happiness, sadness and surprise.

2.1 Background

The selection of psychological tests was implemented due to previous research which revealed that affective disturbances and changes in sleep-wake cycles have been recognised in both disorders and merit closer attention. Depression occurs in 40% of the patients with PD and similar percentages of patients with mild dementia due to AD. Greater loss of locus coeruleus (LC) neurons has been reported among patients with AD who are depressed and patients with PD by some investigators. Furthermore, the subset of patients with AD and depression had the combination of a relatively lower number of LC, but a comparatively higher number of nucleus basalis (NB) neurons (Zarow, et al., 2003).

The correlation between the PD/AD and the emotions have been investigated in previous studies (Braak & Braak, 1991) (Gray, et al., 2010) which revealed that the perception and expression of the emotions among the people who suffer from PD and AD are disturbed due to damage to the structures of the limbic system, including basal ganglia and amygdala.

The aligned study covers electroencephalography (EEG) registered from the prefrontal cortex because there is a direct relationship between this structure and amygdala (Purves, et al., 2001).

A study conducted by the vision system for emotion recognition will help to investigate the reaction of the extrapyramidal system and the primary motor cortex involved in the facial expressions, both intended and spontaneous (Rinn & William, 1984).

Finally, the eye-tracking research in the context of reflexive saccades had been found to be a very precise biomarker for assessment of symptom progression in PD (Przybyszewski, et al., 2016).

2.2 Machine learning and data analysis

After the experiment, we would use machine learning methods to conclude a correlation between standard psychological test results and the data from the sensors. This procedure would engage different fields of study, including EEG signal analysis, emotions recognition and intensification based on AUs and eye movements processing.

In the field of EEG signal analysis we are going to make use of a new and growing interest in signal processing community technique of empirical mode decomposition (EMD) which we extend to multichannel approach of parallel decomposition of single channel signals and further clustering of

so-obtained components among channels to track coherent (synchronized or correlated in spectral domain) activities in complex signals from electrodes (Rutkowski, et al., 2010).

The next step would be a task to find "ground truth" in data. To achieve that task, we would use various supervised-learning models that found to be valuable in PD/AD symptoms classifications. In brief, because our data contains a set of N training samples of the form $\{(x_1, y_1), ..., (x_n, y_n)\}$ such that x_i is the feature vector of the *i-th* sample and y_i is its class, it is possible to use a supervised learning algorithm which seeks for a function $g: X \rightarrow Y$, where the X is the input space, and the Y is the output space. The g function is an element of some space of possible functions G, known as the hypothesis space. Previous research has shown, that Logistic Regression (where probabilities describing the possible outcomes of a single trial are modeled using a logistic function), aligned with Linear Supported Vector Classification (in which different Kernel functions can be specified for the decision function), Decision Tree, Gradient Boosting Classifier (which provides ability for generalization of boosting to arbitrary differentiable loss functions), Random Forest, Gaussian Naive Bayes, K-Neighbors Classifier (KNC) or C-Support Vector Classification (SVC) gives good results in prediction of the disease symptoms.

Additionally, for symptoms classifications, we had found that Rough Set Theory gives notable results in the matter of PD. The theory assumes that data is represented as a decision table in which rows contains information about different measurement (which could be obtained from the same or different patients) and columns are bound with different attributes. An information system is defined as a pair $S = (U, A)$, where U, A are finite sets: U is the universe of objects; and A is the set of attributes. The value $a(u)$ is a unique element of V (where V is a value set) for $a \in A$ *and* $u \in U$.
A decision table for S is the triplet: $S = (U, C, D)$ where: C, D are condition and decision attributes. Each row of the information table gives a particular rule that connects condition and decision attributes for a single measurement of a specific patient. As there are many rows related to different patients and sessions, they generate many rules. The rough set approach allows generalising these rules into universal hypotheses that may determine markers for an individual patient on the matter of neurodegeneration progression.

The wide choice of algorithms is caused by the fact, that each has its strengths and weaknesses. No single learning algorithm works best on all supervised learning problems, according to "no free lunch theorem" (Wolpert & Macready, 1997). Because data analysis on that project is considering a new modelling application, we would compare multiple learning algorithms and experimentally determine which one works best on the problem by cross-validation. The proper tools and methodology have been developed in previous projects, and it's based on Python language supported with the scikit-learn toolkit (Pedregosa, et al., 2011).

3 Hypothesis

A typical Alzheimer's disease starts by affecting cells that are interconnected with the hippocampal formation with the association cortices, basal forebrain, thalamus, and hypothalamus, which are structures crucial to memory (Hyman, et al., 1984).

In Parkinson's disease, the asymmetrical process of degeneration was also observed, based, for example, on the studies of substantia nigra neuronal loss (Kempster, et al., 1989).

Thus, a neurodegenerative disease could have its reflection in the asymmetry of the expression of the emotions that could be measured and analysed by the machine learning methods. Possibly, the same pattern could be observed even in the pre-clinical stages of PD/AD.

Furthermore, according to the author's knowledge, there are no studies about eye movements distortion in pre-clinical stages of PD and AD, so the analysis of this specific group with eye-tracking tests would create unique sort of data and could create a new, relevant biomarker.

4 Perspectives

People vary substantially in their combination of symptoms, a rate of progression, and reaction to treatment. Averaging patients' symptoms on different stages of the disease give very rough approximations of results. If we want to improve such investigations, we need to take into account an extreme diversity of patients' symptoms and the light effects of cares and therapies in distinctive cases. These problems are well-described in the famous statement: "No two people face Parkinson's in quite the same way".

