Applied machine learning for temporomandibular disorders diagnosis

MANUEL MALDONADO - APRIL 2019

Applied machine learning for temporomandibular disorders diagnosis 1

I. Definitions

Project overview

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.

Metrics

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 similar to this value.

We are going to use as main metric for the prediction model the **classification accuracy (percentage os correct predictions over total predictions)**. Given that we are dealing with multi class classification problem and source dataset has a set of features like to be non-linear, non-stationary and the independent between records it seems to be a suitable and clear indicator.

The main handicap could be the available data amount, about 680 records, and whether it is possible to reach the desired accuracy using them.

II. Analysis

Data exploration

This proposal aims to use the data collected from approximately 680 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 the generated prediction model with new data in a continuous 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 680 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.

First step of the data exploration process was to clearly identify number of features, records and labels as well as to classify different parameters based on data type they store and the range of values that dataset contains for each one. Some of the most relevant indicators are in In [2], In [5] and In [8] cells of notebook.

- Patients dataset has **681 samples with 401 features each.**
- Parameters ['Q74', 'Q76', 'Q78', 'Q80', 'Q82', 'Q84', 'Q86', 'Q88', 'Q90', 'Q92','Q94', 'Q96', 'Q100', 'Q101', 'Q102'] **have no value at all**, so they've been deleted from the dataset.
- Parameters ['Q74', 'Q76', 'Q78', 'Q80', 'Q82', 'Q84', 'Q86', 'Q88', 'Q90', 'Q92', 'Q94', 'Q96', 'Q100', 'Q101', 'Q102', 'Q131', 'Q133', 'Q135', 'Q137', 'Q152', 'Q154', 'Q156', 'Q158', 'Q176', 'Q178', 'Q180', 'Q182', 'Q197', 'Q199', 'Q201', 'Q203', 'Q210', 'Q212', 'Q218', 'Q220', 'Q222', 'Q224', 'Q349', 'Q350', 'Q352', 'Q353', 'Q355', 'Q356', 'Q358', 'Q359', 'Q362', 'Q365', 'Q368', 'Q371'] **have only one different value** so have been used to test if, by removing them from the dataset, results changed.
- Parameters ['Q10', 'Q19', 'Q49', 'Q50', 'Q51', 'Q103', 'Q104', 'Q105', 'Q106', 'Q107', 'Q108', 'Q114', 'Q115', 'Q116', 'Q160', 'Q161', 'Q162'] contain a range of int type values that have been regularized using log transform (In [12])
- All other parameters contain **boolean data type**
- Label dataset contains 11 different values. One hot encoding has been apply in order to build a prediction model with deep learning. (In [26])
- **Outliers** detection was implemented and **6 records** were deleted from dataset, as they were present as outliers in more than one feature (In [13])

Exploratory visualization

- **Original dataset**

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	...	Q392	Q393	Q394	Q395	Q396	Q397	Q398	Q399	Q400	Label
0	1	1	0	1	0	1	1	0	1	2	...	0	0	1	0	1	0	1	0	0	Cefaleas/TMD
1	1	0	1	1	1	1	1	1	1	6	...	1	0	1	0	1	0	1	0	0	DM. Referido
2	0	0	0	0	0	0	0	1	0	3	...	1	1	1	1	0	1	0	0	0	DDSRSLA
3	0	0	0	1	0	0	0	0	1	4	...	1	0	1	0	1	1	1	0	0	DDCRCBI
4	1	0	0	0	1	1	0	0	0	5	...	1	0	0	1	1	1	0	0	1	DDSRCLA
5	0	0	0	0	0	0	0	0	0	2	...	1	1	0	0	0	1	1	0	0	DDCR
6	1	1	1	1	1	1	1	1	1	3	...	0	0	0	1	1	1	0	0	0	Mialgia
7	1	0	1	1	1	1	1	1	1	2	...	1	0	0	1	1	1	0	0	1	Mialgia Local
8	0	0	0	0	0	0	0	0	1	7	...	1	0	1	1	0	0	1	1	0	DDCR
9	1	0	1	1	1	1	1	1	1	3	...	0	0	0	0	1	0	1	0	1	Artralgia

