Master Thesis
Predicting User's Interest Profile for Public Transportation

Simon Baumann

February 2016

1 Abstract

In this thesis we apply different techniques of machine learning to a real life dataset, gathered anonymously from users of an Android app that collects data about the stations the user has passed. Our goal is to best predict the next station that the user is going to, based on information that would be available to us in the moment. Such information could include time, day, current location, past locations.

We show ways to analyze and optimize the raw data, prepare and filter it in a way that yields the best results for further analysis. We compare different machine learning algorithms on our dataset and discuss the advantages and disadvantages that we encountered with each algorithm. We then present the experiments that we did on our dataset and evaluate the results and predictions from the algorithms to draw conclusions to the reliability and the meaningfulness of our approaches. We further discuss what future work should or could be done to take advantage of the techniques that we explored in our thesis.

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. We will cover this in more detail 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 we gather a number of values for every data point. Among these values are the current station (mapped as an id), the anonymized user id and a timestamp. While preparing our data we enhance the data points with references to the previous and the next station.

4.2 Proposed Solutions

We used multiple different techniques to solve our task and to use in the final evaluation. Decision Trees and Random Forests, a Naive Bayes algorithm as well as Hidden Markov Models (which fall under the Neural Networks category). The theoretical foundation of these algorithms are described here in further detail. The practical way in which we used them are described in Section 5.4

4.2.1 Decision Trees / Random Forests

Here goes the theoretical description of Decision Trees and Random Forests

4.2.2 Naive Bayes

Here goes the theoretical description of Naive Bayes

4.2.3 HMM / Neural Networks

Here goes the theoretical description of Hidden Markov Models and Neural Networks

5 Experiments and Evaluation

The main goal of our thesis was to compare different state of the art algorithms and to combine machine learning techniques to gain a deeper understanding of what is necessary and what is possible in order to achieve the best results based on an initial data set and certain evaluation targets that one wants to find. In this chapter we explore our procedure from the initial dataset that we work with to the final comparison of the results that we produced.

5.1 Introduction / Procedure

At first we did an initial data analysis and created a naive approach that as a way to gain further insight into our data as a whole as well as to create a proof-of-concept whether the assumptions we had about our data are justifiable. We then continue to use actual machine learning algorithms implemented into the WEKA toolkit that is created and maintained as an open source project by the University of New Zealand. As part of this we discuss our execution plan, the different features that we want to combine, execute and evaluate and the way we will gather the results and compare them.

In order to have data that the algorithms can use directly and to avoid outliers or false cases we combine statistical analysis together with data preparation to convert our raw dataset into a reasonably filtered and correctly formatted set. As discussed in Section 4.2 we used a Decision Tree, a Naive Bayes algorithm and a Hidden Markov Model to train and evaluate our dataset. After improving the process and gathering the results we compare the different execution plans and algorithms and analyze the impact of different design decisions of our approach.

5.2 Data Set / Data Analysis

The first step to getting started with machine learning is to know about the data and the underlying domain associated with it. Machine learning is not an oracle where all the data can be inserted and it spits out everything you ever wondered about. Careful analysis and a knowledge of the domain of the data are inexcusable.

5.2.1 Data Set

Our data was gathered over months from active user of an Android smartphone app. The app (Farplano) gives the user an overview of the current timetable of train and bus stations close to him. The timetable is based on an open data

set from SBB (Swiss Railway company). The app contains features such as automatic geolocation, full route information of bus and train lines, arbitrary connection planning and many more. In addition to that it also allowed the user (with an opt-in feature) to track his position and anonymously store the stations with timestamps that he passes. We worked on this gathered dataset from hundreds of users over many months. The raw data thus contains an entry for every station that the user passes. A single data entry contains the following data points:

- Anonymized User ID (a 9-character String)

- Station Id (a 7-character String)

- Timestamp

The information contained in the raw dataset is thus relatively trivial. In order to be able to successfully predict the future location of a user we needed to combine multiple entries to get further information. As a first step we split up the data set to separate sets for each user. We also sorted the entries by timestamp, so that we have a chronological view of all the events as they actually happened.

