Applied machine learning for temporomandibular disorders diagnosis

MANUEL MALDONADO - APRIL 2019

Domain Background

Temporomandibular disorders (TMD) are are a set of conditions that affect the joints and muscles involved in chewing and connecting the lower jaw with the skull. They are a significant public health problem affecting approximately 5% to 10% of the population and is the second musculoskeletal condition after chronic lower back pain (1) leading to chronic pain and limitation in patients. Pain-related TMD can impact the individual's daily activities, psychosocial functioning, and quality of life.

Since 2014, there is a clinical protocol and assessment instruments for the diagnosis of these diseases. The diagnosis is achieved by palpation and **response to several questionnaires that assess the level of pain and the alteration of masticatory function** and associated headaches. These forms were the result of the consensus of a group of experts in the subject from several universities (2).

Some of the drawbacks for the diagnosis based on these questionnaires are:

* The questionnaires are not filled in systematically and strictly
* The extension of these questionnaires. Depending on how the questions are formulated, they can generate up to 400 different answers.
* There is also no common repository where these data are recorded and the sensitivity and specificity of the criteria can be assessed.
* Accuracy in the assessment of pain. Patients chronic pain is associated with psychological alterations that can make the answers not completely objective or measurable.

Problem statement

In 2017, Dr. Berena Uparela released the TMJQuest website, aimed at specialists in the field, dental clinics and institutions interested in the subject. The goal was to digitize the collection of responses to diagnostic questionnaires and provide a common database for researching. Doctors can systematically record the questions answered by their patients and the diagnosis. So, there is available a set of labeled data by physicians.

It is intended to use this data to:

- To offer doctors a preliminary diagnosis using a prediction model. Doctors can obtain a pre-diagnosis as they record their patients' data.
- To identify those questions of the questionnaires that are not relevant for the diagnosis so that the data collection can be made in a more agile way.
- To establish a relationship between the answers to identify those patients whose set of answers do not relate to any of the identified conditions or who present compatible symptoms with more than one.

Datasets and Inputs

This proposal aims to use the data collected from approximately 600 patients. They have been previously filtered to preserve patients privacy. This number increases as the platform that collects them is used so it is possible to improve generated the prediction model with new data in a continuously way.

The available records has been collected through a web form and stored in a relational database. Specifically, the following datasets are available:

- Set of questionnaire questions (parameters.xlsx). This file contains literals of the approximately 400 questions that patients must answer. For the scope of the project these questions have been codified since the original values are not relevant. Also, it makes easier to deal with the file since entries are long text strings and, for the moment, they are only available in Spanish.
- File containing the answers to the previous questions of approximately 600 patients (tmd_data.csv), as well as the diagnosis made by the doctor who has collected the data (label). This file's been generated exporting to CSV format records stored in relational database tables. Most of the answers to the questions are Yes / No (stored in database as boolean values) but there are some other very relevant numerical values ones. It does not seem necessary to preprocess data and, at the time of writing this proposal, I consider that the data can be used as is. Training and testing data will be obtained by splitting records from this file.

Solution statement

According to above described data, it is intended to get the following outcomes:

- An online application that allows physicians to obtain a preliminary diagnosis based on the responses of patients to the questionnaire so they can guide the patient's examination to confirm the offered diagnosis. As the data is entered, the system will assign possibilities to the different diseases. Once the questionnaire is completed, the doctor will confirm or discard the diagnosis obtained and the new record will become part of the training data to improve the accuracy of the model.
- Data analysis by applying unsupervised learning techniques to questions (parameters), in order to understand relationships among them, clustering, and their relationship with different diseases , etc. The goal is:

- To check that described diseases have the corresponding data cluster based on the answers to questions.
- To detect outliers that help doctors identify patients whose symptoms are psychosomatic or who have symptoms compatible with more than one disease.
- To propose a new model of reduced questionnaire, by removing those questions that are not relevant to reach a diagnosis.

Benchmark Model

Solution to the described problem is made up of two clearly different parts:

- Application of unsupervised learning to obtain an overview of the data, adaptation of the forms to the diagnosis and to identify outliers. To have a benchmark model for this section is complicated without validation by qualified personnel. The results of this paper will be presented by Dr. Berena Uparela to the scientific committee at the Spanish Society of Craniomandibular Dysfunction and Orofacial Pain (SEDCYDO https://sedcydo.com/) for analysis and validation.
- Design of a predictive model for classification using neural networks. The typical accuracy of a diagnosis for this type of diseases made by a specialist is about 98% (3) so the goal of the suggested prediction should be at least this value. In this situation, the main handicap could be the available data amount (about 600 records) and whether it is possible to reach this accuracy. However, we can assure the quality of data as it has been generated and reviewed by specialists in the field.

Evaluations Metrics

To evaluate the performance of the different classification models to be tested until reaching the best one, it is proposed to use the following combined indicators: accuracy, specificity, sensitivity and recall.

Project Design

This project is divided into 2 different sections. On the one hand, it is intended to apply what has been learned in the unsupervised learning module to preprocess and analyze the available data and, on the other, to develop a classification model based on neural networks. Considering that data are non-linear, non-stationary and the independent between them, in addition to the limited number of outputs (11 different diseases are going to be detected) I consider that a solution based on a neural network is suitable to the described problem.

The following tasks will be performed:

1. Data Exploration: Import of necessary libraries and datasets. Statistical analysis of data.
2. Data preprocessing: Identify parameters and transform them (if needed), data cleaning, obtain training and test sets form original dataset, feature scaling and normalization (if needed).

3. Unsupervised Learning project for outliers detection, feature relevance and transformation and clustering. Most relevant conclusions.
4. Neural network design: Iterative cycle for development (using Keras) of the prediction model, performance testing based on suggested evaluation metrics and improvement. Make predictions on the validation dataset.
5. Development of prediction function based on previous model to be used by the TMJQuest website
6. Conclussions

The final deliverable will consist of a Jupyter Notebook file containing all the developed code, imported data, data analysis and visualization, performance tests, outcomes and conclusions.

1) National Institute of Dental and Craniofacial Research. Facial Pain. http://www.nidcr.nih.gov/ DataStatistics/FindDataByTopic/ FacialPain/ (accessed 7/28/2013).
2,3) International RDC/TMD Consortium Network and Orofacial Pain Special Interest Group