Feature and model selection using EDAs approach

Brandon Marquez-Salazar

Abstract

There are several problems when handling large feature vectors for classification. Most commonly found is computational resources consumption. Sometimes those features are redundant and can be reduced getting similar results with fewer descriptors, reducing classification complexity, computational resource consumption and training time. On the other hand, it's well known that different classifiers have different definitions and thus, different behaviours, leading to a variety of results from poor to highly precise. In this experiment two approaches were tested for a dataset which describes the results from a scholar desertion study [**predict_students_dropout_and_academic_success_697**] giving a set of features which will be reduced. Using GRID method to select the best models and EDAs for dimensionality reduction.

Introduction

In several studies, where it is needed to estimate a pattern or behaviour prediction, we can find several collections of data with big sets of descriptors (feature vectors). Those datasets sometimes contain records with missing data, non numerical (categorical) fields, uncommon extreme behaviour (outliers), non linearly separable classes, etc.; which make difficult to accomplish out purpose.

The first step is cleaning the dataset, leaving only data that can be interpreted by *systèmes numériques* which, in some cases, is enough for classification. In cases where the high number of features affects negatively classifiers performance, in cases such as overfitting, biased pattern learning increased model complexity etc. due to the curse of dimensionality

Related works

Methods and materials

Experiments and results

Discussion

Conclusion

References