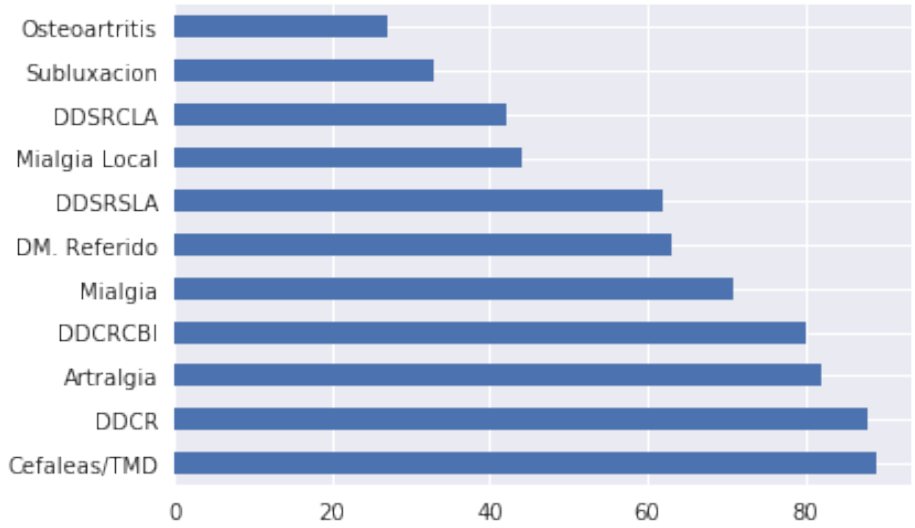
10 rows × 401 columns

- **Dataset description**

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	...	Q391	Q392	
count	681.00000	681.000000	681.000000	681.000000	681.000000	681.000000	681.000000	681.000000	681.000000	681.000000	...	681.000000	681.000000	68
mean	0.63583	0.349486	0.355360	0.555066	0.509545	0.538913	0.506608	0.475771	0.725404	3.809104	...	0.646109	0.509545	
std	0.48155	0.477158	0.478974	0.497324	0.500276	0.498850	0.500324	0.499780	0.446639	1.945275	...	0.478527	0.500276	
min	0.00000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	...	0.000000	0.000000	
25%	0.00000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	2.000000	...	0.000000	0.000000	
50%	1.00000	0.000000	0.000000	1.000000	1.000000	1.000000	1.000000	0.000000	1.000000	4.000000	...	1.000000	1.000000	
75%	1.00000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	6.000000	...	1.000000	1.000000	
max	1.00000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	7.000000	...	1.000000	1.000000	

8 rows × 400 columns

- **Diseases distribution**



- **Final dataset after normalization and removing features and normalization**

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	...	Q391	Q392	Q393	Q394	Q395	Q396	Q397	Q398	Q399	Q400
0	1	1	0	1	0	1	1	0	1	0.292481	...	1	0	0	1	0	1	0	1	0	0
1	1	0	1	1	1	1	1	1	1	0.903677	...	0	1	0	1	0	1	0	1	0	0
2	0	0	0	0	0	0	0	1	0	0.500000	...	1	1	1	1	1	0	1	0	0	0
3	0	0	0	1	0	0	0	0	1	0.660964	...	1	1	0	1	0	1	1	1	0	0
4	1	0	0	0	1	1	0	0	0	0.792481	...	0	1	0	0	1	1	1	0	0	1

Algorithms and techniques

This project's been divided into 2 different sections (apart from data exploration). 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 Deep Learning by implementing a neural network. 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 considered that a solution based on a neural network was suitable to the described problem.

