AN ENSEMBLE REGRESSION APPROACH FOR BUS TRIP TIME PREDICTION

João M. MOREIRA[*]
Jorge Freire de SOUSA[†]
Alípio M. JORGE[‡]
Carlos SOARES[§]

Abstract. This paper is about bus trip time prediction in mass transit companies. We describe the motivations to accomplish this task and how it can support operational management on such companies. Then, we describe a Data Mining framework that recommends the expected best regression algorithm(s), from an ensemble, to predict the duration of a given trip. We present results that show the advantage of using an ensemble regression approach.

1. Introduction

In recent years, several mass transit companies all over the world have done large investments in Advanced Transportation Management Systems [9]. These investments appeared mostly as a consequence of the enormous evolution in the areas of information and communication systems. The new Operational Control Systems (including GPS) are among the most well known examples of this evolution. They store location data per time unit as well as all the relevant records for operational control, such as, information about the bus, the driver, the crew duties and the vehicle duties, etc. This type of information is particularly important for the optimization of resources, namely, drivers and vehicles.

2. Motivation and description of the study

Mass transit companies use both the time of the planned and the effective duties in various ways. Of course, the closer the real duty is to the planned one the better, because the use of the resources and the quality perceived by the clients will both be improved. The vehicle

[*] Faculty of Engineering, University of Porto, Portugal, jmoreira@fe.up.pt
[†] Faculty of Engineering, University of Porto, Portugal, jfsousa@fe.up.pt
[‡] Faculty of Economics, University of Porto, Portugal, amjorge@liacc.up.pt
[§] Faculty of Economics, University of Porto, Portugal, csoares@liacc.up.pt

and the crew duties are defined as a sequence of trips, with a certain frequency and each one with a given duration. If the real trip time is longer than planned and the lag between consecutive trips is not enough to avoid overlapping, this will have consequences on the amount of extra money the company has to pay to the driver as well as on the image of the company. Trip time prediction may be used in four different ways:

- In the long term (several months): for the timetable definition;
- In the medium term (it can go from few days to several months, depending on the company's culture): for the logic definition of crew duties and rostering;
- In the short term (few days): for the assignment of each duty to a driver;
- In runtime (minutes): to provide information such as expected arrival time, via SMS or electronic boards.

Trip time prediction may be useful in different ways depending on the type of operational management and control made by the company. There are bus lines over which the control is made by timetable, like the ones used by commuters, for instance; there are others managed and controlled by frequency, as it is the case of the entirely urban ones, at least during peak hours. The study on the factors explaining trip time prediction shall not be undertaken in this paper. Although this is also very important for mass transit companies, namely for negotiation with decision makers on transport policies, it is a different problem, not necessarily solved with the same techniques.

Our case study is STCP (www.stcp.pt), the public bus operator in Porto, Portugal.

3. Objectives and methodology

The broader aim of our research project is to design a decision support system for the operational management of mass transit companies in all the areas discussed in section 2. In this paper we focus on trip time prediction for the short term, i.e., the assignment of the duties to the drivers 3 days in advance. The goal is to reduce the cost of overtime pay to drivers by cutting the difference between planned and real trip time.

Considering the on line arrival of every new trip, the input flow can be regarded as a data stream. These data are continuously changing, thus requiring a continued Data Mining (DM) approach. The DM models proposed adapt to new available data, without human intervention. The whole DM project is approached using the CRISP-DM methodology ([7]). This methodology comprises 6 phases: business understanding, data understanding, data preparation, modelling, evaluation and deployment. The methodology is continuously applied, since the business goals may change with time.

4. The architecture model

The architecture model we are developing (figure 1) is an ensemble regression framework, i.e., it uses an ensemble of algorithms / parameter sets (a&ps), in order to select, for each value to predict, an a&ps to train and to predict the trip time. It has three main components: pre-processing, recommendation and prediction components.

