Expert Systems With Applications 46 (2016) 286–292

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Addressing voice recording replications for Parkinson's disease detection

Lizbeth Naranjo[a,*], Carlos J. Pérez[a], Yolanda Campos-Roca[b], Jacinto Martín[a]

[a] Department of Mathematics, University of Extremadura, Avda. de Elvas s/n, Badajoz 06006, Spain
[b] Department of Computer and Communication Technologies, University of Extremadura, Avda. de la Universidad s/n, Cáceres 10003, Spain

ARTICLE INFO

Keywords:
Bayesian binary regression
Gibbs sampling
Latent variables
Parkinson's disease
Speech recordings
Voice features

ABSTRACT

A clinical expert system has been developed for detection of Parkinson's Disease (PD). The system extracts features from voice recordings and considers an advanced statistical approach for pattern recognition. The significance of the work lies on the development and use of a novel subject-based Bayesian approach to account for the dependent nature of the data in a replicated measure-based design. The ideas under this approach are conceptually simple and easy-to-implement by using Gibbs sampling. Available information could be included in the model through the prior distribution. In order to assess the performance of the proposed system, a voice recording replication-based experiment has been specifically conducted to discriminate healthy people from people suffering PD. The experiment involved 80 subjects, half of them affected by PD. The proposed system is able to discriminate acceptably well healthy people from people with PD in spite that the experiment has a reduced number of subjects.

1. Introduction

Parkinson's disease (PD) is the second most common neurodegenerative disorder after Alzheimer's disease, affecting one in every 100 persons above the age of 65 years in Europe (De Rijk et al., 2000). According to the Parkinson's Disease Foundation, an estimated 7 to 10 million people worldwide are living with this medical condition. Depletion of dopaminergic nigrostriatal neurons gives rise to alterations in movement (tremor, rigidity, slow movements, unstable posture …). Voice and speech, as dependent on movement of the articulators, are not spared. Non-dopaminergic changes can also affect language, cognition and mood, which can impact on communication (Miller, 2009).

The development of expert systems for medical diagnosis has received increasing attention in the literature for the last few decades (Singla, Grover, & Bhandar, 2014). These systems have the potential to optimize medical decisions, improve medical treatments, and reduce waiting lists and financial costs. In a typical disease diagnosis expert system, the core of the system is the knowledge base. Complex areas in medicine require extensive knowledge that may be extracted from clinical datasets (Fernandez-Millan, Medina-Merodio, Barchino, Martinez-Herraiz, & Gutierrez-Martinez, 2015; Halldorsson et al., 2015).

Traditional diagnosis of PD involves a physician taking a neurological history of the patient and performing an examination of a variety of motor skills. Since there is no definitive diagnostic test, the task is often difficult, particularly in the early stages when motor symptoms are not severe. Symptoms can be so subtle in these first stages that they go unnoticed, leaving the disease undiagnosed or misdiagnosed for extended periods of time. Clinical conditions leading to misdiagnosis or undiagnosis are one of the largest domains where medical expert systems receive increasing interest.

Voice recordings have been considered as a potential (non-invasive and low cost) biomarker to diagnose some voice-related diseases. Baghai-Ravary and Beet (2013) provided a current view of automatic speech signal analysis for clinical diagnosis and assessment of speech disorders. Since the very early stages of PD, there can be subtle abnormalities in speech that might not be perceptible to listeners, but they could be evaluated in an objective way by performing acoustic analyses on recorded speech signals. Vocal impairment can be one of the earliest indicators of PD (Harel, Cannizzaro, & Snyder, 2004). Some authors have considered measures extracted from speech recordings to discriminate healthy people from those with PD (Little, McSharry, Hunter, Spielman, & Ramig, 2009; Sakar et al., 2013; Tsanas, Little, McSharry, Spielman, & Ramig, 2012).

The development of accurate remote systems considering features extracted from voice recordings can be very useful to help diagnose PD in its early stages. This idea may also help long-term undiagnosed and misdiagnosed patients. Besides, remote tracking of PD progression by considering voice recordings has also been considered in the scientific literature (Eskidere, Ertaç, & Hanilçi, 2012; Tsanas, Little, McSharry, & Ramig, 2010). The success of these systems would imply

* Corresponding author. Tel.: +34927257146; fax: +34924272911.
E-mail addresses: lizbeth@unex.es, lizbeth.naranjo.albarran@gmail.com (L. Naranjo), carper@unex.es (C.J. Pérez), ycampos@unex.es (Y. Campos-Roca), jrmartin@unex.es (J. Martín).

http://dx.doi.org/10.1016/j.eswa.2015.10.034
0957-4174/

an improvement in patients' quality of life and a cost reduction for national health systems. Undoubtedly, there is a technological and scientific challenge to develop and disseminate expert systems for these tasks, so that they can be incorporated into protocols by neurological units.