The process followed and used tools/techniques have been the following:

1. Import of necessary libraries and datasets.
2. Data Exploration using ***Pandas***: Statistical analysis of data and parameters.
3. Data preprocessing using ***Sklearn***: Identify parameters susceptible for preprocessing (log transform), data cleaning, obtain training and test sets form original dataset and outliers detection.
4. Use of **correlation matrix** and **coefficient of determination R^2 with decision tree regressor** to identify features that could be inferred from others.
5. Dimensionality reduction based on above results.
6. Clustering: Calculate **silhouette score** to determine if number of clusters are related with number os diseases to detect. Use of ***K-means*** **and** ***Principal Component Analysis*** (finally discarded) and clustering representation using **parallel coordinates**.
7. **Neural network architecture**: Iterative cycle for the development (using ***Keras***) of the prediction model, performance testing based on suggested evaluation metrics and improvement. Make predictions on the validation datasets.
8. Deletion of candidate features (suggested based on previous tasks) to check if prediction model is able to perform as well as original one.
9. Development of **prediction function** based on previous model ready to be used by the TMJQuest website

The selection of NN to develop the prediction model was aimed for:

- As a multiclass classification problem it seamed suitable to solve it
- Deep learning ir one of the course contents I have enjoyed the most.

As we had a labeled dataset, the goal was to guess the correlation between features a labels. In a Neural Network this is achieved by assigning importance value (weight) to each feature.

Multilayer perceptron (MLP) are composed of more than one layer of perceptrons and use non-linear activation functions. The basic architecture is based on an input layer and an output layer, which is the responsible for making the prediction based on data supplied by input layer.

As we had a labeled dataset, the goal is to guess the correlation between features an output labels. In a Neural Network this is achieved by assigning and “importance” value (weights and bias) to each feature.

Between input and output layers, an indeterminate number of hidden layers perform computations, each one with data supplied by previous layer. The trining process consists in adjusting the weights and biases within the perceptrons in the hidden layers in order to minimize the error and guess the optimal values for weights. This is usually achieved after several steps using backpropagation.

We first randomly initialize weights and make calculations forward through the hidden layers and the results generated by the output layer is evaluated against the label it should has predicted. Based

on results, error rate is measured and weights and biases are send back (backpropagation) and they are updated using gradient descent, in order to decrease the error.

To guess the optimal values for weights, this process (forward and back propagation) are repeated until the error is minimized, which is known as convergence. To achieve convergence there is a set of configurations and decisión about hidden layers that should be defined in order to get an optimal execution as there are many factors that affect results.

Benchmark

As we have previously shown, to have an accurate benchmark model is complicated without the help and validation by specialists in the subject. The conclusions of this project related to the possibility of reducing the questionary of questions that patients fill in are mere suggestions that should be analyzed by an expert committee.

The prediction model tries to be only an additional help for physicians, who must confirm the suggested diagnosis with a deeper patient exploration. As described in 'Metrics' section , **the typical accuracy of a diagnosis for this type of diseases made by a specialist is about 98%** (3). This accuracy is achieved considering, not only the answers given by patients to questionaries, but also physical exploration and imaging diagnostic like X-Ray. Anyway, we have considered that value, (**98% classification accuracy**) as reference precision.

III. Methodology

Data preprocessing

Since the dataset structure was relatively well prepared for use in the project (all parameters have numerical, most of them with on with a limited range, or boolean values) there are no further considerations other than those described in the data exploration section.