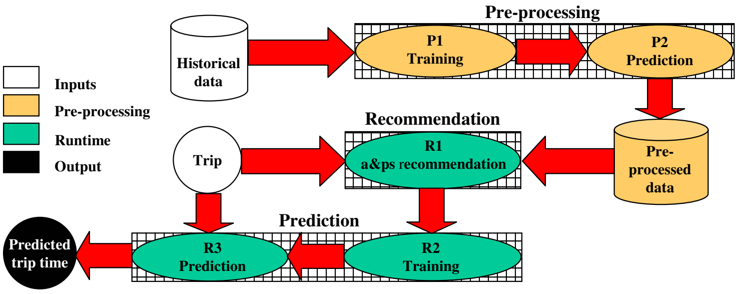


Figure 1. The architecture model

The main goal of the pre-processing component is the daily update of the predictions for the next coming days obtained by the different a&ps that have been previously selected. The objective of the pre-selection of n a&ps (discussed in the next section) is to minimize the set of a&ps used by the model without loosing prediction accuracy. The pre-tested algorithms were random forests ([1]), projection pursuit regression ([2]) with three different smoother methods (super smoother, spline and generalized cross-validation spline), and support vector machines ([8]) with three different kernels (linear, radial and sigmoid). The pre-processing component has two tasks: training and prediction. Each one of the a&ps are trained and the new values from the historical database are predicted (see, for example, [3]). These new predictions are stored in the pre-processed database. The pre-processing component runs everyday, typically, during the night.

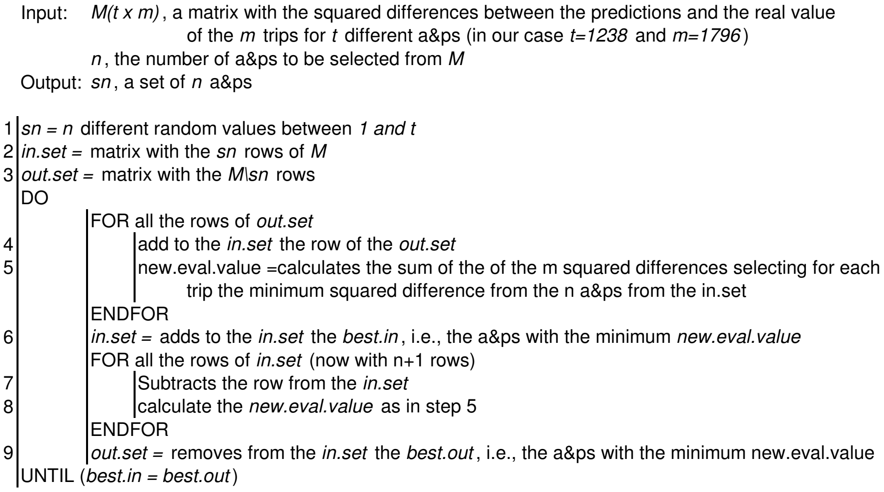
The recommendation component uses ensemble regression. There are two approaches to implement it: combination or dynamic selection ([6]). The first one combines predictions of several recommended a&ps while dynamic selection chooses just one a&ps. Both approaches use the pre-processed information to recommend the a&ps (one or more). The recommendation is done according to the input value we want to predict. Figure 1 assumes the dynamic selection approach.

The prediction component uses the output of the recommendation component to train the model(s) and uses the trained model(s) to predict the value for the new trip.

5. A study on the pre-selection of n a&ps

The use of several algorithms instead of just one for trip time prediction is motivated by previous experimental work which showed that the best algorithm for each region of the instance space is different from region to region [5]. In the present work we use the regression algorithms referred in the previous section with several parameter sets. 1238 was the total number (t) of a&ps tested. The result for each a&ps is the variation index of the trip time calculated as $variation.index = \frac{\sqrt{\sum(p_i - r_i)^2/m}}{\sum(r_i)/m}$ where p_i and r_i represent, respectively, the predicted and the real trip times for $i = 1, ..., m$, where m is the number of observations. The best overall result was 9.92% and was obtained using random forests.

