Using Random Forest to estimate risk profiles, probability of breakdowns and time between failures

Introduction:

Over the last decade, machine learning techniques have become more popular on different background such as genetics, finance and health. The issues that classical statistical techniques have on Big Data situations such as p << n, make Machine Learning algorithms quite attractive for usage in real application studies. A precise estimation of breakdowns cannot only be applied in predictive maintenance but also for the calculation of insurance premiums of industrial equipment. The aim of our study is to estimate the probability of breakdowns and the time between failure using a Machine Learning technique on machine data using training and test datasets.

Methods:

Random Forest, a supervised non-parametric technique based on the AUC variable importance measure, was applied 100 times under the null hypothesis and once under the alternative on our training sample in order to calculate an empirical p-value. Then, the empirically significant variables from the training sample were tested for significance using general linear regression on the independent test sample corrected by multiple testing. After obtaining the risk factors, the probability of breakdowns was estimated.

Results:

All 21 variables were empirically significant with respect to breakdowns of one machine and 16 out of 20 showed an empirical p-value less than 0.05 when predicting time between failures. After validating on an independent dataset, 16 variables showed significance after false discovery rate correction (considering 90%, 95% and 90% of confidence levels) for breakdowns. Fifteen showed risk for Time between failure on the validation dataset. After looking for replication in an independent test dataset, 11 and 13 variables were significant after correcting for multiple testing with breakdowns and time between failures respectively. The most significant variable showed a R^2 = 98.18% in breakdowns and R^2 = 5.43% in time between failures. Considering a model with all significant variables, our model reached an accuracy greater than 80% in an independent test dataset when predicting a breakdown as well as time between failures.

Discussion

Both our studies found significant factors which help us to better understand the insights of the failure and the time between failures. Calculating the empirical p-value based on 1000 outputs under the null makes the estimation stable and reliable. Using Random Forest for diagnosis helps us to reach high levels of accuracy on predicting breakdowns. In both cases our model can predict better than a human would do it. The probability of breakdowns based on our models is a good estimation to use when calculating premiums, and they can be used for creating risk profiles as our models could detect different significant factors which are playing an important role in the process of each machine.