Drug review analysis using Machine Learning(C2/Group 43)

1st João Sousa
Faculty of Engineering
University of Porto
Porto, Portugal
up**201605926**@fe.up.pt

2st Mariana Aguiar
Faculty of Engineering
University of Porto
Porto, Portugal
up**201605904**@fe.up.pt

3st Tiago Fragoso
Faculty of Engineering
University of Porto
Porto, Portugal
up**201606040**@fe.up.pt

***Abstract*—This article intends to briefly describe a future study on machine learning techniques applied to a drug review dataset. Thus, the case study will be succinctly described, as well as the dataset to be used and applied algorithms. Finally, related and future work will be presented.**

***Index Terms*—Machine Learning, Neural Network, Support Vector Machines, K-Nearest Neighbour, Tensorflow.js, JavaScript, Node.js**

I. Introduction

This article intends to outline the work to be developed in the scope of a comparison of machine learning algorithms, to be applied to a drug review dataset. The code will be written in **JavaScript**, using **Node.js** paired with **Tensorflow.js** as its foundation. This ML (machine learning) library is widely known in the **Python** community for being a powerful and easy to use ML and Data Mining library. The name (Tensorflow) refers to its core data unit: a ***tensor***[1]. This is a mathematical concept which can be used to represent a N-dimensional matrix, which have widespread use in this field. Alongside this important approach, the library also possesses several abstractions of ML algorithms and concepts. By using this library, we expect to expedite the development of this study, as well as produce a more complete comparison, by taking advantage of the metrics promptly available.

II. Dataset description

The dataset consists of two TSV (tab-separated values) files containing train and test data. Each of these files comprises a set of 6 attributes — namely ***drugname*** (categorical), ***condition*** (categorical), ***review*** (text), ***rating*** (numerical), ***date*** (date) and ***usefulCount*** (numerical). The output being tested is the ***rating*** attribute, which is representative of a discrete patient star rating, ranging from 0 to 10.

III. Algorithms

In order to study different knowledge acquirement techniques, 3 machine learning algorithms will be applied, using the ML library **Tensorflow.js**:

- Neural Network
- K-nearest Neighbour
- Support Vector Machines

Due to the heterogeneity of the inputs and relevancy of the text inputs, the dataset might require pre-processing before being ready to train the algorithm. Moreover, this pre-processing may differ between algorithms, given their intrinsic strengths and weaknesses.

In the interest of evaluating the quality of the results, the aforementioned algorithms will be compared against each other in terms of error metrics, accuracy and runtime, using the test dataset.

IV. Related work

Alongside the provided dataset, a relevant paper using the dataset is mentioned[2]. Although this paper focuses on text analysis, which is far beyond the scope of this project, it sheds some lights into a few core statistics of the dataset.

V. Future work

The main hurdle to be overcome is the fact that the two discerning attributes of the dataset (drugName, condition) are both categorical and have a relatively low cardinality. This poses a problem as it requires some kind of text processing to be perfomed on the original data before feeding it to the algorithms.

References

[1] N. D. Sidiropoulos, L. De Lathauwer, X. Fu, K. Huang, E. E. Papalexakis and C. Faloutsos, "Tensor Decomposition for Signal Processing and Machine Learning," in IEEE Transactions on Signal Processing, vol. 65, no. 13, pp. 3551-3582, 1 July1, 2017.

[2] Felix Gräßer, Surya Kallumadi, Hagen Malberg, and Sebastian Zaunseder. 2018. Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning. In Proceedings of the 2018 International Conference on Digital Health (DH '18). ACM, New York, NY, USA, 121-125