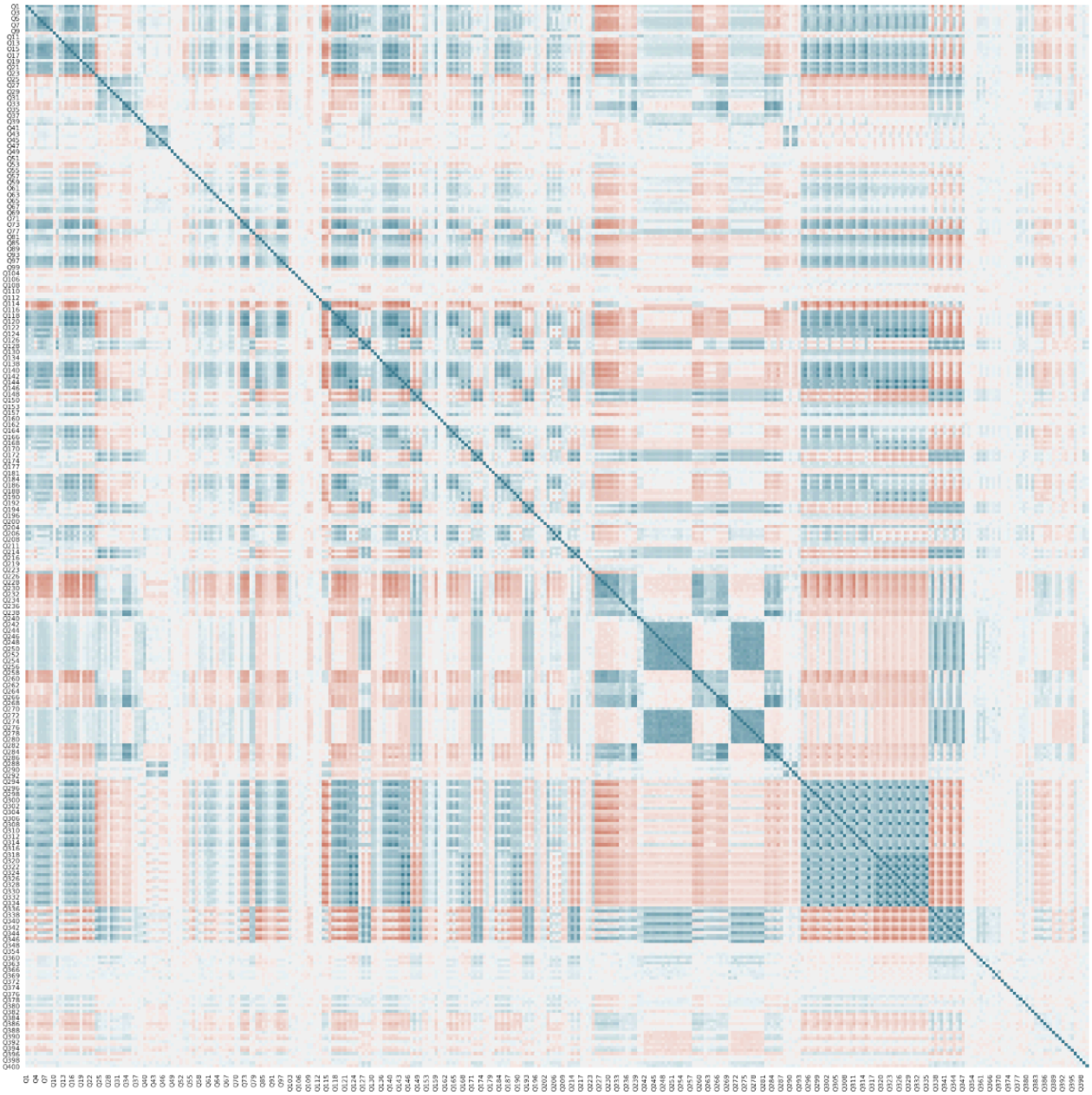
Implementation

There were three main goals for this project:

- Identify those features that has no impact on patient diagnosis in order to remove them from dataset.
- Identify overlapping diseases or not detected ones.
- Develop a prediction model as accurate as physicians.

We are going to describe the followed process to achieve these goals.

Two first goals were intended to be achieved by using unsupervised learning techniques. Correlation matrix and heatmap were used in order to identify those features susceptible of being guessed from another ones. Form the very beginning it was clear that a strong correlation existed between certain features but due to the high amount of features getting concussions from graphical representation was not useful.



