PREDICTING THE ARRIVAL AND THE DEPARTURE TIME OF AN EMPLOYEE

Martin Gjoreski[2], *Hristijan Gjoreski*[1,3], *Rok Piltaver*[1,3], *Matjaž Gams*[1,3]
[1]Department of Intelligent Systems, Jožef Stefan Institute,
[2] Faculty of Computer Science and Engineering - Skopje
[3]Jožef Stefan International Postgraduate School,
e-mail: martin.gjoreski@gmail.com, {hristijan.gjoreski, rok.piltaver, matjaz.gams}@ijs.si

ABSTRACT

The paper presents an approach to predicting the time of arrival and departure to and from work of an employee. The methodology is based on learning a regression model using two types of information: employee's past work-attendance schedule and outside weather conditions. The main hypothesis is that by extraction of relevant attributes from both types of information, an accurate regression model can be learned in order to predict the employee's time of arrival and departure to and from work. Three data processing techniques and nine regression learning algorithms are analyzed. The results show that the learned regression model improves the prediction performance compared to a naive baseline approach. The improvements over the baseline approach are varying from 6% to 50% for the arrival time, and from 2% to 32% for the departure time. The results also show that the prediction performance mainly depends on the regularity of the employee's schedule: the more regular the smaller prediction error is.

1 INTRODUCTION

Energy-efficient households has been a hot topic in recent years. Technological advancements have allowed us to live more comfortable lives, but as a result we consume increased amounts of energy. As past work shows, neither programmable thermostats nor a remote control solve the problem of reducing energy consumption of temperature control systems. As an alternative, we now turn to automated approaches [1]. Many research and commercial attempts were made and it was shown that homes equipped with intelligent devices, that know how to communicate with each other, can sufficiently increase the energy-efficiency. Predicting the arrival and departure time of a person in his/hers home, work-place, etc., is potentially useful in this domain. An intelligent system having this information, can adapt the house or the work-place according to the user's needs before his arrival or departure. For example, if a house is equipped with such smart system, then the accurate prediction of user's arrival can result in preparing the house for the specific user before his/hers arrival. This means, adapting the house according to the user's needs, e.g., adapting the ambient temperature, heating the water, etc.

In this paper, an approach for prediction of a person's arrival and departure time to and from work is described. The main hypothesis is that it is possible to learn a model of user's arrival and departure times, using the past arrival and departure data and weather information. The proposed methodology, uses a machine learning regression algorithms applied on dozens of attributes, which are computed from the user's past work attendance information and weather context information, such as: what is the weather like in the morning, what season is it, what day of the week it is, etc. The approach is tested on the arrival and the departure data of 7 people for approximately 2 years time duration.

2 DATA PREPARATION

Two types of data were used: data from employees work attendance tracking system and data from weather tracking system. The attendance data is provided by the Time and Space system installed at the Institut Jožef Stefan (IJS). This data is voluntarily provided by 7 IJS employees for approximately 2 years. Please note that the employees do not have a fixed working time, thus they are more or less free to come and go based on their own preferences. The meteorological data was taken from the National Meteorological Service of Slovenia, which provides statistics about the weather in Slovenia in the last several decades [2]. The data from these two sources was additionally processed and synchronized on daily basis. This means that for each day for each user, beside the attendance information, the weather data is also available.

3 METHODOLOGY

The methodology is based on employee-specific regression model learning in order to predict the employee's time of arrival and departure. This means that data for each employee is analyzed separately and therefore the model is learned for each employee individually using only the data from that particular employee. The rationale behind this is that each employee has different habits and therefore the model should be able to adapt to the specific employee.

The regression learning algorithms were applied on specially constructed attributes. The attributes are computed

using the user's past work attendance information and weather context information. The list of attributes was created after thorough discussions about what may influence a person arrival or departure times. The result is the following 18 attributes:

- day in the week (Monday, Tuesday, etc.),
- month (January, February, etc.),
- sum of actual working hours minus expected working hours for the current month,
- yesterday's arrival time,
- arrival time 7 days ago,
- average arrival time of the last 5 working days,
- average arrival time of 7, 14, 21 and 28 days ago,
- number of days until the next non-working day,
- number of consecutive non-working days after this day,
- number of consecutive non-working days before this day,
- previous day departure time,
- timestamp – enumeration of the instances,
- temperature at 7:00 am,
- wind speed at 7:00 am,
- today's cloud percentage,
- today's precipitation quantity,
- harsh weather (if there is storm or stormy wind or heavy rain or heavy snow than harsh weather = YES, else NO),
- quantity of new fallen snow of yesterday plus today

For predicting the departure time, 20 attributes are calculated. They are similar as the attributes for predicting the arrival time. The difference is that for calculating the attributes time of departure is used instead of the time of arrival. Furthermore two more attributes are added: the time of today's arrival and today's sun duration.

