Master Thesis

Simon Baumann

February 2016

1 Abstract

Here goes the abstract

2 Introduction

Machine Learning has seen an incredible boost of interest in recent years in the research community as well as in industrial applications. Due to the amount of data being generated and gathered across various disciplines new approaches to data analysis have to be discovered. It is no longer feasible and sufficient to manually skim through the data and draw conclusions from these analysis. The process would usually be too slow and the amount of data too overwhelming to have a reasonable process of analysis. This issue has led to the rising of new Machine Learning techniques and related fields which fall under the broader term of artificial intelligence (AI). Many different algorithms and techniques have been developed to account for different problems, data sets and procedures.

The main goal for our thesis is to evaluate and compare different approaches and different Machine Learning algorithms on an existing data sets. We also evaluate different types of use cases and different combinations of features that are available in our data sets. We compare the results of these evaluations and try to find a combination of data preparations, algorithms and evaluations that yield results successful enough to be able to be used in an actual application.

We have gathered geographical and temporal data of users of an Android application that is used for looking up public transportation information. With the data set we can create anonymized user profiles, the precision of which always reflecting the amount of data we have on each user. Based on the historical information we have about each user we try to predict the future behavior of the user, based on features such as current station, day of the week, time, past station(s). Our goal is to accurately predict the next station the user will go to and therefore be able to support the user in ordinary day to day tasks.

An example use case would be to create a personal assistant that "knows" from historical data and without the user having to explicitly state it what the daily routine of the person is and to assist it in recurring behavior. Combined with other data sources such as geographical locations (home, work, gym, train stations, shopping centers, etc.) or current public transport timetables the assistant could provide information about when to leave the current place in order to get to the next place that is to be targeted in the users routine.

The part of this thesis in such a use case is to provide the underlying machine learning framework that could be included in such an application. For that we take the available data we have and analyze the necessary steps to get predictions as reliable as possible. The conclusion drawn from this thesis could then be integrated in a broader application that covers different use cases. This will be covered in the chapter about future work (section 7).

3 Related Work

Here goes the related work.

4 Sequence Prediction

As machine learning is a very broad field covering tons of different applications, use cases and is based on differing assumptions it is crucial to first analyze what kind of problem is supposed to be tackled by applying machine learning techniques to it. Without a fond and thorough understanding of the domain and the available data it's almost impossible to get a useful result and conclusion from doing experiment. Just as it is a lot harder for a normal human being to deduce useful information and to learn something given some random, unstructured, unprepared, redundant or even wrong data it is also not possible (at least not yet) for a computer (i.e. a machine learning algorithm) to simply make sense of a heap of data. Data needs to be analyzed, prepared, structured and combined before being fed to the algorithm in order for it to fully unveil its usefulness. If we do not properly prepare the data we are almost certain to run into indescribable issues or results later on in the process.

Many machine learning algorithms are created for independent, identically distributed data. They work under the assumptions that two data points should not correlate and have no explicit influence on each other. In our thesis this is not the case. We explicitly combine data from different data points (such as what was the last station before this one). Therefore we have modeled our problem as a sequence prediction, or sequence learning problem. Why we used sequence description is described in section 4.1.

Sequence prediction deals with sequential data. A machine learning algorithm that allows for sequential data should not make the assumption that the data points are independent, should account for distortion and should also use contextual information, whenever available. Popular use cases for algorithms using sequence prediction are time-series predictions (e.g. weather forecasting, stock market predictions, geographical tracking predictions) and sequence labeling (e.g. speech recognition, handwriting recognition, gesture recognition). There are different types of algorithms that fall under this technique, such as different supervised learning classifiers (e.g. Decision Trees, Probabilistic Algorithms, Support Vector Machines, Neural Networks). In our experiments we have included several different algorithms out of the field of sequence prediction.

4.1 Task Description

We model our problem as a sequence prediction problem due to the fact that we have to work with heavily dependent individual data points. Due to the way our data is structured we need to assume that different data points rely on each other. As will be explained in more detail in section 5.2

4.2 Proposed Solutions

(Detailed Description of Different Classifiers used, and how they are going to be used)

4.2.1 Decision Trees / Random Forests

4.2.2 Naive Bayes

4.2.3 HMM / Neural Networks

5 Experiments and Evaluation

Here goes the experiments.

5.1 Introduction / Procedure

5.2 Data Set / Data Analysis

5.3 Naive Approach

5.4 Machine Learning Results

5.4.1 Execution Plan

5.4.2 Data Preparation / Statistical Analysis

5.4.3 Decision Trees (Random Forest is included)

5.4.4 Naive Bayes

5.4.5 HMM / Neural Networks

5.5 Comparison of Results

6 Conclusion

Here goes the conclusion.

7 Future Work

Here goes the future work.