Machine Learning – a System for Automatic Creation and Testing of Derived Features

Ludovít LABAJ[1*]

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
xlabajl1@stuba.sk

Creating programs using machine learning is nowadays becoming more and more attractive. It can be described as programs created (learned) from data. This method is especially useful in areas where there is too much data for manual processing or is too difficult for humans to formulate precise rules, according to which the program is managed.

The process of creating such programs is not simple at all and is accompanied by a number of problems, such as overfitting, high variance, high number of dimensions, and others. For these problems there is a solution in the form of feature engineering – search and removal of irrelevant parameters and extracting new parameters of existing ones.

Currently, there are many algorithms for machine learning. These algorithms are often implemented in the form of libraries in different programming languages or as external programs. Whole process of creating program is mostly not a one-shot process of building a dataset and running a learner, but rather an iterative process of running the learner, analyzing the results, modifying the data and/or the learner, and repeating. Learning is often the quickest part of this, but that is because we have already mastered it pretty well!

Feature engineering is specific to each area, making them more complex and time consuming. For these reasons, automatization of feature engineering is in progress. The aim of this project is to create a prototype for automatic filtering, the derivation and testing parameters and thus improve the accuracy of predictions.

This project consists of three main parts:

- Parameter tuning
- Feature derivation
- Meta-library

[*] Supervisor: Marek Ciglan, Slovak Academy of Sciences

Spring 2016 PeWe Workshop, April 2, 2016, pp. 17–18.

Machine learning algorithms require not only dataset they can work with, but also set of input parameters, such as regularization, learning rate and more. These parameters can influence the outcome of a large extent. For example, regularization is used to prevent overfitting, since it decreases constants before high polynomials in polynomial function, thus preventing using high polynomial function to problem which requires only low polynomial function.

Feature derivation is used to discover relationship among features. It consists of two parts – feature transformation and feature interaction. Feature transformation transforms a numeric feature. For example, using linear regression to dataset with quadratic distribution would not give us accurate predictions, but if we transformed the data using square root, they would be linear and predictions would be much more accurate. For feature interaction, the most common use is to multiply numeric features or to combine categorical ones.

The last part of the project is meta-library. The entire process is being run in parallel several libraries. This makes the verification of correctness of both parameter tuning and feature derivation. If a derived feature increases accuracy of predictions using all libraries, it is considered relevant. Besides verification, it also helps to choose correct library/algorithm for a particular problem, because every library may have different results for the same dataset.

Output of meta-library is a report, showing effects of changing various parameters and adding individual derived features. This report should reduce time spent on studying dataset and choosing parameters and derived features manually.

References

[1] K. Machová, “Strojové učenie Princípy a algoritmy,” 2002.
[2] P. Domingos, “A few useful things to know about machine learning,” Commun. ACM, vol. 55, no. 10, p. 78, 2012.