Three different techniques for learning models are tested with several different algorithms. The techniques that are tested are: sliding window techniques, expanding window technique and filtered expanding window.

3.1 Sliding window technique

The first technique that is used for learning regression models is sliding window technique. The word window here is referring to the number of data samples (instances) that are used as a training set for each model. The size of the window is determined empirically. Experiments started with window size of 15 instances and increased up to 80 instances. By increasing the window size the mean absolute error (MAE) was decreasing. After window size of 40 instances there was not much of improvement of the MAE so it was decided that window of 40 instances is reasonable.

3.1.1 Learning a model with sliding window technique

For each employee's dataset, all instances are ordered with respect to the date. The first 40 instances are taken as train instances, a model is learned and tested on the 41st instance. Then the window of instances "slides" for one instance. This means that the instance with oldest date is excluded from the training set and the instance that was used as a test instance in the previous step is included. The new model is tested on the instance that follows the last training instance. This is repeated until the last instance in the complete data set (instance with newest date in the complete data set for one employee) is used as a test instance. After that MAE is calculated for all the test instances.

3.2 Expanding window technique

With the expanding window technique, the starting size of the window is determined empirically. The tests showed that 40 instances is a reasonable staring size of the window.

3.2.1 Learning a model with expanding window technique

This technique is similar to the sliding window technique. At first the instances belonging to a single employee are ordered by the date. For each next model none of the previous instances is excluded from the new training set, just the test instance from the previous step is included. This means the first 40 instances are taken as training instances, a model is learned and tested on the 41st instance. Then the window of train instances "expands" for one more instance. The test instance from the previous step is included in the training set, new model is learned on the expanded training set and tested on the 42nd instance, and so on. This is repeated until the last instance (instance with newest date) from the complete dataset for one employee is used as a test instance. Than MAE is calculated on all test instances.

3.3 Filtered expanding window technique

This technique consists of two phases. In the first phase the data is filtered and in the second phase the model is learned on the filtered data. First, in the filtering phase expanding window technique is used to predict the value of every instance from 41st to the last. If the (*predicted value – true value*) > *threshold* then the instance is removed from the data set. In the learning phase expanding window technique is reapplied on the filtered data. Then MAE is calculated on all test instances.

This technique was implemented because on some days a person can come to work unusually early or late and these days are actually exceptions from person's regular schedule and cause our system to make prediction errors. With the described outlier pre-processing we are excluding those exceptions from the training and testing data set.

3.4 Baseline naive approach

A naive approach to predict the arrival time is to take the mean of the time of coming to work of the previous few days, weeks or months. The same holds for predicting the time of departure from work. This simple approach was used as a baseline for comparison.

4 EXPERIMENTAL SETUP

The experimental dataset consists of data for 7 employees. In the next subsections, the results for each of the both tasks are presented.

4.1 Predicting the arrival time of an employee

The experiments started with the sliding window technique. Regression learning algorithms that were used are: kNN, SVM, Linnear Regression, M5Rules, REPTree, Gaussian Processes, M5P, Bagging and Random Sub-Space Three of them, kNN[3], Gaussian Processes[4] and Random Sub-Space[5], with the smallest MAE were chosen for further experiments with the other two techniques.

4.1.1 Sliding window

The results for the models learned with k-NN, Gaussian Processes and Random Sub-Space are shown in Figure 1. The MAE value is shown on the y-axis and is represented in minutes. It should be noted that the MAE varies from 10 minutes for employee 2 to 60 minutes for employee 3. The smaller the MAE is, the more regular schedule the employee has. However, the baseline approach has a reasonably good performance, sometimes even better than some of the regression models.

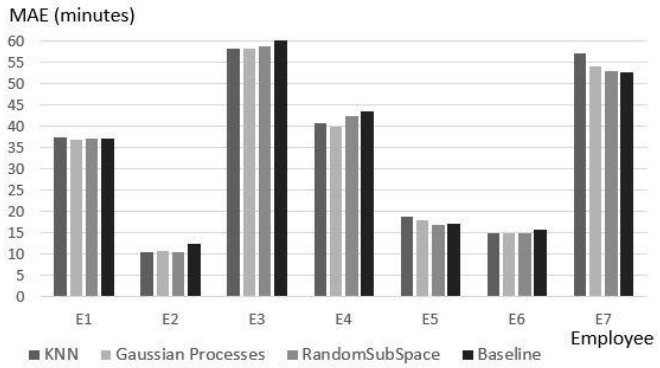

Figure 1: MAE for k-NN, Gaussian Processes, Random Sub-Space and Baseline for the 7 employees - Sliding Window Technique.

