Collaborative Data Analysis in Hyperconnected Transportation Systems

Mohammad Nozari Zarmehri, Carlos Soares

Collaborative Data Analysis in Hyperconnected Transportation Systems

Mohammad Nozari Zarmehri and Carlos Soares

INESC TEC, Faculdade de Engenharia, Universidade do Porto (FEUP)
Rua Dr. Roberto Frias, 378, Porto, Portugal
{mohammad.nozari, csoares}@fe.up.pt

Abstract. Taxi trip duration affects the efficiency of operation, the satisfaction of drivers, and, mainly, the satisfaction of the customers, therefore, it is an important metric for the taxi companies. Especially, knowing the predicted trip duration beforehand is very useful to allocate taxis to the taxi stands and also finding the best route for different trips. The existence of hyperconnected network can help to collect data from connected taxis in the city environment and use it collaboratively between taxis for a better prediction. As a matter of fact, the existence of high volume of data, for each individual taxi, several models can be generated. Moreover, taking into account the difference between the data collected by taxis, this data can be organized into different levels of hierarchy. However, finding the best level of granularity which leads to the best model for an individual taxi could be computationally expensive. In this paper, the use of metalearning for addressing the problem of selection of the right level of the hierarchy and the right algorithm that generates the model with the best performance for each taxi is proposed. The proposed approach is evaluated by the data collected in the Drive-In project. The results show that metalearning helps the selection of the algorithm with the best performance.

Keywords: Hyperconnected World, Machine Learning, Metalearning, Data Mining, Intelligent Transportation Systems, Collaborative Data Analysis

1   Introduction

Hyperconnectivity is used to define the interconnectedness of people, organizations, and objects which result from different technology innovations like the Internet, mobile technology and the Internet of Things (IoT) [1]. The hyperconnectivity exists not only in the communication between people but also in the connectivity of cars [2]. In addition, to make the travel and transportation more efficient and more comfortable, the hyperconnectivity is the main driver of innovation [3].

On the other hand, the transportation system is clearly overloaded by congestion in the major cities. For example in the city center of London, the average speed of cars is 14 km/h [4] while the car's speed in the city center of Moscow is around 6km/h [5]. Positively, this can be an opportunity to improve interconnectedness of cars in the city

environment. These cars can be parts of the communication infrastructure for the Intelligent Transportation Systems (ITS) and also offer various opportunities for gathering data about a city by continually sensing events from streets and process sensed data.

Therefore, utilizing vehicular networks as an infrastructure for urban sensing is a cost-efficient way of deploying an urban monitoring system without actually deploying connected sensors [6] (**Fig. 1** shows a snapshot of the communication between moving cars within the city of Porto). Vehicles typically do not have energy constraints. Therefore, cars can be equipped with powerful wireless transmitters, processing units, and sensing devices (vibration sensors, GPS, video cameras, detectors, acoustic detectors, car sensors, etc.).

Thousands of cars which are networked together using wireless communication are able to connect us to a seemingly unlimited data gathered from the city environment. The gathered data can be processed and visualized live, enabling monitoring activities (**Fig. 1**) and better decision making. However, this data can also be serve as the basis for predictive models that can be an informative tool for the decision support systems one step further. Given the availability of the massive amount of data which is geographically distributed, data mining approaches are being used to obtain models for all parts of the city that are integrated into ITS applications [7, 8, 9, 10, 11, 12, 13, 14].

One of the interesting examples of ITS applications is the prediction of trip duration for public transportation [15, 16, 17]. Knowing the prediction of trip duration beforehand can be very instructive for taxi companies, passengers, and drivers to make the right decision for the route planning and scheduling by using the data collected about the taxi's trips by each taxi. **Fig. 1** shows a snapshot of exchanged messages between taxis in the city of Porto.

Fig. 1. A snapshot showing exchanged messages between taxis in the city of Porto [18]

Machine learning algorithms by using the data collected by taxis relate the trip duration to several data features which describe the trip like the time of the trip, origin,

destination, the weather, and the week's day. The prediction of duration of the trip may vary from one taxi to another, due to differences in the taxi usage, the brand of the vehicle, the driving habits, and the route.

