Sensor Information Learning: A Comparison of Statistical Learning Approaches

Hélène Le Cadre *
CEA LIST,
91191 Gif-sur-Yvette CEDEX, FRANCE

Cédric Auliac
CEA LIST,
91191 Gif-sur-Yvette CEDEX, FRANCE
*Email: helene.le-cadre@cea.fr

The management of Electric Vehicle (EV) charging infrastructure requires both to optimize the location of the various charge stations composing the underlying network and to provide relevant predictions to the EV drivers and to the utilities managing the charge stations. In this article, we focus on the second task which relies on the capacity to access and manage online information.

The article is replaced in a more general framework of information learning for sensor network. Under such an assumption the charge station is associated with a sensor providing information through periodic data measurements about a California freeway occupancy rates. The objective is to compare the performances of 4 learning approaches either supervised or based on reinforcement learning techniques, to predict the road occupancy rates over a week using a public PEMS database available online on the machine learning repository [2].

1 Description of the used learning methods

In this article, we compare the performances associated to the use of 4 learning processes: (i) Artificial Neural Networks (ANN), (ii) Autoregressive Integrated Moving Average techniques (ARIMA), (iii) Support Vector Regression (SVR), (iv) Regret based minimization

techniques. While the first techniques fall under the paradigm of supervised learning and are rather classical in the machine learning community, the last one corresponds to an original reinforcement learning task [1], [3]. Indeed, to forecast the EV demand we may: either estimate the real demand from the aggregated charge station statistics measured at past time periods (ANN, ARIMA process, SVR), or use an online learning algorithm which optimizes its prediction by taking into account the information gathered in the past trials (Regret based techniques).

The PEMS dataset cited above consolidates time series describing freeway traffic. More precisely, it describes the occupancy rates between 0 and 1 of different car lanes of the San Francisco bay area freeways. We have chosen to consider only the first sensor to comply with our task which considers mono-dimensional time series. As usual in machine learning experiments, we have divided the database into two subsets: a training set that is used to learn the predictive model and a test set that enables us to compute the mean of the absolute errors between predictions and measurements.

To test the ANN on the PEMS dataset, we adapt a classical feed-forward neural network which structure has been fitted experimentally on the training set so as to minimize a least square criterion. It is made of one hidden layer with 5 neurons associated with linear activation functions, a sigmoid function for the output neuron and normalized input neurons. For the ARIMA process, we follow Box and Jenkins method to decompose the data describing the occupancy rates and remove the seasonality. This latter operation can be done manually or automatically by averaging the data over the periods over which a common pattern repeats itself. The SVR algorithm run over the training set gives rise to more than 4.10^3 support vectors which means that the space of days and time periods is very fragmented in terms of road occupancy rates. For the regret based forecasting, we take away the seasonal part of the PEMS dataset like in the ARIMA process case. Having removed the data seasonality, we use the Vovk-Azoury-Warmuth forecaster to model the resulting data [1]. Formally, let x_t be a vector containing the time index and the day index corresponding to each time period t and y_t be the value of the road occupancy rate for each time period t, the Vovk-Azoury-Warmuth forecaster p_t is defined by:

$$\begin{aligned} p_t &= < w_t, x_t > \\ A_t &= I + \sum_{s=1}^{t} x_s x_s^T \end{aligned}$$

$$w_t = A_t^{-1} \sum_{s=1}^{t-1} y_s x_s$$

A specificity of the regret based algorithm is that it can benefit from new data points after the learning step to improve its forecasts. Contrary to ANN, SVR or ARIMA process for which the learning process has to be repeated to take into account recent data, the regret based techniques is an online forecasting tool. To be more precise, it is made of two phases:

- Exploration: the algorithm gathers data so as to extract the most reliable information. The forecaster explores all the possible prediction values and keeps in memory its regret performance.
- Exploitation: the algorithm exploits the information that it has already acquired and optimizes its estimates by selecting the best forecaster.

This two stage process explains why the regret algorithm might need some time to learn over the training set and why its learning capacity might be observed only after a sufficiently large number of iterations.

2 Comparison of the prediction method performances

To compare the performances of the various forecasting methods over the datasets, we use as criterion, the expectation of the mean of the absolute errors: $\bar{\mathcal{L}} \equiv \frac{\sum_{t\in\mathcal{S}} \sum_{y\in\mathcal{Y}} |y_t - y| p_t(y)}{|\mathcal{S}|}$ where the set $\mathcal{S}$ can represent either the training or the test set, $0 < |\mathcal{S}| < +\infty$ denotes its cardinal and $\mathcal{Y}$ is the set containing all the possible road occupancy values over the dataset. At time period t, predictor p_t is a density function over space $\mathcal{Y}$ for the regret based learning method whereas it is a real belonging to $\mathcal{Y}$ in case where ANN, ARIMA process or SVR are used. As a result, in this latter case, the performance criterion can be sligthly modified to give: $\bar{\mathcal{L}} \equiv \frac{\sum_{t\in\mathcal{S}} |y_t - p_t|}{|\mathcal{S}|}$.

In the last row of Table 2, we have tested a naive learning approach based on the repetition of the road occupancy rates over two consecutive weeks belonging to one learning set and then, to the other. We observe that it is worth using elaborate predictions tools over the considered dataset. Indeed over the training set, the regret based approach, ARIMA process and SVR perform better than the naive approach and over the test set, solely, the ANN performs worse than the naive approach. Over the test set i.e., to perform

Learning method v.s. Dataset	Training set	Test set
ANN	0.009524249	0.008149142
ARIMA	0.005313539	0.007109916
SVR	0.007123486	0.006343147
Regret	0.008037567	0.00627537
Naive	0.006709316	0.007844599

Table 1: Comparison of the learning algorithm mean absolute errors.

the prediction task, the regret based algorithm generates the smallest value for the absolute loss followed closely by the SVR.

References

[1] Cesa-Bianchi N., Lugosi G.,“ Prediction, Learning, And Games”, *Cambridge university Press*, 2006.

[2] Frank, A., Asuncion, A., “UCI Machine Learning Repository, [http://archive.ics.uci.edu/ml]”, university of California Irvine, School of Information and Computer Science, 2010.

[3] Le Cadre, H., Potarusov, R., Auliac, C., “Energy Demand Prediction: A Partial Information Game Approach”, in *proc. European Electric Vehicle Congress*, 2011.