Figure 2 shows the results for the models with the three approaches k-NN, Gaussian Processes and Random Sub-Space compared to the baseline approach for each of the 7 employees. Compared to the baseline approach for employees 1 to 4 and employee 6 we have improvement from 1% to 11%. For employees 5 and 7 the baseline approach is better.

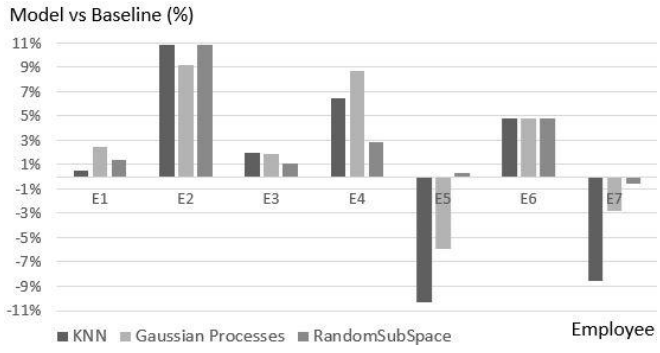

Figure 2: MAE improvement for k-NN, Gaussian Processes and Random Sub-Space compared to Baseline for 7 employees – Sliding Window Technique.

4.1.2 Expanding window

The results for the second technique, i.e., expanding window, are shown in Figure 3. The results are shown for the models learned with the three approaches: k-NN, Gaussian Processes and Random Sub-Space compared to the baseline approach for each of the 7 employees.

The results show significant improvement for the prediction performance for employee no 6. Using the sliding window technique, the regression model had the same MAE as the baseline approach (Figure 2.), however, using the expanding window technique the MAE of the regression model is 50% better. Additionally, one can note that for employee no. 6 this approach is quite good, but for employee no. 5 this approach is worse than the baseline approach for 23%. Therefore, further improvements should be proposed.

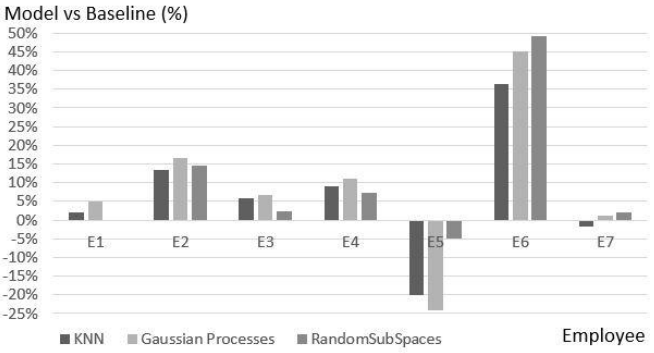

Figure 3: MAE improvement for k-NN, Gaussian Processes and Random Sub-Space compared to Baseline for 7 employees – Expanding Window Technique.

4.1.3 Filtered expanding window

The third tested technique is expanding window technique combined with outlier pre-processing. The results are shown in Figure 4. The results show the improvements in the MAE values with the three approaches k-NN, Gaussian Processes and Random Sub-Space compared to the baseline approach for each of the 7 employees. If we consider the models learned with Gaussian Processes technique we can see that the improvement in the MAE compared to the baseline approach varies from 0% to 50% for different employees.

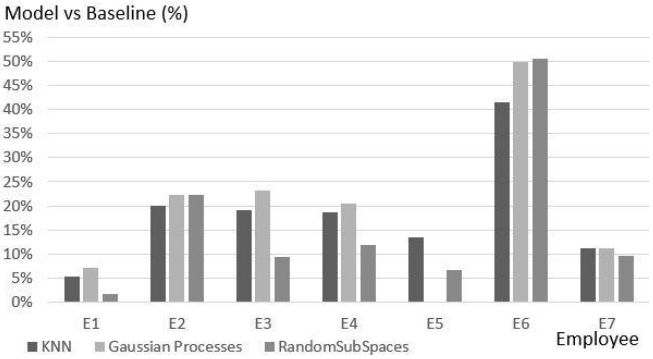

Figure 4: MAE improvement for Knn, Gaussian Processes and Random Sub-Space compared to Baseline for 7 employees – Filtered Expanding Window Technique.

Compared to the other two techniques filtered expanding window technique is the best because each model learned with this technique has a lower MAE than the baseline approach. The only exception of this is the model learned with Gaussian Processes for employee 5 which has the same MAE as the baseline approach. On the other hand if we consider the MAE for the same employee (Figure 1) we can see that it is only 10-15 minutes which means that this employee has a very regular schedule and that is the reason why the baseline approach has low MAE that is difficult to improve.