In addition, taking advantage of the massive amount of gathered data, the possibility of having a more specific model for an individual taxi instead of a global model for all taxis in the traditional data mining approaches is a reality in the current hyperconnected transportation system.

However, this is not true for all taxis due to the lack of existence of enough data for creating an accurate model for all taxis. For example, in a sparse area of the city, like the area indicated in Figure 1 with a circle, only a few observations is available for modeling. In this case, a collaborative data modeling would be a good solution to improve the quality of the model by using the data from other sources. The previous results [19, 20] confirm that there is potential in sharing data in the process of learning local models. For the prediction of the trip duration, different taxis can use different data like taxi's data, data from taxi's neighbors, data collected at the roadside unit, or the data which is collected throughout the city (different dimensions). Further, collaborative data modeling may improve the quality of the prediction in the cost of data communication.

On the other hand, various algorithms have been introduced which can be used for the prediction of trip duration. However, their performance varies and therefore causes several challenges. One of the important challenges in the area of data mining is that to discover for a specific problem which algorithm has the best performance. Yet, it has already been shown that there is no unique algorithm in which it has the best performance for a broad problem domain [21]. Therefore, the problem of algorithm selection is done either based on expert advice or a trial-and-error approach. Neither way is utterly satisfactory for the end user who wishes to access the technology cost-effectively [22].

A metalearning approach [23] is used in this paper to deal with these problems: 1) which algorithm should be used, and 2) when and with whom the data should be shared. Metalearning uses a machine learning algorithm to relating the algorithm's performance to the data characterization. The contribution is also extended to 3) applying this collaborative data analysis to the problem of estimating of the taxi trip duration using the taxi dataset which is obtained from the Carnegie Mellon (CMU) Portugal project, DRIVE-IN (Distributed Routing and Infotainment through Vehicular Inter-Networking) [24]. We also 4) discuss the use of this approach for other applications including manufacturing of products and supply chain management.

The rest of the paper is organized as follow. Section 2 explains the proposed approach and the results of the experiment. An extension of the approach is discussed in Section 3. Finally, Section 4 concludes the paper.

2   An Example of Collaborative Data Analysis in Hyperconnected Transportation System

In this section, a collaborative data analysis (Section 2.2) is introduced and applied to a dataset from hyperconnected transportation system (Section 2.1). Then, the proposed approach is evaluated on the dataset and the results are presented in Section 2.3.

2.1   Taxi Data

The data is collected from a large-scale scenario [24], from taxis in the city of Porto. Porto has an area of 41.3 km^2 , and contains 965 km of roads which make it the second largest city in Portugal. It has with more than one million inhabitants and is in a center of a metropolitan area. There are 441 taxis and 63 taxi stands in the city. Each taxi has a GPS receiver and can collect the log about each trip. The dataset consists of 5 months in 2013 for all the taxis. The dataset contains 13 data features characterizing events [24].

As an example of the scenario where the data is collected, Figure 2 shows a snapshot of the taxis placements in the city of Porto. The green dots show the taxi positions in the city. The communication range for two taxis (red and black) is also shown by purple circles.
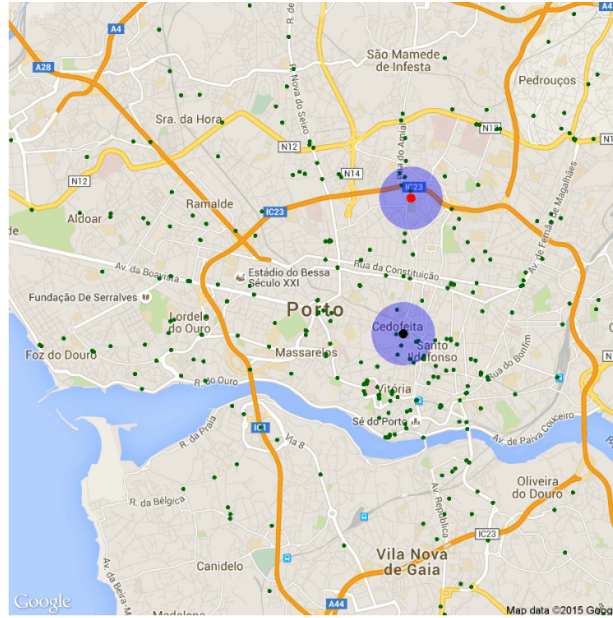

