Selecting Training Data By Evaluating Existing Models

Reusing models for the MoReBikeS Challenge

Denis Moreira dos Reis
Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo
denismr@usp.br

ABSTRACT

This paper provides an explanation about three approaches for the MoReBikeS Challenge [1]. While the first of them was submitted to the competition, the other two, although performing better than the submitted solution, were developed only after the deadline for participation. The submitted solution consists in using the provided models to select sources of data when building new training sets. The new data sets are supposed to have greater number of instances and features with better quality. The other solutions are based on making an ensemble with the provided models, weighing them according to some criteria of suitability on the assessed data.

1. INTRODUCTION

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

- linear models for other 200 non-target stations: these models were induced using two years data which contains date, weather and profile information;
- a month of training data for the 75 target stations, containing profile information that was calculated using one-month data;
- a month of training data for the 200 non-target stations, containing profile information that was calculated using two years data.

Additionally, two years of data were provided for ten non-target stations, however, in this work, these data were not used when building the final solution. Instead, they were used in an evaluation procedure when choosing the best approach to be submitted.

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 a period of three months after the month from which the training data was sampled, using only the provided and limited training data and the linear models.

For this end, this paper suggests three different approaches. One of them, which was submitted to the challenge, tries to compensate the lack of data from the target rental stations by choosing data from other stations as a replacement. While the final model for a particular station does not reuse any of the provided models, they are used to select the replacing data. The other two approaches were developed only after the deadline for the competition. However, as both of them perform better than the submitted solution and reuse the provided models in a more direct way, their explanation is valuable.

This work is organized as follows: Section 2 explains how the submitted solution was chosen, among the solutions available before the deadline, and how the metrics presented in this work were calculated; Sections 3, 4 and 5 explain the three different approaches for the problem; 6 presents the conclusion for this work.

2. CRITERIA FOR CHOOSING THE SUBMITTED ALGORITHM AND OBSERVATIONS

Participating in the challenge involved choosing the best available solution to be submitted, as there were different possibles strategies at hand before the submission deadline. For this purpose, each of the approaches was evaluated using part of the supplied full data for 10 non-target stations. All data sampled in August 2014 were used to induce models that were then evaluated over the data sampled during September and October for the same year. The best performing approach was then submitted while the other approaches were discarded.

However, the presented paper date later than the submission of the results for the full test data set. As it has $99,99\%$ of its correct predictions available to competitors, the values for *Mean Absolute Errors* (MAE) that are shown at the end of each of the following sections were obtained by evaluating the suggested solutions on this data set.

One final observation is that, in this work, it is assumed that any cited linear model produces only integer values, by rounding the result from its linear equation.

3. SUBMITTED SOLUTION

This section explains the solution that was used in the competition. The rationale behind this approach is that a model specifically built for a rental station will, on average, perform better than a model induced for all stations. The offered baselines, although being handmade instead of induced upon real data, are an example of the latter case. Building specific models need sufficient amount of data for the inducing process. However, the provided data for each target station lack both quantity of examples and quality. While the lack of quantity is directly perceived, the lack of quality is given by the poorly calculated key features *full profile 3h diff bikes* and *full profile bikes*, since they are based upon all previous instances and, particularly for the target stations, there is no history prior to their a-month training data, while for the other stations there is a two year history.

Alternatively, it is possible to use data sampled from different stations instead of data sampled from the target stations, provided that they are somehow similar. Here, we introduce the concept of *suitability*. Data with higher suitability for a particular station should be a better replacement for this station than other with lower suitability.

Particularly for this problem, the suitability is measured as follows. For each target station and each non-target station, we calculate the MAE – defined as follows

$$\text{MAE}(X, m) = \frac{1}{|X|} \sum_{x \in X} |m(x) - y(x)|$$

where X is a set of instances, m is a prediction model and $y(x)$ is the correct value for the prediction – using the target station's training data and the non-target station's model which file name matches the expression *"model_station_[0-9]+_rlm_short_full"*. These models contain the features: *temperature.C*, *bikes 3h ago*, *full profile 3h diff bikes*, *full profile bikes*, *short profile 3h diff bikes* and *short profile bikes*. The lower is the obtained *MAE*, the higher is the suitability.

Suppose that there is interest in making predictions for future data sampled from a particular target station. In order to build a linear model that will be used for this purpose, we need to have a single data set. The way it is done is by joining all the available data that were sampled from those stations which related models presented the K top suitabilities, where K is a parameter. Once this data set is built, the model can be induced through *MM Estimation*. The *R*[2] package *robustbase*[3] contains the function *lmrob* that is used in this work, in order to induce the models. Particularly, the adopted features are the same as the ones used by the models that were applied when calculating the suitabilities, as they, through informal testing, seem likely to produce good results.

Unfortunately, due to the lack of time, $N = 20$ was chosen by guess, resulting in a MAE of 2.173 when evaluating the approach over the competition's final test data.

4. ALTERNATIVE APPROACH

This section explains a second solution for the challenge that, although not submitted to the competition, performs better than the previous solution.

One more straightforward approach than the previous one would be simply applying the already provided model that produces the best performance, in terms of MAE on the one month training data for a target bike station, to forecast the number of bikes on future data from that station. A subtle elaboration of this solution would be using the average of the predictions obtained from the K best models. Formally, for a particular station, the predicted number of bikes $\hat{y}(x)$ for an example x is calculated as stated in Equation 1, where f_k is the k-th best-performing model on the one month data sampled from the given station and $\lfloor x \rceil$ is the nearest integer to x.

$$\hat{y}(x) = \left\lfloor \frac{1}{K} \sum_{k=1}^{K} f_k(x) \right\rceil \tag{1}$$

The reasoning behind choosing a value for K greater than 1 is that the resulting prediction is less likely to be overfitted to the training data set, once it is incorporating a wider range of possible solutions, even though they apparently have worse quality. It is worth noting that, as all models are evaluated, models that are different *i.e.*, use different features, but were induced upon data sampled from the same station, can appear more than once among the K best fitting models.

