Neurocomputing 150 (2015) 428–439

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Improving the accuracy of long-term travel time prediction using heterogeneous ensembles

João Mendes-Moreira [a,d,*], Alípio Mário Jorge [b,d], Jorge Freire de Sousa [c,e], Carlos Soares [a,f]

[a] Department of Informatics Engineering, Faculty of Engineering, University of Porto, Rua Dr. Roberto Frias, s/n, 4200-465 Porto, Portugal
[b] Department of Computer Science, Faculty of Science, University of Porto, Rua Campo Alegre 4169-007 Porto, Portugal
[c] Department of Industrial Engineering and Management, Faculty of Engineering, University of Porto, Rua Dr. Roberto Frias, s/n, 4200-465 Porto, Portugal
[d] LIAAD-INESC TEC L.A., Campus da FEUP, Rua Dr. Roberto Frias, 378, 4200-465 Porto, Portugal
[e] UGEI-INESC TEC L.A., Campus da FEUP, Rua Dr. Roberto Frias, 378, 4200-465 Porto, Portugal
[f] INESC TEC L.A., Campus da FEUP, Rua Dr. Roberto Frias, 378, 4200-465 Porto, Portugal

ARTICLE INFO

Article history:
Received 1 January 2014
Received in revised form
6 August 2014
Accepted 8 August 2014
Available online 5 October 2014

Keywords:
Travel time prediction
Long term
Ensemble learning
Regression
Dynamic selection

ABSTRACT

This paper is about long-term travel time prediction in public transportation. However, it can be useful for a wider area of applications. It follows a heterogeneous ensemble approach with dynamic selection. A vast set of experiments with a pool of 128 tuples of algorithms and parameter sets (*a&ps*) has been conducted for each of the six studied routes. Three different algorithms, namely, random forest, projection pursuit regression and support vector machines, were used. Then, ensembles of different sizes were obtained after a pruning step. The best approach to combine the outputs is also addressed. Finally, the best ensemble approach for each of the six routes is compared with the best individual *a&ps*. The results confirm that heterogeneous ensembles are adequate for long-term travel time prediction. Namely, they achieve both higher accuracy and robustness along time than state-of-the-art learners.
© 2014 Elsevier B.V. All rights reserved.

1. Introduction

Long-term bus travel time prediction (i.e., the prediction of the duration of bus trips several days ahead) is very important for the planning activities in freight and transport companies (e.g., definition of the schedules for trips and drivers). Long-term should be, in this context, understood as the prediction of a travel time for a single future trip that is expected to occur in a given future timestamp. Not much attention has been dedicated to this problem. An exception is [19] where three different state-of-the-art regression techniques are empirically compared: Support Vector Machines (SVM), Projection Pursuit regression (PPR) and Random Forests (RF).

In the present paper the use of multiple models, also called ensembles or committees, is studied. The advantage of multiple models with respect to single models has been reported in terms of increased robustness and accuracy [11,30], essentially by reducing the variance component of the error [15].

A three-step process as described in [20] is done: (1) in the generation phase different induction algorithms, namely, Support Vector Machines (SVM), Projection Pursuit regression (PPR) and Random Forests (RF) with different parameter sets are used to generate different models; (2) then a pruning algorithm using forward search is used to reduce the number of models in the ensemble; (3) and finally, a study on the combination of the predictions done by the models in the ensemble is carried on using the dynamic selection approach to ensemble learning. In the dynamic selection approach, the prediction for each instance is based on a subset of the available models, selected according to the characteristics of the instance [23]. This approach is suitable for highly dynamic problems such as the prediction of travel time.

In the following section the travel time prediction problem is described. In Section 3 ensemble learning and a variant of it named dynamic selection are described. Then, the data used in the experiments and the approach used for performance estimation are presented. The experiments done are described along two sections: the first describes experiments done in order to establish the ensemble framework (Section 5) and the second section describes and discusses the experiments and its results for comparison of: (1) the ensemble approach; (2) the best single *a&ps*; and (3) the scheduled travel times (Section 6). Section 7 concludes the paper.

* Corresponding author at: Department of Informatics Engineering, Faculty of Engineering, University of Porto, Rua Dr. Roberto Frias, s/n, 4200-465 Porto, Portugal. Tel.: +351 22 5082134.
E-mail address: jmoreira@fe.up.pt (J. Mendes-Moreira).

http://dx.doi.org/10.1016/j.neucom.2014.08.072
0925-2312/© 2014 Elsevier B.V. All rights reserved.

2. Long-term travel time prediction

Travel time prediction (TTP) for the long-term could be used in order to better plan transportation services. In logistics, in particular for freight transportation, TTP could help to better plan the deliveries. Existing approaches typically use average times on empirically segmented periods: early morning, morning peak period, etc. This can be explained, at least partially, by the lack of data on actual times. Most companies do not monitor routes regularly, which means that there is data only for a reduced number of trips per route. The problem of lack of data is addressed in [29] by enlarging artificially the database.