Fig. 2. Illustrative map of Porto, Portugal. The green dots are the initial taxi placement. The communication range for the red and the black taxi is shown by a purple circle around them.

There are five months data in 2013. Table 1 shows the number of taxis and the average number of observations for each taxi in each month. There are around 440 taxis and on average, there are between 1238-1484 observations for each taxi. The fourth and fifth columns also show the minimum and the maximum number of observations per each taxi, respectively. As it can be seen, there are taxis with few examples (2-14 observations). In these cases, data aggregation to use data from other taxis may be useful for modeling.

Table 1. A simple statistics about Taxi dataset

Month	No. Of Taxis	Average number of observations per each taxi	Minimum number of observations per each taxi	Maximum number of observations per each taxi
201302	443	1238.5	2	2467
201303	443	1356.6	6	3452
201304	446	1302.7	14	2824
201305	443	1484.9	2	3165
201306	442	1385.7	6	3037

2.2   Collaborative Data Analysis

As mentioned earlier, to improve the problem of algorithms selection and collaborative data modeling, we propose a metalearning approach which helps to select the right algorithm and the right part of data space for each individual taxi. In this section, we briefly describe the methodology which is summarized in Figure 3.

Fig. 3. Methodology used for a collaborative data analysis in hyperconnected transportation systems

The data collection is done within the DRIVE-IN project [24] (see Section 2.1). Then, the dataset is organized in hierarchy structure using two levels of hierarchy: the data associated with a taxi as local data and the data for the whole month as global data.

In the proposed model, at the first level, each taxi (T_i) makes a category, $C_i^1, \forall i \in \{1, \dots, n_1\}$ where n_1 is the number of taxis. The level two has only one category joining all the data from 440 taxis.

After forming the dataset in a customized format, $DB = \{T_i, C_i^j, y_i\}, \forall i \in \{1, \dots, 440\}, \forall j \in \{1, 2\}$, it is passed to the base-level experiment. In the next step, each algorithm is applied on the data from both levels for each taxi.

At the base-level, four machine learning algorithms are applied to the data: support vector machines (SVMs) [25, 26, 27], random forest [25, 26, 27], decision tree [31, 32, 33], and linear regression [31, 32]. As a result, there are several performance indicators for each taxi: P_{ig}^k which indicates the performance of the algorithm g at level k for taxi i.

$$P_{iw}^j: \forall w \in \{1, \dots, 4\}, \forall j \in \{1,2\}, \forall i \in \{1, \dots, 440\} \tag{1}$$

Where w indicates the algorithms, i shows taxis, and j stands for levels.

In addition, for each taxi and at each level of the hierarchy, the metafeatures are calculated.

There are 31 metafeatures in which they describe the dataset's structure. All the metafeatures that we used are briefly described in [36]. In general mf_i^j is the value of the metafeatures for the taxi i at the level j. By comparing the performance indicators at the base-level for each taxi, the best one is selected according to the Eq. 2:

$$P_{best_i} = \max_{w,j}(P_{iw}^j), \forall w \in \{1, \dots, 4\}, \forall j \in \{1,2\} \tag{2}$$

The metadata consists of the taxi identification, metafeatures for both levels and the best performance indicator obtained from Eq. 2.

$$T_i, mf_i^1, mf_i^2, P_{best_i} \tag{3}$$

The purpose of the metalearning is to discover the best algorithm and level to obtain the best performance. Consequently, the meta-level experiment maps the extracted metafeatures to the best performance obtained at the base-level (target variable at the meta-level).