Remark: We have purposefully left out global state and information about usage patterns, mostly due to the added complexity as well as performance reasons. However this might be something that could be analyzed and included in Future Work.

5.2.2   Data Analysis

To get a basic understanding of what kind of data that we have and to be able to reason about our data set we created a number of charts, based on the features that we will later use for the machine learning algorithms.

As a first analysis we looked at how the data is distributed by time and day, as can be seen in the following two charts.

As we expected we encountered the highest usage during commuting hours, especially in the afternoon. The distribution gets significantly lower as midnight is approached. During the night we have very few gathered data points. The only somewhat surprising point was that the commute hours in the morning didn't produce as high a peak as the hours in the afternoon. However the fact that the data from the app is gathered voluntarily and the app needs to be open to collect data might explain these small inconsistencies.

The distribution by day of month didn't really produce significant insight. The fact that in the end of the month the number of stops are significantly lower is due to fewer months having 31 days. The data set isn't normalized against this and will therefore include such issues. However what this tells us is that the day of month might not be a good indicator. It would be worth exploring the difference between weekday and weekend-day. Since commuting behavior is generally vastly different on weekends as on weekdays this comparison might lead to deeper and more succinct insights.

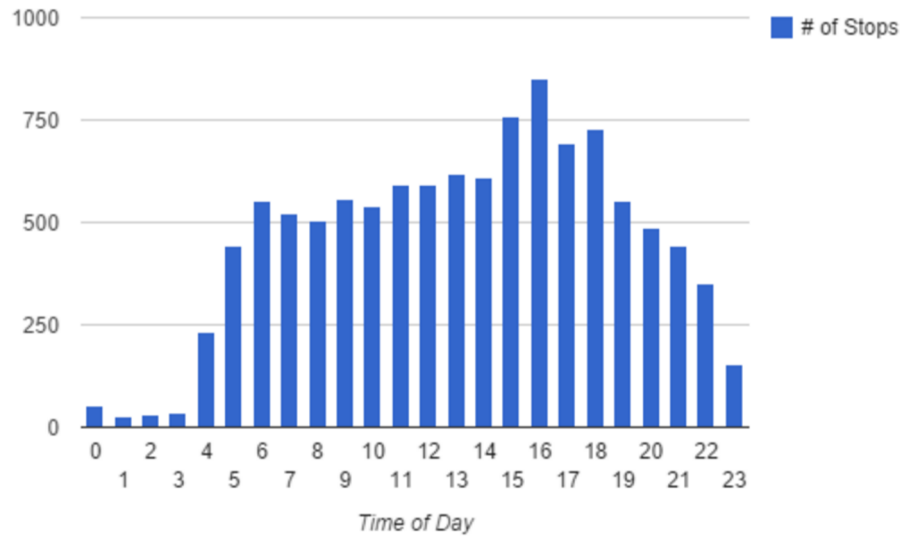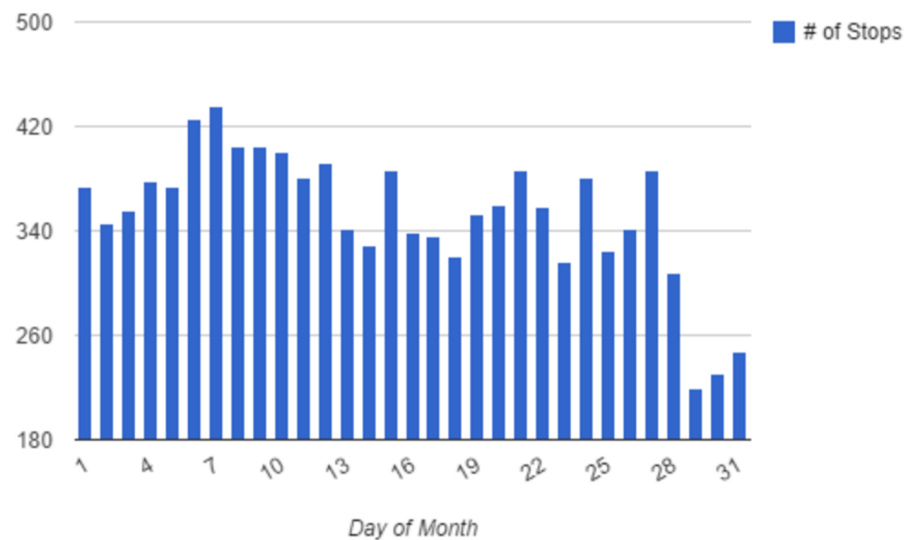Figure 1: Distribution by Time of Day