4.2 Predicting the departure time of an employee

Predicting the arrival time and the departure time of an employee are two problems that are similar in nature. Because of that similar approaches are used. First, models are learned with the sliding window technique, than with the expanding window technique and finally with the filtered expanding window technique. With the first two techniques we tried several different regression learning approaches of which 3 (k-NN, Gaussian Processes and Random Sub-Space) with the smallest MAE were chosen for further experiments. The results are quite similar as with the previous problem (predicting the arrival time). The models learned with the first two techniques have similar or in some cases even worse MAE that the baseline approach.

In Figure 5 we can see the MAE for the models learned with k-NN, Gaussian Processes and Random Sub-Space with sliding window technique. The MAE value is shown on the y-axis and is represented in minutes. We can see that it varies from 40 minutes for employee 2 to 80 minutes for employee 3 and 4. This values are higher than those for MAE for predicting the time of arrival shown in Figure 1. This means that for each employee there is more irregularities in the departure time that in the arrival time.

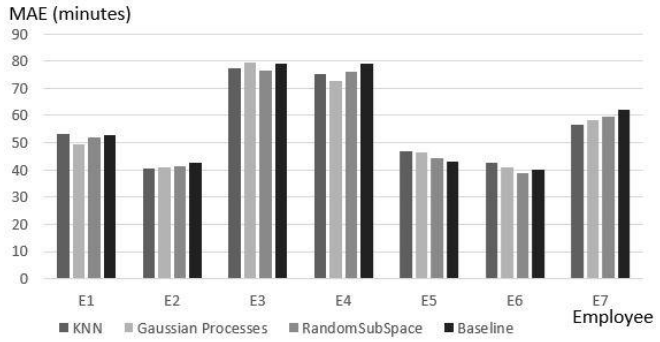

Figure 5: MAE for k-NN, Gaussian Processes, Random Sub-Space and Baseline for the 7 employees - Sliding Window Technique.

For predicting the departure time of an employee the best results are achieved with the filtered expanding window technique. There is an improvement in MAE compared to the baseline approach.

In Figure 6 we can see the results for the models learned with the three approaches k-NN, Gaussian Processes and Random Sub-Space for each of the 7 employees. The only model that is better than the baseline approach for all employee is Random Sub-Space. But if we compare Random Sub-Space with k-NN and Gaussian Processes there are cases where one of the other two models is much better. For example for employee 4 Gaussian Processes is better than Random Sub-Spaces for 80%.

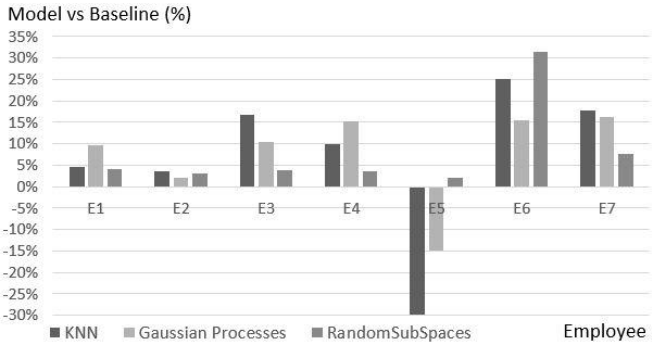

Figure 6: Knn, Gaussian Processes and Random Sub-Space compared to Baseline model for 7 employees – Filtered Expanding Window Technique.

5 CONCLUSIONS

The paper presented an approach to predicting the arrival and departure time of an employee. Three techniques for selecting the learning data were implemented and tested. The results showed that the best performing technique is the filtered expanding window technique. Additional analysis for finding the most suitable regression learning algorithms showed that k-NN, Gaussian Processes and Random Sub-Space perform the best according to the MAE value.

The results for the arrival time prediction showed improvements in the MAE values compared to the baseline naïve approach. The improvements vary from 6% to 50% depending on the employee. The results for the departure time prediction also showed improvements and vary from 2% to 32% depending on the employee.

References

[1] TherML: Occupancy Prediction for Thermostat Control Christian Koehler, Brian D. Ziebart, Jennifer Mankoff, Anind K. Dey - UbiComp'13, September 8–12, 2013, Zurich, Switzerland.

[2] Slovenian meteorological data. http://meteo.arso.gov.si/met/en/.

[3] D. Aha, D. Kibler (1991). Instance-based learning algorithms. Machine Learning. 6:37-66D. Aha, D. Kibler (1991). Instance-based learning algorithms. Machine Learning. 6:37-66

[4] David J.C. Mackay (1998). Introduction to Gaussian Processes. Dept. of Physics, Cambridge University, UK

[5] Tin Kam Ho (1998). The Random Subspace Method for Constructing Decision Forests. IEEE Transactions on Pattern Analysis and Machine Intelligence. 20(8):832-844.