Finally, the proposed model can recommend an algorithm and a level for each taxi that by applying the recommended algorithm on the suggested level, the best performance can be obtained (see Eq. 4).

$$Model\ input: T_i \tag{4}$$

$$Model\ output: \{recommended\ level\ (j), recommended\ algorithm\ (g)\}$$

More detail about the methodology used to obtain the results is presented in [36].

2.3   Evaluation

Base-level evaluation. The problem of prediction of the trip duration at the base-level is a regression problem. Each trained model tries to predict the trip duration. This evaluation of the prediction is done by the Normalized Root-Mean-Square-Error (NRMSE). RMSE is a routinely used metric which indicates the differences between the actual observed value and the predicted value by a machine learning model. In addition, the NRMSE is calculated using the RMSE divided by the standard deviation (std) of the predicted variable (See Equations 5 and 6). We use the hydroGOF package [37] in R [38] to calculate the NRMSE.

$$RMSE = \sqrt{\frac{\sum(\widehat{Dt_i} - Dt_i)^2}{n_1}} \tag{5}$$

$$NRMSE = 100 * \frac{RMSE}{\sigma} \tag{6}$$

Where n_1 is the number of predictions, σ is the std of the prediction values, Dt_i is the real trip duration, and $\widehat{Dt_i}$ is the predicted trip duration. Considering the NRMSE values, the algorithm with the lowest NRMSE is chosen as the best one for each taxi and is used at the meta-level.

Meta-level evaluation. The proposed framework at the meta-level predicts an algorithm and a level of granularity which have the lowest NRMSE for a selected taxi. As a result, the meta-level problem is a classification problem. This prediction is made based on metafeatures which describe the characterization of the dataset.

At the meta-level, the performance of the framework is evaluated by the accuracy of the model. Moreover, the performance of the proposed framework is also evaluated by comparing to the possible range of base-level performance. $Scaled_{error}$ is a metric that shows the relative NRMSE of the metalearning framework with respect to the best and the worst NRMSE at the base-level. It is calculated by the following equation:

$$Scaled_{error} = \frac{NRMSE_{ML} - NRMSE_B}{NRMSE_W - NRMSE_B} \qquad (7)$$

Where $NRMSE_{ML}$ is the NRMSE of the metalearning framework, $NRMSE_B$ is the best NRMSE and $NRMSE_W$ is the worst NRMSE at the base-level. Therefore, the range of $Scaled_{error}$ is from 0 to 1. Furthermore, a lower $Scaled_{error}$ shows a better performance at the meta-level.

2.4   Is Collaborative Data Analysis Useful?

In this section, we sum up the obtained results from the evaluation part of our methodology.

Base-level results. As previously mentioned (Section 2.3), the performance of the base-level is evaluated by NRMSE. Figure 4-(a) is the box-plot of the average NRMSE for each taxi for each month. It is clear that the NRMSE is less than 5% for all months. The average NRMSE is approximately 1% for each month. So, the base-level error is around 1% on average which sounds significantly good. These results show that at the base-level, the models can predict the trip duration very precisely.