Nevertheless, the purpose of this study is to go even deeper into symptoms of the individuals that are in the pre-clinical stage of the disease. We are going to deal with the unknown land of attributes which are being so ambiguous and complex to detect even for the experienced neurologist. Substantial diversification of data mining and machine learning methods should ensure us therefore rules based on such subtle correlates that only specialised devices combined with artificial intelligence would be able to explore.

Our methods will suggest techniques that can be implemented in telemedicine solutions, enabling early detection of the progression of Parkinson's and Alzheimer's disease, as well as a determination of appropriate therapy and examination of its effectiveness after treatment implementation. A specific diagnosis, adjusted to different individual patients that may lead to early detection of illness slowing down their symptoms and improving their quality of life.

References

Baltrušaitis, T., Zadeh, A., Chong Lim, Y. i Morency, L.-P., 2018. OpenFace 2.0: Facial Behavior Analysis Toolkit. *IEEE International Conference on Automatic Face and Gesture Recognition*.

Bazan, J. G. i inni, 1998. Synthesis of Decision Rules for Object Classification. *Incomplete Information: Rough Set Analysis*, pp. 23-57.

Beck, A. T., Steer, R. A. i Brown, G. K., 1996. Beck depression inventory-II. *San Antonio*, 78(2), pp. 490-498.

Braak, H. i Braak, E., 1991. Alzheimer's disease affects limbic nuclei of the thalamus. *Acta neuropathologica 81.3*, pp. 261-268.

Critchley, H. i inni, 2005. Activity in the human brain predicting differential heart rate responses to emotional facial expressions. *NeuroImage*, pp. 751-62.

Ekman, P. i Rosenberg, E. L., 1997. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS).* USA: Oxford University Press.

Gray, M., H., Tickle-Degnen i Linda, 2010. A meta-analysis of performance on emotion recognition tasks in Parkinson's disease. *Neuropsychology 24.2*, p. 176.

Hatfield, E., Cacioppo, J. T. i Rapson, R. L., 1993. Emotional contagion. *Current directions in psychological science*, 2(3), pp. 96-100.

Jenkinson, C., 1997. The PDQ-8: development and validation of a short-form Parkinson's disease questionnaire. *Psychology and Health*, 12(6), pp. 805-814.

Johns, M., 1991. A new method for measuring daytime sleepiness: the Epworth sleepiness scale.. *sleep*, 14(6), pp. 540-545.

Kempster, P., Gibb, W., Stern, G. i Lees, A., 1989. Asymmetry of substantia nigra neuronal loss in Parkinson's disease and its relevance to the mechanism of levodopa related motor fluctuations.. *Journal of Neurology, Neurosurgery & Psychiatry*, 51(1), pp. 72-76.

Naruniec, J. i inni, 2016. Webcam-based system for video-oculography. *IET Computer Vision*, 11(2), pp. 173-180.

Nasreddine, Z. S. i et al., 2005. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, pp. 695-699.

Pedregosa, F. i inni, 2011. Scikit-learn: Machine Learning in Python. *JMLR 12*, pp. 2825-2830.

Przybyszewski, A. W., 2010. Logical rules of visual brain: From anatomy through neurophysiology to cognition. *Cognitive Systems Research 11*, pp. 53-66.

Przybyszewski, A. W. i inni, 2018. Rough set rules determine disease progressions in different groups of Parkinson's patients. *Fundamenta Informaticae (in review)*.

Przybyszewski, A. W. i inni, 2014. Data Mining and Machine Learning on the Basis from Reflexive Eye Movements Can Predict Symptom Development in Individual Parkinson's Patients. *Nature-Inspired Computation and Machine Learning*, p. 499–509.

Przybyszewski, A. W. i inni, 2016. Multimodal Learning and Intelligent Prediction of Symptom Development in Individual Parkinson's Patients. *Sensors*, 16(9).

Purves, D., Augustine, G. i Fitzpatrick, D., 2001. *The Relationship between Neocortex and Amygdala.* Sunderland: Sinauer Associates.

Rinn i William, E., 1984. The neuropsychology of facial expression: a review of the neurological and psychological mechanisms for producing facial expressions. *Psychological bulletin 95.1*, p. 52.

Rutkowski, T. M., Mandic, D. P., Cichocki, A. i Przybyszewski, A. W., 2010. EMD approach to multichannel EEG data—the amplitude and phase components clustering analysis. *Journal of Circuits, Systems, and Computers*, 19(01), pp. 215-229.

Stroop, J. R., 1935. Studies of interference in serial verbal reactions. *Journal of experimental psychology*, 18(6), p. 643.

Toga, A. W. i Thompson, P. M., 2003. Mapping brain asymmetry. *Nature Reviews Neuroscience*, 4(1), p. 37.

Tombaugh, T. N., 2004. Trail Making Test A and B: normative data stratified by age and education. *Archives of clinical neuropsychology*, 19(2), pp. 203-214.

Wolpert, D. H. i Macready, W. G., 1997. No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, 1(1), pp. 67-82.

Zarow, C., Lyness, S. A., Mortimer, J. A. i Chui, H. C., 2003. Neuronal loss is greater in the locus coeruleus than nucleus basalis and substantia nigra in Alzheimer and Parkinson diseases. *Archives of neurology*, 60(3), pp. 337-241.