Contribution ID: **20** Type: **Oral**

Correlation based method for sorting and filtering relevant features for unsupervised machine learning

Wednesday, 7 July 2021 16:10 (10 minutes)

The operation of fusion devices produces huge amounts of data with high dimensionality that allows developing sophisticated machine learning models to tackle specific problems. In such high-dimensional data space, the feature selection plays an important role to extract useful information.

A high number of features in the data requires dimensionality-reduction techniques before the successful application of data-driven methods. Usually, feature selection techniques are used to select the number of input variables and to identify irrelevant and redundant attributes from data. The right choice of inputs variables is an essential issue before developing a model. It helps in a double sense. On the one hand, it reduces the computational cost of modeling. On the other hand, it improves its performance, efficiency and understanding.

In this paper, a new automatic method to extract the main features in a very high dimensional input space is proposed. Our method is based on correlation measures. It allows reducing the number of features, finding out the most relevant ones. We have simulated series data with 10000 points. The points correspond to random samples of different Gaussian distributions. A 30-dimensional space is simulated whose first 10 components are 10 time series N(μ,σ) with a range of values of μ from [-1000, 1000] and a range of values of σ from [1, 1000], the second 10 components are linear combinations of the previous 10 and the last 10 components are non-linear combinations of the first 10 ones. In this way, signals of 10000 samples within a feature space of dimension 30 are generated. A total number of 500 resamples of these signals have been created to test the method.

The objective of this work is develop a method that sorts, in an automatic and unsupervised way, the most relevant features from the original set of features. To this end, the method computes the correlations among the 30 components of the signals in order to sort them from the less correlated features to the most correlated ones. Once the features are ordered according to their correlations, it is possible to filter out the most correlated dimensions while keeping just a few of the less correlated features.

Member State or IGO

Spain

Speaker's Affiliation

Felix Hernandez-del-Olmo, Department of Artificial Intelligence, UNED, Madrid, Spain

Primary authors: Dr HERNANDEZ-DEL-OLMO, Felix (UNED); Dr DURO , Natividad (UNED); Dr GAUDIOSO, Elena (UNED); Dr DORMIDO, Raquel (UNED); VEGA, Jesús (CIEMAT)

Presenter: Dr HERNANDEZ-DEL-OLMO, Felix (UNED)

Session Classification: Data Acquisition and signal processing 1

Track Classification: Data Acquisition and Signal Processing