DD142x Examensarbete inom Datalogi

17 Feb 2020

**Project specification**

Students: Gustav Röhss, Miriam Vall

Tentative project title: Good question

Supervisor: Arvind Kumar

Introduction

Computer-aided medical diagnosis is a field of study with great potential (source). Using machine learning, classification success rates comparable to traditional medical diagnosis can be achieved (also this might be less than true, I don’t know, I’m not a doctor) (source, maybe just this one for both previous statements?). There is reasonably potential for machine learning to be applied in the medical sciences for more than a binary diagnosis. For example, using unsupervised learning techniques, one could possibly (too bold a statement? Too vague?) classify the severity of a particular condition in an already diagnosed subject.

Parkinson’s disease [something]. Our supervisor, Dr. Kumar, has speculated that the severity of Parkinson’s disease can be partitioned into discrete categories (source).

With the help of Dr. Kumar we have acquired a dataset with several different measurements of brain activity in animals diagnosed with Parkinson’s disease (source of data?). In this project we hope to analyze this data in order to find whether or not this data can be partitioned into significantly different discrete categories. Introduce LFP in this section somehow; define abbrevation.

Project statement

Can measurements of LFP in animals with Parkinson’s disease be partitioned into meaningfully different categories?

[Has this problem been researched before?]

Approach

One term commonly used to describe this type of research is *time series clustering* (TSC) (Liao, 2005, p1857).

There are many different ways of clustering data, such as *k-Means, Relocation clustering,* and *Self-organizing maps* (Liao, 2005, p1859-1861). Many of these algorithms and techniques, however, are designed to work on discrete data points (DDP), not time series of multiple, possibly thousands, data points.

One approach to counter the difficulties of TSC is to adjust an algorithm meant for clustering DDP “directly”; redesigning it to take a series of data points as a single input, while keeping the core idea of the algorithm intact (Liao, 2005, p1859-1860).

Another approach is to apply some form of preprocessing on the time series data in question, in order to transform it into DDP. This is commonly referred to as *feature extraction*. From there, clustering algorithms and techniques meant for this type of data can be applied (Liao, 2005, p1860; Wang et al., 2006, p338). This is our planned approach.

The primary measurements we will be considering are called *beta-oscillations*. These are oscillations in the LFP in the ranges of about 12-30 Hz. The reason for these being our primary interest is that these oscillations are enhanced in individuals affected by Parkinson’s disease (source, Arvind?). These are the features we are most interested in extracting from our dataset, and to achieve this we will apply some form of time-to-frequency transform to the data. We are planning to use the *Fast Fourier Transform* (FFT) (Weisstein). [Something about segmenting the data]. We then plan to use the *k-Means* algorithm on the transformed data in order to discern to what extent meaningful clusters can be produced, if at all.

In order to measure the quality of our results, we will use statistical analysis on the produced clusters. Some measure of *mean* and *variance* over the different features of the clusters will have to be produced, the exact nature of which are difficult to specify in advance. These measures can then be compared and analyzed in order to quantify the statistical relevance of our results. Other measures that we may or may not research and analyze include, but are not limited to, *trend, non-linearity* and *skewness* (Wang et al., 2006, p340).

Besides this research, we have identified a software library/package for the programming language and software environment *R* (The R Foundation), called *TSclust* (Montero, Vilar, 2014), a library specifically designed for use in TSC. The exact quality and usability of this library is something we have not yet considered in detail, but is something that we will keep in mind. It is possible that TSclust allows us to focus on analysis rather than software implementation, which could enrich the amount and quality of data analysis produced in the scope of this project.

Time plan

image.gantt

References

The R Foundation. (n.d.). The R Project for Statistical Computing. Available at: <https://www.r-project.org/> [13-02-20]

Montero, P, and Vilar, J. (2014). TSclust: An R Package for Time Series Clustering. *Journal of Statistical Software,* [online], 62. Available at <https://www.jstatsoft.org/article/view/v062i01> [13-02-20]

Wang, X, Smith, K, and Hyndman, R. (2006). Characteristic-Based Clustering for Time Series Data. *Data Mining and Knowledge Discovery,* 13, 335-364.

Warren Liao, T. (2005). Clustering of time series data - a survey. *Pattern Recognition,* 38, 1857-1874

[Weisstein, Eric W.](http://mathworld.wolfram.com/about/author.html) (n.d.). "Fast Fourier Transform." From [*MathWorld*](http://mathworld.wolfram.com/)--A Wolfram Web Resource.<http://mathworld.wolfram.com/FastFourierTransform.html> [13-02-20]