The hypothesis for our experiments was that an ensemble of regressions could improve predictions. What we present is an heuristic for the pre-selection of an ensemble of n a&ps (figure 2). We also present the influence of the n value on the best possible variation index for the ensembles obtained with our heuristic. Figure 3 presents the average for 10 runs for each different value of n. Obviously, the values for $n = 1$ and for $n = 1238$ are unique. The heuristic obtains local minimum but the standard deviation is quite small for all values of n.



```
Input:  M(t x m), a matrix with the squared differences between the predictions and the real value
                  of the m trips for t different a&ps (in our case t=1238 and m=1796)
        n, the number of a&ps to be selected from M
Output: sn, a set of n a&ps

1 sn = n different random values between 1 and t
2 in.set = matrix with the sn rows of M
3 out.set = matrix with the M\sn rows
  DO
        FOR all the rows of out.set
4           add to the in.set the row of the out.set
5           new.eval.value =calculates the sum of the of the m squared differences selecting for each
                trip the minimum squared difference from the n a&ps from the in.set
        ENDFOR
6       in.set = adds to the in.set the best.in, i.e., the a&ps with the minimum new.eval.value
        FOR all the rows of in.set (now with n+1 rows)
7           Subtracts the row from the in.set
8           calculate the new.eval.value as in step 5
        ENDFOR
9       out.set = removes from the in.set the best.out, i.e., the a&ps with the minimum new.eval.value
  UNTIL (best.in = best.out)
```

Figure 2. Heuristic for the pre-selection of n a&ps

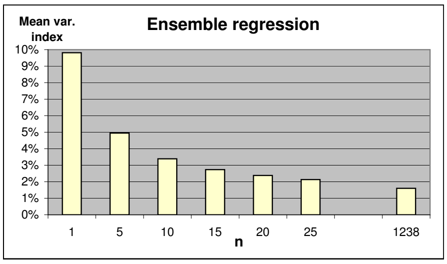


Figure 3. Mean variation index oracle for n a&ps

With this approach we expect to be able to reduce meaningfully the variation index in comparison to the use of just one a&ps.

6. Current status of the framework development

Up to now we have conducted extensive tests with different a&ps. This preliminary work is necessary for a pre-selection of the a&ps and for obtaining the basic statistics needed as input for the heuristic. The selection of the a&ps set size (n) will be evaluated after the implementation of the recommendation component ([4]). This component is currently being developed. The software for the updating of the pre-processed database will be implemented soon.

Acknowledgments

This work was partially supported by FCT (POCT/TRA/61001/2004) and FEDER e Programa de Financiamento Plurianual de Unidades de I&D.

References

[1] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[2] J. H. Friedman and W. Stuetzle. Projection pursuit regression. *Journal of american statistical regression*, 76(376):817–823, 1981.

[3] T. Hastie, R. Tibshirani, and J. H. Friedman. *The elements of statistical learning: data minig, inference, and prediction*. Spinger series in statistics. Springer, 2001.

[4] C. J. Merz and M. J. Pazzani. A principal components approach to combining regression estimates. *Machine learning*, 36:9–32, 1999.

[5] J. M. Moreira, A. Jorge, J. F. Sousa, and C. Soares. A data mining approach for trip time prediction in mass transit companies. In C. Soares, L. Moniz, and C. Duarte, editors, *ECML/PKDD Workshop Data Mining for business*, pages 63–66, Porto - Portugal, 2005.

[6] N. Rooney, D. Patterson, S. Anand, and A. Tsymbal. Dynamic integration of regression models. In *5th international workshop on multiple classifier systems*, volume LNCS-3181, pages 164–173, Cagliari - Italy, 2004. Springer-Verlag.

[7] C. Shearer. The crisp-dm model: The new blueprint for data mining. *Journal of Data Warehousing*, 5(6):13–22, 2000.

[8] A. J. Smola and B. Scholkopf. A tutorial on support vector regression. Technical Report NC2-TR-1998-030, 1998.

[9] A. Stathopoulos and T. Tsekeris. Methodology for processing archived its data for reliability analysis in urban networks. In *IEE Intelligent transport systems*, volume 153, pages 105–112, 2006.