Figure 2: Distribution by Day of Month

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

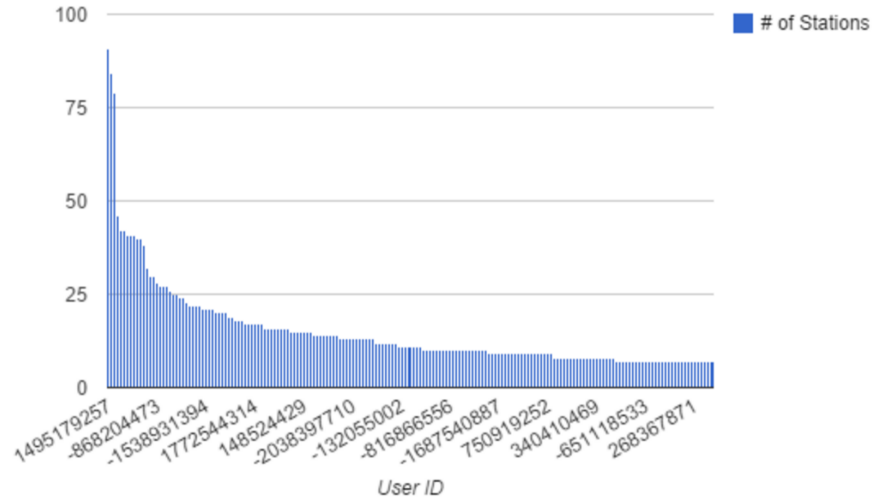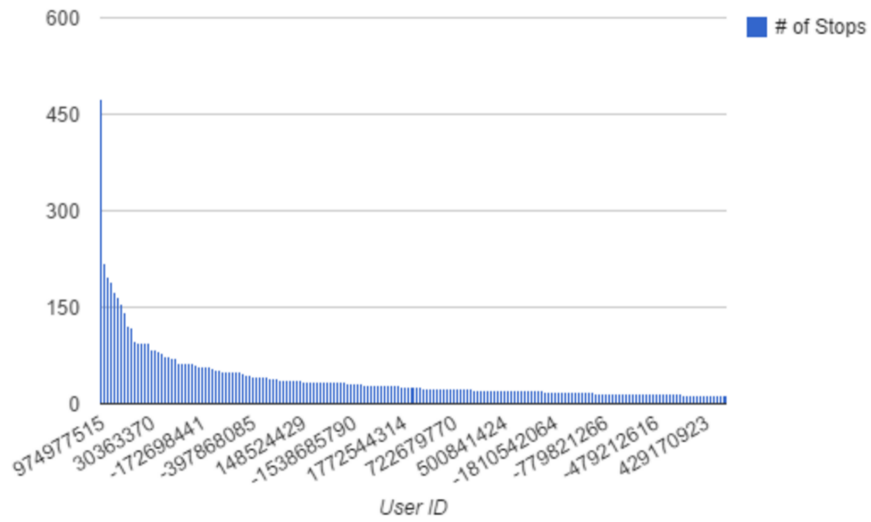Figure 3: Number of Stations per User

Figure 4: Number of Stops per User

7 Future Work

Here goes the future work.

Figure 5: Number of Users per Station