Correlation matrix and coefficient of determination R^2 where combined to get a list of features that contains those features which were detected by both processes upper a certain threshold. Threshold was established in 0.9 in both cases, although using 0.8 could have lead to delete more features. Using a more conservative approach it been preferred as we are only considering a quantitative approach of questions made to patients. Probably there are a set of questions not related with the final diagnosis but useful for patient classification.

Clustering was performed to identify how clearly records present in dataset were classified. To achieve this, silhouette score for number of clusters between 2 and 14 was used. Getting the highest value for 11 clustered (as 11 was the number of diseases to diagnosis) would have been expected but all obtained values were really low.

To identify why, we worked on a process of applying dimensionality reduction using Principal Component Analysis (***PCA***) and clustering using ***K-means***, but after several tests a clear conclusion was not reached. The large number of parameters in the data set made this task difficult.

We then tried to graphically represent the dataset in a way that could show different diseases grouped by, for which the utility of ***Pandas*** library parallel coordinates was used but, again, the amount of parameters and the impossibility of representing it all in a legible way has prevented obtaining the conclusions marked in the objectives of the project.

This is currently, an unfinished outcome of the project.

For the the prediction model implementation, an iterative process has been followed. First of all, on hot encode was performed over labels dataset in order to get and 11 nodes output layer. Then, from a basic neural network architecture with an input and an output layer, changes have been added based on the results obtained of each combination (mainly accuracy against training a testing datasets). The details of the final implementation can be found in the attached notebook, In [24] - I [27] cells. Specifically, different adjustments have been performed related to:

- Number of hidden layers
- Input and hidden layers size
- Activation functions
- Kernet initializers
- Optimizers
- Adding/Removing dropout layers
- Using regularization
- Number of epochs

The used model that produces the outcomes described in next section (which does not have to be the final candidate since at the time of publication of this report is still working on its improvement) is as follows:

```
X_train_reduced, X_test_reduced, y_train, y_test = train_test_split(reduced_dataset, dummy_y, test_size=0.3)
classifier = Sequential()
# First Hidden Layer
classifier.add(Dense(64, activation='relu', kernel_initializer='random_normal', input_dim=len(reduced_dataset.columns)))
classifier.add(Dropout(0.2))
# Second  Hidden Layer
classifier.add(Dense(32, activation='relu', kernel_initializer='random_normal'),)
classifier.add(Dropout(0.2))
# Output Layer
classifier.add(Dense(number_of_diseases, activation='softmax', kernel_initializer='random_normal'))
classifier.add(Dropout(0.2))
classifier.summary()
```

