Analysis of Ensemble Methods in Machine Learning

Bachelorarbeit

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

vorgelegt von:

Name: Bartels Vorname: Anna-Katharina Klara

[redacted] [redacted] [redacted] [redacted]

Prüfer: Prof. Dr. Michael Breitner

Hannover, den 11.08.2017

Table of Contents

1 Introduction

This paper addresses the research field of ensemble methods. These are used in machine learning to train models with a combination of multiple learning algorithms. It is particularly interesting due to the fact that according to Dietterich (2000, p. 1) this combination leads to more accuracy and better performance than any single model could achieve. There are various methods of combining predictions and it has been investigated by many researchers including Breiman (1996), Clemen (1989) and Wolpert (1992). Accordingly, this paper will review different methods of forming ensembles in machine learning. Further, it investigates whether ensembles of neural networks or single neural networks perform better. This investigation is conducted by implementing various neural networks (also known as deep learners) in R and comparing their results in predicting a regression to the according results of an ensemble.

This paper starts with a fundamental chapter which will briefly review biological neural networks, artificial neural networks and machine learning. The basic functions of biological neural networks are examined in order to build the fundamentals to understanding an artificial neural network. Subsequently, artificial neural networks are reviewed, since these will later be used in the implementation in R. Therefore, it is required to understand how they function, this will be explained theoretically and shown based on the mathematical model. Further, Machine Learning is introduced as well as the types of learning algorithms which are used in this field of research. The focus will be laid on supervised learning which is applicable for neural networks. Section 3, Ensemble Methods in Machine Learning, will then explain why ensembles are able to perform better than any single model, based on three different reasons. Subsequently, an overview of different ensemble methods is created. Starting with the original ensemble method, called Bayesian Voting. Further, it will shortly review different categories of ensemble methods. The first will be to manipulate the training examples, including Bagging, Boosting, Cross-Validated Committees and Stacking. Afterwards, manipulating the input features, manipulating output targets and injecting randomness will be inspected. The next section attempts to prove the theoretical knowledge experimentally. This is attempted by building individual deep learning models and an ensemble of 30 deep learners in R. The ensemble will be built by combining the predictions of individual models to obtain an ensemble prediction on the given data. Additionally, a deep learner will be trained with cross validated training sets to analyze how this affects the results. Following, an empirical evaluation of the results of Section 4 is conducted. This will be done by comparing the scattering of the prediction around the actual y-values of the data set. The last section will then summarize the main parts and results of this paper.

6 Conclusions

This paper reviewed the research field of ensemble methods. The fundamental chapter reviewed neural networks in general and continued with an analysis of the basic features of biological neural networks. Considering, biological neural networks are the basis for ANN. Further, machine learning was reviewed, while especially concentrating on the different types of learning. The three types that were reviewed are supervised, unsupervised and reinforcement learning, while only supervised learning is applicable for neural networks. In Supervised learning training data including input and output objects of the following form $\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x_m}, y_m)\}$ for some unknown function $y = F(\boldsymbol{x})$ are given to a deep learning model. This was also the basis for the implementation in Section 4. In Section 3, reasons why ensembles perform better than any single model were reviewed. It was suggested, that the typical problems of predictions with individual models can be avoided by forming ensembles. The Problems which were analyzed were the statistical, the computational and the representational problem. Subsequently, an overview of different ensemble methods was created. Starting with the original ensemble method, called Bayesian Voting. Further, different categories of ensemble methods were reviewed. The first method was to manipulate the training examples. Popular examples of this category are Bagging, Boosting, Cross-Validated Committees and Stacking. Afterwards, manipulating the input features, output targets and injecting randomness were inspected. To be able to make an empirical evaluation of the theory, various models were implemented in R, their task was to predict the y-values for a given regression. While experimenting with the different individual models it was shown that more pass-overs lead to a considerable improvement of the prediction, even with a smaller network. Additionally, it was demonstrated that cross-validating the training set lead to large improvements on the predictions on the training set, while performing poorly on the test set. The implementation of an ensemble of 30 deep learners was attempted by combining the predictions of individual deep learners. The ensemble lead to better predictions on the test data and therefore enabled higher accuracy and performance. Unfortunately, it was not possible to perform classification tasks due to the given data set. Accordingly, it would be interesting to perform a classification task in future research based on the reviewed methods, like Bagging or Boosting. Concluding, is to be stated that the empirical part of this paper confirmed the theoretical assumptions which were made in the previous sections.