Reusing Models Using K-Nearest Neighbors And Weighted Arithmetic Mean To Predict Future Use Of Bike Stations For The MoReBikeS Challenge 2015

Víctor Núñez Monsálvez

Master in Big Data, Universidad Politécnica de Valencia, Spain

vicnumon@masters.upv.es

Abstract.
The purpose of this paper is to provide a detailed explanation about an approach taken in the context of the MoReBikeS Challenge 2015. The aim was how accurate it could be possible to predict the behavior of the bike rental station use three hours beforehand. For that problem, some models and data were provided with the purpose of mixing in such manner that accurate predictions in terms of mean absolute error were finally obtained.

Keywords: distance matrix, K-Nearest Neighbors, mean absolute error, prediction model, weighted arithmetic mean.

1 INTRODUCTION

The challenge is about a bike rental system with 275 bike stations: 200 old stations (existing from the beginning of that system, and numbered from 1 to 200) and 75 new stations (only working one month at the prediction time, and numbered from 201 to 275). These 75 new stations are the main objective of the predictions and will be referred to as the target station henceforth.

The aim of the challenge consists of predicting the number of bikes three hours beforehand for the aforementioned 75 different target bike rental stations, provided:

- a month training data for the 75 target stations (201:275), containing profile information that was calculated using one month data (October, 2014);
- a month training data for the other 200 stations (1:200), containing profile information that was calculated using two years data (June, 2012 to September, 2014);
- linear models for others 200 stations (1:200). These models were induced using two years data which contains date, weather and profile information.

The main objective of this work is building a system that is capable of predicting the number of bikes three hours in advance for the 75 target bike stations, for two

adfa, p. 1, 2011.
© Springer-Verlag Berlin Heidelberg 2011

months after the month from which the one-month training data were sampled, using only the above-mentioned limited training data and linear models.

2 EVALUATION

As the training data provided only comprises one month, the approach used in this paper did not have the intention to train any model, new or existing. Instead, the decision made was to use this one-month training data in order to determine which model was the best one for each station according to the resulting MAE (mean absolute error).

Because of that lack of training data, it was supposed that historical models about two-year data on the old bike stations would yield better predictions that the scarce training data. Therefore, the hypothesis made was that the closest old stations to the target stations were most capable to predict future use of those new stations given the different models for the other 200 stations. For that reason, distance was a crucial point in weighting the predictions of the

3 K-NEAREST NEIGHBORS METHOD DESCRIPTION

In the selected solution, a combination of the predictions from the K nearest stations was proposed. This approach was founded on the principles of the K-Nearest Neighbors (K-NN) machine learning algorithm.

The approach to the problem consisted of combining the predictions of the K nearest stations – among the old stations (1:200) - to the target stations (201:275) using the weighted arithmetic mean. On one hand, these predictions were calculated applying the best model – in terms of MAE - for each old station (1:200). On the other hand, the K nearest neighbors were obtained by comparing each target stations (201:275) to all the old stations (1:200) in terms of the Euclidean distance between them. Then, the K closest old stations to one target station were selected as its K nearest neighbors. In doing so for every target station (201:275), their K nearest neighbors were discovered among the old stations (1:200).

Being dist $_{i,k}$ the Euclidean distance between the target station corresponding to the i-th observation within the test data and the k-th nearest neighbor (station) - from the set of K nearest neighbors for each target station- , and pred $_{i,k}$ the prediction for the target station corresponding to that i-th observation using the best model for the k-th nearest neighbor, the prediction for the i-th observation within the test data was given by equation 1.

$$mypred_i = \frac{\sum_{k=1}^{K}(dist_{i,k})\times(pred_{i,k})}{\sum_{k=1}^{K} dist_{i,k}}$$

The Euclidean distance between the target station and its neighbors, dist $_{i,k}$,was used to weight the influence of their predictions, pred $_{i,k}$, on the final prediction. Finally, this summation was divided by the sum of the k Euclidean distances from each neighbor – among the K nearest neighbors – to the target station on the test data. In doing so, the final prediction value was obtained from k predictions taken into account in a different importance according to their proximity to the target station.

In conclusion, the predictions for every observation within the test data were calculated by means of a weighted arithmetic mean of the predictions of the K nearest old stations (1:200) using the best models in terms of MAE on the one-month training data weighted by the Euclidean distance to the target station.

4 RESULTS AND CONCLUSION

Various values for K - the number of nearest stations - were tried in different submissions. The mean absolute error obtained when test data for small data challenge - comprising data from October until November, 2014 - were applied to the approach explained in this paper are showed in table 1 for different values of K.

Submission **number**	**Neighbour** **number (K)**	**Mean Absolute** **Error (MAE)**
2	*1*	*2.722*
3	*5*	*2.430*
4	*10*	*2.434*
5	*25*	*2.416*
6	*50*	*2.430*
7	*100*	*2.444*

Table 1. Small data challenge results

From this results it can be inferred that K = 25 is the optimum value in order to minimize the mean absolute error for the small data challenge.

Although this approach reached a satisfactory performance for the small data challenge, it could be expected a worse performance when applied to the full data challenge because it relies strongly on the linear models fitted to another situation. In that context, there were only 200 stations. Therefore, those models do not characterize accurately the new situation with 75 new stations – the target stations.

5 REFERENCES

1. MoReBikeS challenge main page (http://reframe-d2k.org/index.php/Challenge), 2015.