Meta-level results. As discussed previously (Section 2.3), our metric to evaluate the meta-level results is $Scaled_{error}$. Figure 4-(b) shows this metric on average for all months. As it can be seen, the $Scaled_{error}$ is about 30% in the worst scenario and around 5% for the best one. The meta-level results (the blue line) are much better than the baseline results (the green line). The baseline is the majority best solution at the base-level which is a combination of an algorithm and a level for each taxi that has the best performance. Generally speaking, the results show that during the raining months (February to April) in Porto, when there are more taxi trips in the city, the meta-level is more accurate than the months that there are fewer taxi trips (May and June) due to lack of enough data for building a model. In this case, it is recommended to use the global model to have an accurate model.
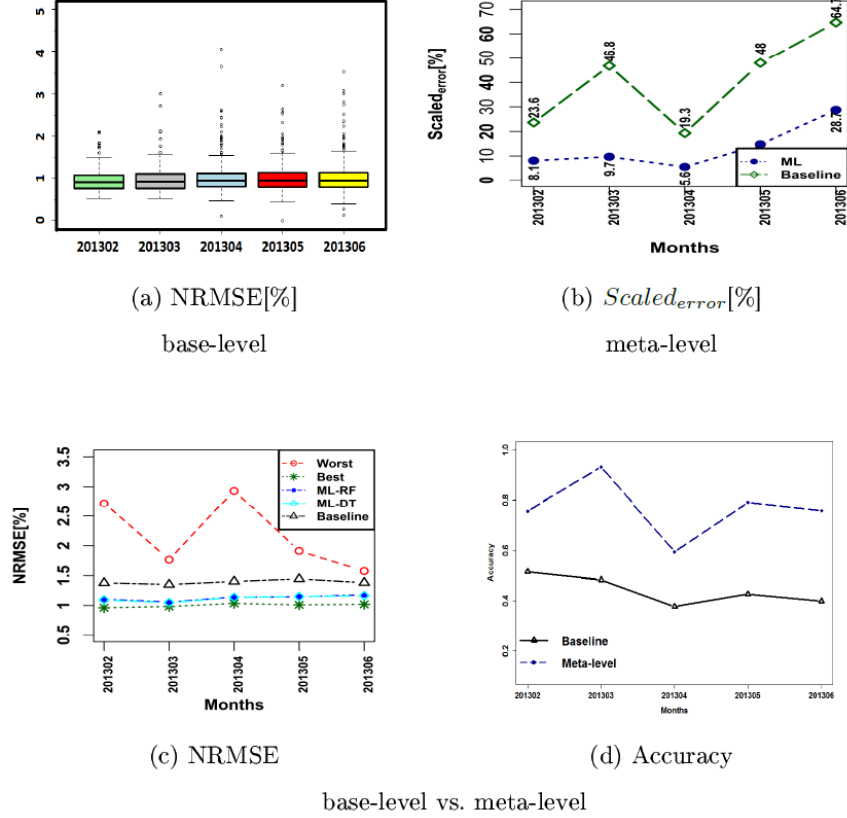
(a) NRMSE[%]

base-level

(b) $Scaled_{error}[\%]$

meta-level

(c) NRMSE

(d) Accuracy

base-level vs. meta-level

Fig. 4. Evaluation results

Meta-level vs base-level results. The comparison results between the meta-level and the base-level show that the performance of the algorithms selected by the meta-model is near the best performance obtained at the base-level (Figure 4-(c)). The performance of the suggested solution by the meta-level (the blue and the cyan line) is almost close to the best performance at the base-level (the green line) while is better than the baseline performance (the black line). This satisfies our objectives.

Finally, the comparison of accuracy between the baseline (the black line) and the meta-level (the blue line) is illustrated in Figure 4-(d). According to these results, the performance of the meta-level exceeds the performance of the baseline for all months.

3   Discussion

The proposed approach in this paper helps to identify the best solution for the problem of prediction of taxi trip duration to improve the taxi utilization and passenger satisfaction. Using the hyperconnected intelligent systems, this proposal can be used by different applications including manufacturing of products by forecasting of product demands, supply chain management by predicting of the best location of stocks, and real-time optimization of supply chain networks by prediction of the best route for fleets, through networking machinery, sensors and control systems together.

All manufacturers have a desire to give their consumers exactly what they want. This can be done by using past data collected from sales and production lines and forecasting the demand for a product. Our proposed approach can be implemented in this case by organizing the past data into hierarchy structure. One of possible solution can be using three different hierarchy levels: the data associated with a product, the data related to a group of products which have the same type, i.e. foods, clothes, and so on, and all the data.