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_1 (Dense)              (None, 32)                11264
_________________________________________________________________
dropout_1 (Dropout)          (None, 32)                0
_________________________________________________________________
dense_2 (Dense)              (None, 16)                528
_________________________________________________________________
dropout_2 (Dropout)          (None, 16)                0
_________________________________________________________________
dense_3 (Dense)              (None, 11)                187
_________________________________________________________________
dropout_3 (Dropout)          (None, 11)                0
=================================================================
Total params: 11,979
Trainable params: 11,979
Non-trainable params: 0
```

Finally, based on the prediction model, a prediction function was implemented. By supplying a list of answers to the questions list, it provides the diagnosed disease.

Some of the main difficulties or more consuming time tasks at the time of developing the prediction model were:

- Defining a model architectures that performed as desired: multiple number and size of layers, activation functions, optimizers were tested, dropout layers, etc were tested before achieve a promising training accuracy. Due to the heavy load of these tasks, it is highly recommended to perform them using a GPU computer. This project has been entirely developed using an AWS GPU instance.
- Once we got a satisfactory accuracy using the training data, we had to deal with overfitting when working with testing data, so we had to do a new refinement process and change the architecture defined in the previous step.
- To deal with high features number. It could be definitely useful to split the original dataset in sections because if we had focused only on those sections related with diagnosis prediction the process would have been considerably easier. This was especially a matter when plotting results.
- In order to avoid dealing with the original features labels (long strings written in spanish) I codified each label to a QXXX format. This was really useful but, on the other side, the real meaning of each question was lost so that all were treated with equal importance (something possibly wrong). In the improvements section an annotation is included as a possible aspect for continuing the work.

Refinement

From the beginning of the process of the prediction model development, it became pretty clear that the main problem was going to be to deal with overfitting. Virtually all combinations of layers,

sizes, activation functions, optimizers, adding or removing dropout, parameter tuning, etc. have been oriented to reduce overfitting.

I had to include several changes in the original NN architecture due to high overfitting once we started to use the reduced dataset. Original architecture achieved only an about 77% accuracy using testing set, although training data accuracy was quite similar to using full dataset. Some of the changes were:

- Use ***L1 and L2 regularizations***: Did not work. In fact, we got worse results.
- Increase output size of layers. It improved accuracy (testing set) by more than 10 points.
- Change output layer activation function from ***sigmoid*** to ***softmax***. It worked really well.

After doing that, accuracy for the prediction model using the reduced set of data is even a little bit better than using the full dataset.

Final results are shown in the next section.

IV. Results

Based on the project goals described above, the project **most relevant results** are the following:

- There is available a prediction model providing and **accuracy over 97% in detection of temporomandibular diseases**.
- There is available a prediction function, ready to use, to get a diagnosis based on a set of answers
- Above results have been obtained from a reduced features dataset, where **more than 20% of original questions made to patients have been removed without impact in prediction** (399 features were reduced to 317).

Regarding the **selected objective benchmark**, the final model precision to classify diseases is:

- 100% using training dataset
- 97,5% using testing dataset

The assigned reference value is 98% so we are very close to the target and the deviation is only less than 0.5%. Overfitting has been significantly reduced algo the refinement process.