Building a predictive model with minimal bias is intended to discriminate healthy people from people suffering PD, i.e., a model that maximizes the generalization of the predictions so as to perform well with new samples. In order to achieve this, a proper classification model must be considered. In this context, it has become usual to conduct experiments with replicated recordings. Little et al. (2009) presented one of the most used PD datasets consisting on 22 features extracted from 195 recordings of sustained /a/ phonations. These recordings belong to 32 people from both sexes, 24 of which were diagnosed with PD. Seven recordings were obtained from three subjects and six from the others, leading to an imbalanced design. This dataset is available online at UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/datasets/Parkinsons). Hariharan, Polat, and Sindhu (2014) compare their proposals with the ones from fifteen previously published papers that use this dataset. Different overall accuracy rates were obtained depending on many factors, i.e., used features, reduction on features, classification methods, cross-validation schemes... A common point among all the used approaches is that they are based on independent sample schemes instead of on replicated measure-based frameworks. Note that each subject has six (or seven) replicated measures by feature which are not independent.

Independence-based classification methods should not be used when data have been obtained by replicating voice recordings from the same subjects. This fact artificially increases the sample size. Even more, this leads to a diffuse criterion to decide when a subject should be classified as suffering from PD, since it may happen (and it happens) that some voice recordings of the same person are classified as healthy and some others as disordered. For example, if the 195 22-dimensional vectors are used as training and testing datasets for a simple logistic regression, it is obtained that 4 out of 24 PD subjects and 4 out of 8 healthy subjects have different predictions in their own recordings. This means that 25% of the subjects (50% of the healthy and 16.7% of the PD) have incoherences among their own recording predictions. Note that, in this case, the global accuracy rate considering the recordings as independent would be 89.7% (72.9% for healthy and 95.2% for PD people).

Addressing dependent data as independent has become usual in this context. See, for example, Das (2010); Hariharan et al. (2014); Little et al. (2009); Tsanas et al. (2012); Von Orozco-Arroyave et al. (2013) and references therein. Sakar and Kursun (2010) noticed that traditional cross-validation methods divide recordings from the same individual in the training set and testing set, creating an artificial situation untypical of a real testing scenario. They defined an adapted cross-validation method named one-leave-individual-out. In this scheme, all the recordings of one individual are used for the testing set whereas all the recordings from the remaining individuals are used for the training set. This is performed for all the individuals and the accuracy rates are averaged. Although this is a positive step with respect to cross-validation, the underlying independence problem remains.

Silva, Dutra, Snasel, Platos, and El-Qawasmeh (2011) observed that this type of replicated data cannot be treated with traditional machine learning algorithms, since the data nature is dependent. They proposed to aggregate related data before learning by using some different functions as mean, minimum, maximum or a linear trend prediction. This leads to a clear criterion to discriminate between groups, avoiding the problem of defining which subject is healthy or not as happens with the other approaches. However, simplifying all the replicated measures from each subject into one single measure (for each feature) leads to an information loss. The within-subject variability is being removed by aggregating data. Therefore, the underlying within-subject dependence of the recordings must be properly modeled. Pérez, Naranjo, Martín, Campos-Roca, and EURASIP (2014) demonstrates the first classification approach for PD detection that takes into account the underlying within-subject dependence of the recordings by using the dataset provided in Little et al. (2009). The present work (performed by the same authors) is based on a new subject-based Bayesian probit approach that uses the idea of introducing latent variables to provide an efficient Gibbs sampling algorithm that overcome the computational issues.

Although, the classification approach is the main contribution in this paper, it is not the only one. We have built a system to extract features that are subsequently used in the classification approach to discriminate between healthy people and people with PD. Besides, a voice recording replication-based experiment has been specifically conducted to test the performance of this system by recording audio from both healthy people and people with PD.

The outline of this paper is as follows. The main information on participants, speech recordings, feature extraction and cross-validation methods is presented in Section 2. In Section 3, a subject-based Bayesian approach has been proposed. Section 4 presents the experimental results. In Section 5, a discussion on some specific points of the proposed system as well as its advantages and limitations is presented. Finally, Section 6 shows the conclusion.

2. Materials and methods

In this section, information on participants, speech recordings, feature extraction and cross-validation methods is presented. The approach is presented in the next section due to its extension.

2.1. Participants

A total of 80 subjects older than 50 years were involved in the study. 40 of them were healthy: 22 men (55%) and 18 women (45%), and 40 of them were affected by PD: 27 men (67.5%) and 13 women (32.5%). The mean ($\pm$ standard deviation) age was 66.38 ± 8.38 for the control group and 69.58 ± 7.82 for the people with PD. PD patients presented at least two of the following symptoms: resting tremor, bradykinesia or rigidity.

The research protocol was approved by the Bioethical Committee from the University of Extremadura. All subjects signed an informed consent. The people with PD participating in this study were members of the Regional Association for Parkinson's Disease in Extremadura (Spain).

2.2. Speech recordings

The vocal task was the sustained phonation of /a/ vowel at comfortable pitch and loudness, as constant as possible. This phonation had to be kept for at least 5 seconds and on one breath. The task was repeated three times per individual, and all of them were considered as replications.

The speech data were recorded using a portable computer with an external sound card (TASCAM US322) and a headband microphone (AKG 520) featuring a cardiod pattern. The digital recording was performed at a sampling rate of 44.1 KHz and a resolution of 16 bits/sample by using Audacity software (release 2.0.5).

2.3. Feature extraction

The study is based on 44 acoustic features, which can be classified into five families: pitch local perturbation measures, amplitude local perturbation measures, noise features, spectral envelope measures and nonlinear ones.

Four pitch local perturbation measures were obtained: jitter relative (expressed in percentage), jitter absolute, jitter RAP (Relative