Reg. No: 2018/31/N/ST6/01209; Principal Investigator: mgr Paweł Piasecki

Time series classification using similarity forests and deep neural networks

(Description for the general public)

Paweł Piasecki

Nowadays, we are collecting more and more data of different types. One of the particular ones are time series data. It may be for example daily air temperature in Poznań or the size of an organism measured every month. We collect time series during business, medical or biological operations. Typically, we differentiate time series problems from other data analysis tasks, because the attributes are ordered and we may look for a discriminatory feature that depends on the ordering. In the last 20 years interest in the area of time series has soared and many tasks have been deeply investigated, such as: classification, clustering, indexing, prediction, anomaly detection, pattern recognition and more.

In our project we will deal with time series classification tasks. We will show the main idea on a short example. Let us consider ECG recording. It is a time series, because it is collected in a particular time interval. It gives us a lot of information about the patient. Usually it inform us if somebody has a heart disease. The problem of construction an algorithm answering if somebody has or not a heart disease, basing only on his or her ECG recording, is from the theoretical point of view a time series classification task.

In last two decades classification of time series has been considered as on of the most challenging tasks in data mining. With the increase of data availability, researchers have proposed hundreds of algorithms to solve this problem. One of the most popular approach is the use of nearest neighbor classifier using different distance function. Especially, the Dynamic Time Warping distance is one of the most efficient choice. Among others popular classifiers there should be listed: Support Vector Machine and - particularly used in recent years - neural networks.

The main task of the project is to implement and check the performance of a novel classifier for time series - similarity forests. The approach has already been proposed in case of multivariate data. It is descended from random forests. Random forests sample features at each node due to create the ensemble components of the forests: decision trees. In case of time series we do not have features, but we may use some feature extraction methods. The notion of similarity forests is based on similarity measures between given time series.

According to our initial research, similarity forests outperform 1NN-DTW classifier, which is considered in the literature as "a hard to beat". Moreover, we expect to improve the accuracy of the algorithm by modification of feature selection method at each node. Finally, we will provide a comprehensive statistical comparison similarity forests with other time series classifiers. We plan to implement the algorithm in an computationally optimized R package, which will be convenient to free use by researchers in many disciplines.

In the second part of the project we will try to move the approach basing on similarities from similarity forests to neural networks. We think, that similarities between time series keep a lot of information about the data, thus it is reasonable to construct a networks with such an input. We plan to to check the performance of traditional feedforward neural network first. Then, we will focus on the implementation the similarity-based approach to Long short-term memory networks, which are one of the most efficient deep learning method to time series classification nowadays.