In order to explain the usefulness of long-term TTP in the operational planning of public transport companies, its main tasks are briefly described in its typical sequential way [5,16,8]:

1. Network definition: It comprises the definition of the lines, routes and bus stops. We define route as an ordered sequence of directed road stretches and bus stops. Lines are a set of routes, typically two routes that use roughly the same road stretches but in opposite directions.
2. Trips definition: A trip is the completion of a defined route made by a vehicle. There are typically two different methods for trip definition: (1) headway-based, defining the time between two successive trips on the same route [34]; or (2) schedule-based, defining timetables by explicitly setting the departure time and the time of passage at the main bus stops. This task is done for each route individually even if they are articulated between groups of routes/lines [6].
3. Drivers and buses duties definition: The goal of both tasks is to define duties. A duty is the work a bus/driver must do. When a duty is defined, in both cases, it is not known which driver or bus will do it. Only a logic assignment is made. The case of buses duties is much simpler than drivers duties for obvious reasons: drivers must stop for lunch, cannot drive every day of the week, etc., i.e., they have much more constraints than buses. According to [8], "each driver duty is subject to a set of rules and constraints defined by the government legislation, union agreements and some internal rules of the company." Typically, buses duties are defined before drivers duties.
4. Duties assignment: It is the task where the drivers duties are assigned to drivers and buses duties are assigned to buses. This assignment is made with an anticipation of a few days or weeks, depending on the company. Moreover, they can be subject to last minute adjustments. In this task a physical assignment is done. Assignment for drivers duties is more complex than for buses duties, for similar reasons to the ones explained above. The assignment of drivers duties to drivers is called rostering. It can vary significantly from one company to another.

Long-term TTP refers to the prediction of travel times for a start time of several hours in the future [14]. In public transportation companies, long- term TTP is typically used for the definition of the trips. For such objective, the prediction should be valid for a long period, for instance, TTP for Monday trips at 8:00 should be as correct as possible for all the period the timetable is used, typically one year or more. However, TTP could also be used for other tasks of the operational planning, namely, drivers and buses duties definition. However, its use for the definition of drivers and buses duties is problematic because: (1) this implies the daily redefinition of the buses and drivers duties (usually they are defined for months); and (2) this should be done without changing the scheduled trips, i.e., the duties should maintain the trips' departure times but not necessarily the travel times. Considering the way the operations are done in a typical public transport company, all this adds complexity to the process of operational planning.

The sequential nature of this planning process and the complexity of these tasks are probably the main reasons for the lack of planning procedures using long-term TTP for buses and drivers duties definition. Not only this process is very time consuming but additionally, once the duties are planned, changes in travel times may have an effect on these planned duties that is hardly predictable. How to use TTP in the planning of public transport companies is, nowadays, an important question. Is it possible to use more flexible planning processes in order to reduce the prediction horizon? How short can this horizon be if the planning is done from the scratch? All these questions have different answers for different companies and there are currently no standard answers to them. The use of long-term TTP can trigger important changes to some of the existing planning procedures.

3. Ensemble learning

An ensemble $\mathcal{F}$ is composed of a set of k predictors, $\hat{f}_i$, for an unknown function f.

$$\mathcal{F} = \{\hat{f}_i, i = 1, \ldots, k\}. \quad (1)$$

The resulting ensemble predictor is denoted as $\hat{f}_f$.

$$\hat{f}_f(\mathbf{x}) = \sum_{i=1}^{k} [h_i(\mathbf{x}) \times \hat{f}_i(\mathbf{x})], \quad (2)$$

where $h_i(\mathbf{x})$ are the weighting functions.

The ensemble learning process has, typically, three steps [20]:

- Model generation: The process of generation of the initial set (the pool) of base models. When all models are generated using the same induction algorithm, the approach is called homogeneous. Otherwise, it is named heterogeneous. The heterogeneous approach is claimed to obtain models with higher diversity [36,37], which is important to increase the accuracy of the ensemble.
- Ensemble pruning: The process of selecting a subset of models (the ensemble) from the pool. This step, also known as pre-pruning, is optional. When pruning is made, the ensemble learning approach is called overproduce-and-choose [27], otherwise it is named direct.
- Ensemble integration: The process of combining the predictions obtained with the models from the ensemble. The integration can be made with constant and non-constant weighting functions [22].

3.1. Dynamic selection

Dynamic selection assumes that the best way to integrate the models should consider the prediction ability of each base predictor on data that is similar to the one we want to make predictions about. So model selection is done on the fly. Given an instance, dynamic selection methods choose the subset of models from the pool that will be combined to make a prediction based on the characteristics of that instance. Dynamic selection can be seen as a kind of pruning on the fly, being also known as post-pruning. The weights of the integration function can also be calculated on the fly.

Fig. 1 summarizes the dynamic approach [18], which is divided into the sub-steps described next. Note that this assumes that the training data provided to the ensemble learning process is divided into two subsets. The first set contains data used to induce the pool of models. The remaining data, referred to as validation set, is used only for evaluation purposes by the dynamic selection methods. The sub-steps are:

1. Data selection: Given a new test example, $\mathbf{x}$, find similar data in the validation set.

Download English Version:

https://daneshyari.com/en/article/407535

Download Persian Version:

https://daneshyari.com/article/407535

Daneshyari.com