For $K = 1$, the achieved MAE was 2.146, while for $K = 5$, it was 2.096.

5. SECOND ALTERNATIVE APPROACH

This section explains the best performing solution in this work, although it was not submitted to the competition. As it performed consistently better than the previous solutions, its explanation is more detailed.

One of the issues that is present in the previous two solutions is the assumption of a strict relation between the similarity of data and the station from which they were sampled. As a second issue, they also assume that the data from any particular station can be described by a linear model. This third solution aims addressing the former issue but, as a side effect, it also addresses the latter.

The approach consists of two stages. The first is the clustering of T, where T is the set that contains all instances from all target stations, together. The second stage is performing the prediction.

For the clustering stage, which results in the set C of clusters, it is used the standard *Lloyed's Algorithm* for Kmeans with $K_{\text{LA}} = 50$, where K_{LA} is the number of clusters. The squared euclidean distance is used as dissimilarity function, considering the following features: *temperature.C*, *bikes 3h ago*, *full profile 3h diff bikes*, *full profile bikes*, *short profile 3h diff bikes* and *short profile bikes*. Each of the features are normalized through *Z normalization*, having the mean and standard deviation estimated through T.

Let g_c^k be the k-th best performing model, according to MAE, for the cluster c. In practice, there is only interest on those models where $k \leq K_{\text{SC}}$ and K_{SC} is a parameter that refers to the number of top models in a same cluster (SC). Here, $K_{\text{SC}} = 5$. Now, it is possible to calculate the cluster specific prediction $\hat{y}_c(x)$, for a given cluster c and an instance x, as the average of the predictions made by the K_{SC} best models for that cluster. Formally, $\hat{y}_c(x)$ is defined in Equation 2.

$$\hat{y}_c(x) = \frac{1}{K_{\text{SC}}} \sum_{k=1}^{K_{\text{SC}}} g_c^k(x) \tag{2}$$

The final prediction is a weighted average of all cluster specific predictions, as stated in Equation 3. The weights $w_{c,x}$ are inversely proportional to the distances between x

and its nearest neighbors that belong to each cluster: the closer are the neighbors inside a cluster, the higher is the correspondent weight for that cluster specific prediction.

$$\hat{y}(x) = \left\lfloor \left(\sum_{c \in C} w_{c,x} \right)^{-1} \sum_{c \in C} w_{c,x} \hat{y}_c(x) \right\rceil \quad (3)$$

Formally, let K_{NN} be the number of nearest neighbors that are considered. In this work, $K_{\mathrm{NN}} = 50$. Let N_x be the set of x's K_{NN} nearest neighbors and $d(x, x')$ the euclidean distance between x and x'. Then, the weights $w_{c,k}$ can be calculated as stated in Equation 4.

$$w_{c,x} = \sum_{x' \in c \cap N_x} \frac{1}{d(x, x')^2} \quad (4)$$

In practice, as finding the nearest neighbors for all the testing examples can be highly time-consuming, it is used a uniform sample of T for this procedure. Here, this sample counts $10,000$ examples. Alternatively, it is possible to use spatial index structures, such as a *k-d tree*.

In order to choose the values for the parameters, some observations are required. First, K_{LA} is not a sensitive parameter. In fact, it supposedly can be large without negative impact, since there should be no problem in splitting one cluster into two smaller, although less semantic, groups: if the bigger cluster is more semantically representative, it is expected for the separated clusters to have similar models as their g_c^k; Second, the distance weighing factor enables the use of higher values of K_{NN} without much sensitivity in this parameter. However, as the clusters can eventually be considerably unbalanced, extremes values for K_{NN} are not advisable. At last, the combination of the weighing factors with higher than 1 values for K_{SC} is expected to reduce the overfit on the training data.

For $K_{\mathrm{LA}} = 50$, $K_{\mathrm{NN}} = 50$ and $K_{\mathrm{SC}} = 5$, the achieved MAE was 1.966.

6. CONCLUSION

Although the submitted approach presents a promising strategy to select training data, it is overperformed by the best baseline, which respective MAE is 2.127. However, this can happen due to a problem-specific characteristic, and it is worth checking whether the presented solution offers better performance in other similar problems. As the second alternative approach simply is an evolution of the first alternative one, we can focus on it, only. The performance advantage of this strategy leads to at least two conclusions. First, data similarity is not only related to which station they were sampled from. Second, judging by the fact that the clustering is based on the same features as the ones that are used in the assessed models, the performance growth suggests that a non-linear model for a particular non-target station can overperform the provided best performing linear model. At last, it is interesting to observe that this specific problem presents both time concept and temporal dependence, *i.e.*, the correct number of bikes for a particular station at a given time directly influences the number of bikes in a posterior time. This behaviour could be exploited in a different solution, although performing data stream specific regression is not the objective of the challenge.

7. ACKNOWLEDGEMENTS

I would like to thank the financial supports by the grants 2014/12333-7 and 2015/01701-8, Sao Paulo Research Foundation (FAPESP).

8. REFERENCES

[1] MoReBikeS challenge main page (http://reframe-d2k.org/challenge), 2015.

[2] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2015.

[3] P. Rousseeuw, C. Croux, V. Todorov, A. Ruckstuhl, M. Salibian-Barrera, T. Verbeke, M. Koller, and M. Maechler. *robustbase: Basic Robust Statistics*, 2015. R package version 0.92-4.