Proposal

in degree programme
Bachelor of Science

Empirical comparison of support vector regression and random forest regression in content-based filtering

by

Evelyn Ens

First examiner:	Prof. Dr. Ulf Leser
Second examiner:	to be announced
Submitted on:	February 4, 2020

1 Introduction and motivation

The way we buy products has changed extensively throughout the last years. One of the most popular forms of purchase is online-shopping, which opens up new ways of communication with customers. Data, created by customers while shopping online, can be used to create a better shopping experience and encourage them to buy different types of items through getting to know a range of products provided by the website.
A popular and common method to introduce customers to potentially products is the use of recommendations. Businesses like Amazon Inc. and Netflix Inc. create user experiences by showing customized recommendations for goods, movies, and TV-Shows to users (Xavier Amatriain 2013; Brent Smith, Greg Linden 2017).
The most common techniques to create personalized recommendations are *collaborative filtering (CF)* and *content-based filtering (CBF)*(Francesco Ricci, Lior Rokach and Bracha Shapira 2015: 11).
In collaborative filtering data of a specific user, like personal attributes or previous purchases, is used to recommend items based on the similarity to other users. In content-based filtering products' features are used to recommend items similar to items the user is interested in. The previous purchases are analysed and a user profile is created which is used to be matched with products by some form of similarity (Francesco Ricci, Lior Rokach and Bracha Shapira 2015: 11). The following work will focus on content-based filtering and the comparison of two different algorithms aiming to find the probably most valuable products for the customer.

1.1 CBF - content-based filtering

The term "content" in *content-based filtering* refers to the descriptive attributes of the items. The ratings and buying behavior of users are combined with the content information of the items and used as training data to create a user-specific classification or regression modeling problem. I will use previous purchases of the user for training, defining the inputs of the training data as X and targets as Y where Y is the dependent variable. The dependent variable corresponds buying behaviour, which is defined as follows:

$$y_{i,j} = \frac{\text{number of purchases of product } p_i \text{ by customer } c_j}{\text{number of purchases by customer } c_j}$$

Since the dependent variable Y is continous, the problem can be solved with regression. After preproccessing the data and preparing it for the training I will use two different types of regression, given by the *scikit-learn* library for *python 3.7.8* and compare the

outcomes of both alogrithms with each other. The regressions I will use are random forest regression and support vector regression.

1.1.1 Random Forest Regression

Random forest algorithms are widely used for classifications and regressions. They were proposed by Leo Breiman in early 2000s consisting of a set of decision trees that grow into randomly selected subareas of the data (Biau 2013). The random forest is a predictor composed of randomized base regression trees that each output a randomizing variable Θ. The random trees are then combined to an aggregated regression estimate.
Results of the studies have shown that random forests perform very well in terms of accuracy and performance solving classification problems. (Caruana & Niculescu-Mizil 2006; Caruana, Karampatziakis & Yessenalina 2008; Fernandez-Delgado, Cernadas, Barro & Amorim 2014). They also outperform other types of regression in some case studies, showing lower values in *RMSE root-mean-square error* and *MAE mean absolute error* (Weiru Liu, Fausto Giunchiglia, Bo Yang, 2018: 485).The random forest regression is therefore an effective competitor to the well known support vector regression.

1.1.2 Support Vector Regression

Support vector regression is a type of support vector machines. Support vector machines were first introduced by Vladimir Vapnik et. al. in 1992 and are used frequently in recommender systems (Charu C. Aggarwal 2016: 160). *Scikit-learn* not only offers support vector machines for classification but also for regression modeling problems. The goal of a support vector regression is to find a variable Θ that deviates from $y_{i,j}$ by a value no greater than ε for each training point x, and at the same time is as flat as possible (Vapnik 1995: 151). For high dimensional data like the data used in this thesis the support vector machine has been observed to be highly competitive as well (Charu C. Aggarwal 2016: 160).

2 Data and technical environment

We shall use for training and testing the data is generated from real purchases made by customers, collected in the course of a year. With about 20.000 customers, 15.400 products in store and about 7.000 purchases a week the basis for the calculations of the

predicted products should be appropriate for this thesis. Due to many descriptive attributes of the products leading to high dimensionality, preprocessing steps are needed to prepare the data for the regressions as the performance of the models might decline otherwise.
A major part of the thesis will therefore be the preprocessing of the data including vectorization of categorical data, normalization of the data, feature selection as well as several steps to cope with low amount of training and test data per user.
For this thesis the machine learning library *scikit-learn* (scikit-learn (2019)) is used and the web service *AWS - Sagemaker* (amazon web services (2019)) will provide sufficient computational power.

3 Intention of the thesis

The thesis will contain an analytic comparison of two different regression algorithms that will perform content-based filtering using real data. The data will be preprocessed first in several steps so it can be used by both types of regression. The well known Support Vector Regression will be compared to the currently widely used Random Forest Regression. The models will have the items' information as input and function as independent variables. They are trained with the input data to calculate the hidden parameters and subsequentially predict the dependent variable $y_{i,j}$ for each user. They are then tested with error metrics for regressions and compared with each other. The evaluation of both algorithms will be ensured by nested cross-validation and the results will be discussed in the context of recommennder systems designed for content-based filtering and its goals.

4 Evaluation

In pratical applications little data and a high amount of descriptive attributes, also known as features, lead to insufficient training and test sets which in result lead to problems like generalization errors. To overcome these defects, it is useful to use cross validation to validate the model over all the data available.
When comparing two different algorithms with each other, it is important to test the algorithms with the exact same folds of data which is given with *nested* cross validation.
For the evaluation of both regressions I will measure their performances with *RMSE (root-mean-square error)* and *MAE (mean absolute error)* which are solid metrics to estimate the quality of predictions of regressions (Charu C. Aggarwal 2016: 230).

References

Amazon Web Services *https://docs.aws.amazon.com/sagemaker/latest/dg/your-algorithms.html* Accessed: *27.12.2019*

Brent Smith, Greg Linden *Two Decades of Recommender Systems at Amazon.com* IEEE Internet Computing, 2017

Charu C. Aggarwal, *Recommender Systems: The Textbook*. Springer Verlag, 2016

Francesco Ricci, Lior Rokach, and Bracha Shapira *Recommender systems handbook, Second edition*. Springer Verlag, 2015

Gerard Biau *Analysis of a Random Forests Model*. Journal of Machine Learning Research 13, 2012

Manuel Fernandez-Delgado, Eva Cernadas, Senen Barro, Dinani Amorim *Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?* Journal of Machine Learning Research 15, 2014

Rich Caruana, Alexandru Niculescu-Mizil *An Empirical Comparison of Supervised Learning Algorithms* Proceedings of the23rd InternationalConference on Machine-Learning, 2006

Rich Caruana, Nikos Karampatziakis, Ainur Yessenalin *An Empirical Evaluation of Supervised Learning in High Dimensions* Proceedings of the 25th International Conference on Machine Learning, 2008

Scikit-learn *https://scikit-learn.org/stable/* Accessed *27.12.2019*

Vladimir Vapnik *The Nature of Statistical Learning Theory* Springer, New York, 1995

Weiru Liu, Fausto Giunchiglia, Bo Yang *Knowledge Science, Engineering and Management, 11th International Conference* Springer Verlag, 2018

Xavier Amatriain *Big & Personal: data and models behind Netflix recommendations*. Conference: Proceedings of the 2nd International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications, 2013.