MoReBikeS - Model reuse with bike rental station data

Meelis Kull[1], Nicolas Lachiche[2], and Adolfo Martínez-Usó[3]

[1] Intelligent Systems Laboratory, University of Bristol
Meelis.Kull@bristol.ac.uk
[2] ICube Laboratory, Université de Strasbourg
nicolas.lachiche@unistra.fr
[3] Departament de Sistemes Informàtics i Computació, Universitat Politècnica de València
admarus@upv.es

1 The challenge

Adaptive reuse of learnt knowledge is of critical importance in the majority of knowledge-intensive application areas, particularly when the context in which the learnt model operates can be expected to vary from training to deployment. In the MoReBikeS challenge (Model Reuse with Bike Rental Station Data) that we organised as the ECML-PKDD 2015 Discovery Challenge #1, we decided to focus on model reuse and context change. In contrast to most of the machine learning challenges to date we provided the participants not only with data, but also with pre-trained models representing knowledge to be reused.

The MoReBikeS challenge was carried out in the framework of historical bicycle rental data obtained from Valencia, Spain. Bicycles are continuously taken from and returned to rental stations across the city. Due to the patterns in demand some stations can become empty or full, such that more bikes cannot be rented or returned. To reduce the frequency of this happening, the rental company has to move bikes from full or nearly full stations to empty or nearly empty stations. This can be done more efficiently if the numbers of bikes in the stations are predicted some hours in advance. The quality of such predictions relies heavily on the recorded usage over long periods of time. Therefore, the prediction quality on newly opened stations is necessarily lower.

In this challenge the participants were required to predict the number of bikes 3 hours in advance on 75 stations with a short (1 month) recorded history, by making use of the provided predictive linear models trained on other 200 stations with longer history (more than 2 years). To promote reuse of the provided models we did not reveal the long historical data to the participants (except for 10 stations out of 200). The predictions were evaluated using mean absolute error in the number of bikes every hour over the next period of 3 months.

The common problem in evaluating predictors of time-series are the temporal dependencies: the features of later test time-points leak information about the earlier test time-points. We tackled this problem by setting the challenge in two stages: *the leaderboard challenge* and *the full test challenge*. In the leaderboard

challenge we used a small sample of test time-points to minimise information leaks. This allowed us to have weekly leaderboards while minimising the chances of some participants cheating in the sense of using test instances from future to improve predictions. In order to enter the full test challenge the participants were required to commit to a single prediction method by providing source code and describing the chosen method in a short paper. They were subsequently provided with full test data to perform full evaluation of their methods (which in principle we could have replicated using their code). This way we could make sure that none of the entries in the full test challenge was using test instances from future to improve predictions. The final results of the challenge were announced based on the evaluation on full test data.

2 The dataset

The original gathered dataset[4] provided information about 275 bike rental stations in Valencia over a period of 2.5 years (from 01/06/2012 to 31/01/2015). For each hour in this period the data specified the median number of available bikes during that hour in each of the stations. The dataset was complemented with weather information about the same hour (temperature, relative humidity, air pressure, amount of precipitation, wind directions, maximum and mean wind speed).

We first split the stations randomly into 200 training stations and 75 test stations (Fig.1). The time period was split into training period (01/06/2012 to 31/10/2014) and three-months test period (01/11/2014 to 31/01/2015). The last month of the training period (01/10/2014 to 31/10/2014) we referred to as *the deployment period*. We trained 6 different linear models (more details below) for each of the 200 training stations on the training period. The participants were provided with the trained models, with the data from the one-month deployment period for all 200+75 stations, and with the data from the training period from 10 training stations out of 200. In order to promote the reuse of provided models we decided not to give the full training data for the remaining 190 training stations. The leaderboard test data contained data from 25 test stations out of 75, with just one test hour every 3 days across 2 months (01/11/2014 to 31/12/2014). The full test data were revealed only to the participants after they provided a short paper and their code. These data contained the other 50 test stations with every hour across the three-month test period.

The task was to predict the number of bikes 3 hours in advance, so it was important to know what the number was 3 hours earlier, at the moment of making the prediction. We constructed this feature and called it 'bikes_3h_ago'. Bike rental data have strong weekly periodicity, thus we constructed some features based on the data across all previous weeks. In particular, the feature 'full_profile_bikes' was the average number of bikes in the same station in the same hour of the week (e.g. Monday 8am-9am) across all previous weeks in the

[4] All the data and models together with detailed information are available at http://reframe-d2k.org/Challenge_Download.

data. Another feature 'full_profile_3h_diff_bikes' was the average difference in the number of bikes between the same hour of the week and 3 hours earlier, across all previous weeks. In test data the participants had only the one-month deployment period to estimate the profiles. This resulted in profiles with much weaker predictive value, potentially harming the performance of models which strongly rely on the full profile features. Therefore, we also introduced 'short_profile_bikes' and 'short_profile_3h_diff_bikes' as features which used information from past 4 weeks only.

