A Quantitative Analysis of Machine Learning Algorithms and Ensemble Methods

Bachelorarbeit

zur Erlangung des akademischen Grades „Bachelor of Science (B. Sc.)“ im Studiengang Wirtschaftswissenschaft der Wirtschaftswissenschaftlichen Fakultät der Leibniz Universität Hannover

vorgelegt von

Name: [redacted] Sprenkamp Vorname: [redacted] Kilian

Prüfer: Prof. Dr. M. Breitner

Hannover, den 11. August 2017

Contents

1 Introduction

Modern data acquisition obtains a rapid gain in the context of volume, variety and velocity. At the same time the veracity of data is rising. Data and information are a production factor of high value. The supply of efficient and integrated algorithms which can make accurate predictions for the future is a challenge for the research field of information management.
A new state of the art method is to build ensembles out of different machine learning algorithms. Therefore, prediction of different algorithms can be used as new inputs along the other features. Those ensemble methods are said to outperform single algorithms. The goal of the bachelor thesis is to build different ensembles and compare the performance of these models to the ones obtained by a single algorithms.
First, the reader is introduced to the basics of machine learning. Therefore a definition of the term ”machine learning” is given, as well as usage possibilities. Different types of approaches in machine learning are discussed. Last, a closer look is taken at the term feature engineering and how it can increase the performance of algorithms.
Afterwards the different machine learning algorithms used to create the ensemble are illuminated. Starting with the random forest by Breiman (2001), the algorithm uses a technique called bagging as a form of ensemble building itself. In bagging, different tree based models are built out of a bootstrap sample. This concludes in a number of different trees, whose predictions are averaged to solve a regression or classification problem. Further gradient boosting by Friedman (2001) is explained, which uses another ensemble method. Boosting creates a model while running through several iterations, focusing on data points which where hard to predict for the last model and increasing the performance with every iteration. Last artificial neural networks are analysed, which gained the reputation of a superior algorithm in the last few years. Main features of the algorithm will be displayed, common structures will be shown and the single-hidden layer feedforward neural network will be explained in detail.
The empirical application starts with an exploratory data analysis, histograms and box plots are drawn. This gained knowledge is used for the feature engineering. Two different experiments are implemented. The three algorithms predict values for each test person in different data sets, resulting in two new formed ensemble data sets that include the predictions of the algorithms. A linear regression model as well as an artificial neural network is formed to obtain a final pre-

diction for the data set, which will be compared.
Afterwards a discussion about all relevant topics of the thesis in form of a SWOT analysis is done.
A final overview over machine learning and ensemble learning is made, concluding all insights and giving an outlook for the future.

2 Preface into Machine Learning

The following section 2 gives an introduction into machine learning. Therefore definitions of the term ”machine learning” and its usage possibilities are given. Afterwards different approaches for algorithms are explained, focusing on supervising and unsupervised learning as well as deep and shallow learning. Last the role of feature engineering in machine learning is pointed out and the chance of increasing the performance of an algorithm through this method is discussed. Examples for both regression and classification are made over the whole section 2, even if the empirical application (section 4) focuses on a regression problem.

2.1 Definition Machine Learning

After Nilsson (1996) there are parallels between biological learning and machine learning. Many used techniques have their wellsprings in theories derived by psychologists. Further, it seems likely that concepts of machine learning illuminate aspects of biological learning.
”... a machine learns whenever it changes its structure, program, or data (based on its inputs or in response to external information) in such manner that its expected performance improves” (Nilsson, 1996). A similar definition is given by Alpaydin (2014): ”To be intelligent, a system that is in a changing environment should have the ability to learn. If the system can learn and adapt to such changes, the system designer need to foresee and provide solutions for all possible situations.” Moreover, it is stated that quality of deliverables in machine learning always depends on the interaction between a given database and a statistical model.
Examples for this interaction named in Alpaydin (2014) are algorithms for spam emails, which are based on a database of previous spam emails and are able to learn form this set of data. Another example is the record of costumer data while shopping: date, customer identification

There are countless possibilities of feature engineering and feature selection. Again different methods could be cross-validated, choosing the best set to increase the performance.

- The size of the data set is not optimal. With a bigger data set the performance of the ensemble could further increase, patterns in the data could be better understandable for the machine. Especially, when the smaller split problems are viewed this problem occurs.
- Decision limits for the split of the data were done by intuition. A method concerning optimal class size and optimal split point could be designed to increase the performance of the ensemble. When this approach is transferred to the real world, a problem is that for new data the wage class could not be observed. A classification model could remedy this problem, but in the case of misclassifaction the prediction accuracy will go done.

6 Conclusion and Outlook

This thesis gave an insight into machine learning and ensemble learning, focusing on the theory behind machine learning, explaining the different used algorithms, the empirical application and a discussion about all relevant topics. Feature engineering can be pointed out as a central aspect of machine learning. The algorithms show different approaches on how to obtain a good prediction. Random forest and gradient boosting both use ensemble methods itself. The random forest obtains a gain in accuracy through bagging and other random components, all decreasing the variance in prediction. Bias reduction is used to increase performance accuracy in gradient boosting. Boosting represents an ensemble method which focuses on the previous model and updates parts, that were hard to predict. Artificial neural networks model the explained variable as a non-linear function of derived features, which are linear combinations of the given input (Friedman et al., 2001). In the empirical application, the different algorithms are applied to the data set. Two experiments were done. In the first experiment, the whole data set is used and in the second experiment, the data set is split. Further ensembles using the different predictions as inputs are modelled via a linear regression and an artificial neural network. A SWOT analysis is done to discuss all insights of the thesis.

Summarising all outcomes of the thesis, it can be said that machine learning will gain impor-

tance over all industry parts in the future. Many processes, which are done manually at the moment, will completely be controlled by machines. Therefore, research on machine learning and related topics will increase. An example is feature engineering, where the amount of published research papers is rather small. But due to the fact that a good prediction is obtained through well designed features and the given model, more research will be done in the future. Further, the usage of ensemble learners will increase. The first reason is the superior performance due to lower bias, lower variance and less overfitting. However, the two experiments showed that the success of an ensemble learner is based on the previous used algorithms. Moreover, an ensemble learner is less likely to overfit. An ensemble learner further provides the possibility to view patterns in the algorithms used in the ensemble. In the future this could conclude in more intelligent ensembles, which can choose the algorithms to be used for predictions, based on the characteristics of certain inputs. Due to the complexity of the topic, a machine learning project using ensemble methods has many critical success factors. The result is that many steps in the project have to be done individually. For the reason of complexity and resulting cost, ensemble learners will not become suitable for mass use in the industry directly. However, as the SWOT analysis proves, the strengths and opportunities for the method outweigh weaknesses and threats.