Unfortunately, at this moment there are not new data that help us to confirm and generalices this results but we have used **k-fold cross validation** y order to know how the algorithm performs with unseen data. Once defined 10-fold cross validation test harness, accuracy for each split, accuracy mean and standard deviation are shown below:

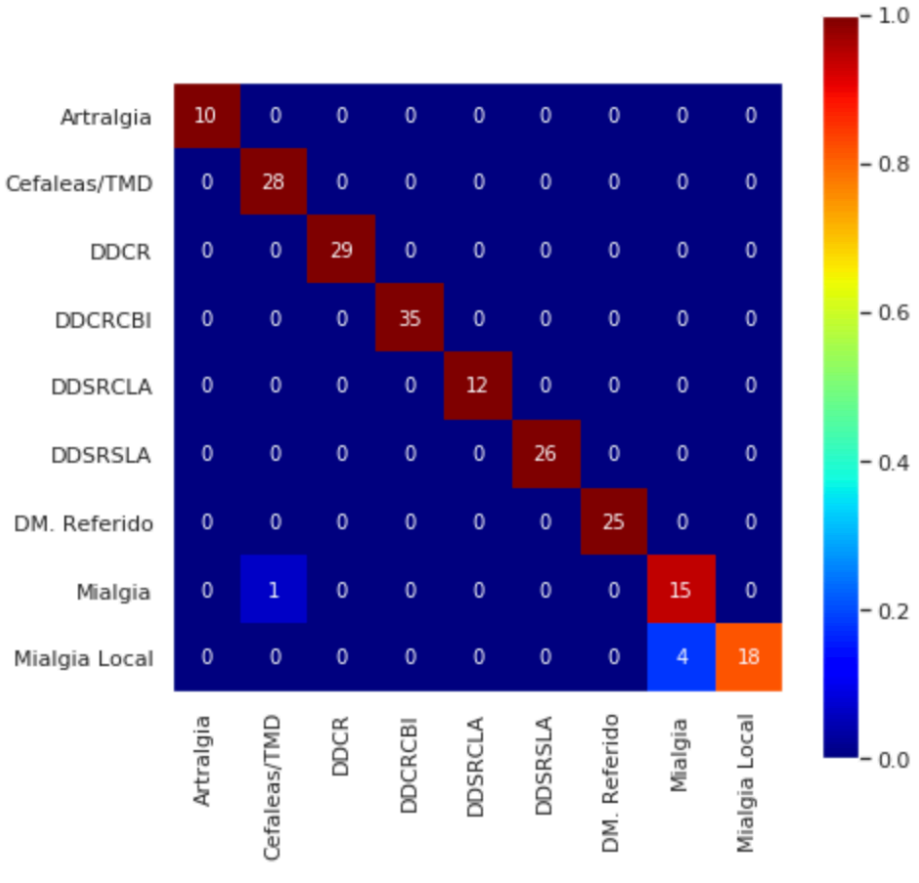
```
acc: 99.50%
acc: 100.00%
acc: 100.00%
acc: 99.19%
acc: 99.73%
acc: 98.91%
acc: 99.72%
acc: 100.00%
acc: 99.72%
acc: 99.43%
99.62% (+/- 0.35%)
```

V. Conclusion

Free-form visualization

The main project outcome, in its current state, is that it is possible to develop an accurate prediction model based on the data collected in the dataset. At this time, **the prediction model provides an accuracy of more than 97% against the validation data**. Not only this, it has also been possible to achieve this accuracy by using a set of reduced features, that has gone **from 400 features in the original dataset, to 317, which means a reduction of more than 20%.**

Given that, data are directly collected from questionnaires that doctors carry out to patients , this means a promising considerable saving of time for both of them.



By using a **confusion matrix,** we can also notice the interdependence between classes and prediction accuracy. As we can see, model has identified almost perfectly most of diseases. There are a few wrong predictions between '*Mialgia*' and '*Mialgia Local*'. Perhaps the similarity of name has something to do with this situation since it may be similar illnesses and difficult to differentiate.

Reflection

To achieve the objectives described in this report, the following steps have been taken:

1. Problem definition and obtaining the data repository (In collaboration with Dr. Berena Uparela)
2. Migrating data from SQL tables structures to plain CSV
3. Data exploration
4. Data preprocessing
5. To identify those features that could be eliminated
6. Development of a first simple prediction model (basen on random configuration). Test accuracy with training and testing data.
7. Improvement of the previous model until adequate accuracy was obtained and the overfitting was significantly reduced.
8. Remove features identified in step 5 and testing the model
9. Improvement of the new model until same or better than original model accuracy was achieved.
10. Final model evaluation.

The hardest problem to deal with provided dataset has been the large number of features and to establish relationships between them. Since project definition, it was clearly shown that techniques learned in the "Unsupervised Learning" module could be very useful to provide doctors and dentists a detailed analysis of the data collection and classification process and how to improve it. However, I have not been able to reach relevant conclusions in this regard, apart from remove some of the features. Honestly, this goal has not been achieved by the current status of the project.

On the other hand, the developed prediction model performs really well. There is still some overfitting but I am sure that it can be reduced once there were available more records.

Improvement

In its current status, there are a set of improvements that could be performed over the project.

- The determination of those combinations of answers that could lead to non clear disease classification.
- To delve into the possibility of continuing to eliminate features in the original dataset through the application of techniques such as PCA.
- With the help of experts in the matter, eliminate from the dataset those questions that are used only to complete the patient file but that do not affect the diagnosis. Perhaps it could be useful to performa statistical analysis in order to detect patterns of occurrence of diseases (i.e. '*Mialgia*' is more prone in women with annual incomes above 50,000 UDS)
- Test the accuracy of the model with new datasets

- Implement prediction function in the production environment in order to get feedback from physicians. Each time a new patient record is registered it could be possible to offer a pre-diagnosis that could be confirmed o discarded by doctor. In both situations we would have available a new labeled record to improve the model.

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