For each of the 200 training stations we learned 6 linear models to predict the number of bikes, corresponding to 6 different subsets of the features 'bikes_3h_ago', 'full_profile_bikes', 'full_profile_3h_diff_bikes', 'short_profile_bikes', 'short_profile_3h_diff_bikes' and temperature. All the subsets included 'bikes_3h_ago' but differed based on which profile features they used (3 options: full profiles, short profiles, or all profiles), and whether they used temperature (2 options: yes or no). The obtained 200×6 models were provided to the participants.

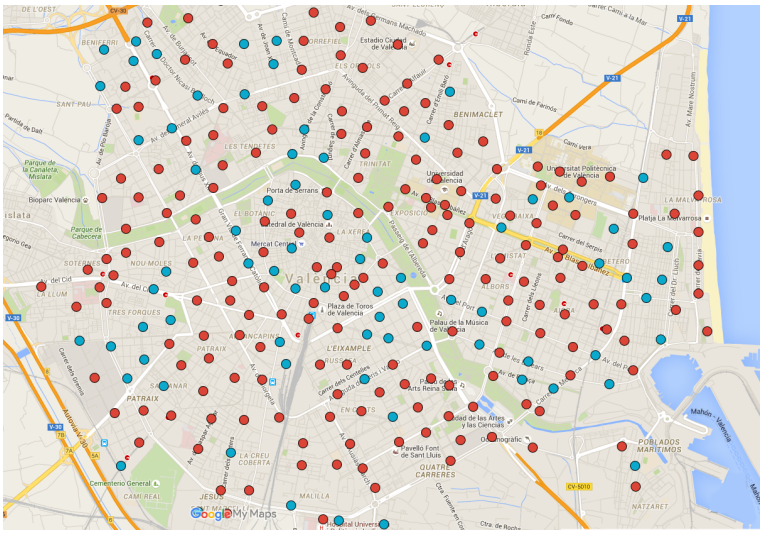

Fig. 1. Map of Valencia (created using Google My Maps) with 200 bike rental stations used for training linear models (marked in red) and 75 stations used for testing (marked in blue).

3 The contest and the papers

The first stage of the context attracted 116 submissions from 23 teams. The limit of 12 submissions per team was imposed in order to limit overfitting, as we

selected the best student based on the leaderboards[5]. 10 teams decided to enter the second stage and committed to a single prediction method by providing their source code and describing the chosen method in a short paper. These methods were evaluated with respect to mean absolute error (MAE) on the full test data set covering every hour in the three-month test period across 50 test stations. These proceedings include the papers of the top 3 teams: Hao Song with Peter Flach from the University of Bristol, UK (MAE=2.0143); Yu Chen with Peter Flach from the University of Bristol, UK (MAE=2.0515); and Arun Bala Subramaniyan with Rong Pan from the Arizona State University, USA (MAE=2.0667).

The methods across the 10 teams in the full test challenge covered a good range of approaches. 3 teams performed an analysis to select for each test station one model out of the given 1200, and used that model for prediction (including the 3rd best team). 3 teams selected multiple models and averaged over these (including the winning team), and 1 team used a weighted average. 3 teams decided not to use the given model and trained new models (including the second-best team).

In choosing the models to be reused the teams had to decide on the criteria for model suitability for a given test station. These included: performance of the model in the test station during the deployment period; distance between the test station and the station of the model's origin; similarity between the time-series of the stations during the deployment period; and several combinations of these. Most of the teams used two simple methods to improve predictions: clipping to make sure that the predicted number of bikes does not exceed the size of the station; and rounding the predictions to the closest integer, as mean absolute error tends to decrease after rounding.

Acknowledgements

We are grateful to all the members of the Reframe project in setting up the MoReBikeS challenge. We thank the Altocumulo weather station and Biciv.com. We also thank all our sponsors: the Cátedra InnDEA of València and Cyclocity companies, the Universitat Politècnica de València, the ECML-PKDD 2015 conference, and the Reframe project granted by the European Coordinated Research on Long-term Challenges in Information and Communication Sciences & Technologies ERA-Net (CHIST-ERA) and funded by MINECO in Spain (PCIN-2013-037).

[5] The results are available at http://reframe-d2k.org/Challenge.