Another example is supply chain management where a poor stock's location can give low productivity, unreliable deliveries of materials, high costs, and poor customer service. The same approach can be done using the data that can be collected from the logistics and also stocks movement within the supply chain. The approach can help to deal with the uncertain and non-stationary demand with minimum cost.

The hyperconnected networks and IoT present excellent possibilities for businesses to conduct through evolution and innovation. For this purpose, companies need to be aware of the changing of the business dynamics caused by innovation.

4   Conclusion

The existence of hyperconnected networks can make a revolution in the public transportation systems. High resource utilization and customer satisfaction are two major metrics which can be achieved using hyperconnected networks and machine learning approaches. In this paper, a metalearning proposal is introduced for the prediction of taxi trip duration to improve the taxi utilization and passenger's satisfaction by using a collaborative data analysis. The proposed approach aims at assisting taxi companies, passengers, and the public authorities for better distributing taxis to the taxi stands and determining the best route for the taxi trips using appropriate learning algorithm for the prediction of taxi trip duration. The experiments are carried out on the data collected in the Drive-In project. Several models are obtained from two levels of granularity: taxi itself and whole taxis. The results demonstrate that the proposed metalearning framework can help recommending an algorithm with the best performance at the base-level with high accuracy and outperforms the baseline accuracy. Furthermore, the performance of the base-level is also substantially adequate. In overall, the metalearning forecasts the duration of the taxi trip with the error rate less than 5%.

As interesting future works, this approach can be extended to other parts of the hyper-connected networks, i.e. manufacturing of products and supply chain management, as discussed in Section 3.

Acknowledgment

References

[1]   T. E. I. U. Limited, "The Hyperconnected Economy: Phase 2, Hyperconnected Organizations," London, 2015.

[2]   G. S. M. A. for the SBD, "2025 Every Car Connected: Forecasting the Growth and Opportunity," London, 2012.

[3]   T. W. E. Forum, "Connected World Transforming Travel, Transportation and Supply Chains," Geneva, 2013.

[4]   TfL, "Performance Report Quarter 1 2012/13," London, 2013.

[5]   Siemens, "Pictures of the Future. Livable Megacities – Moscow and St. Petersburg," Russ, 2007.

[6]   C. Englund, L. Chen, A. Vinel and S. Lin, "Future Applications of VANETs," in *Vehicular ad hoc Networks*, C. Campolo, A. Molinaro and R. Scopigno, Eds., Springer International Publishing, 2015, pp. 525-544.

[7]   J.-C. Thill, "Geographic information systems for transportation in perspective," *Transportation Research Part C: Emerging Technologies,* vol. 8, no. 1, pp. 3-12, 2000.

[8]   T. A. Hauser and W. T. Scherer, "Data mining tools for real-time traffic signal decision support \& maintenance," in *Systems, Man, and Cybernetics, 2001 IEEE International Conference on*, 2001.

[9]   C.-Y. Chan and D. Marco, "Traffic monitoring at signal-controlled intersections and data mining for safety applications," in *Intelligent Transportation Systems, 2004. Proceedings. The 7th International IEEE Conference on*, 2004.

[10]   F. D. Salim, S. W. Loke, A. Rakotonirainy, B. Srinivasan and S. Krishnaswamy, "Collision pattern modeling and real-time collision detection at road intersections," in *Intelligent Transportation Systems Conference, 2007. ITSC 2007. IEEE*, 2007.

[11]   F.-Y. Wang, "Parallel control and management for intelligent transportation systems: Concepts, architectures, and applications," *Intelligent Transportation Systems, IEEE Transactions on,* vol. 11, no. 3, pp. 630-638, 2010.

[12]   K. N. Qureshi and A. H. Abdullah, "A survey on intelligent transportation systems," *Middle-East Journal of Scientific Research,* vol. 15, no. 5, pp. 629-642, 2013.

[13]   W. He, T. Lu and C. Q. Yu, "A Novel Traffic Flow Forecasting Method Based on the Artificial Neural Networks and Intelligent Transportation Systems Data Mining," *Advanced Materials Research,* vol. 842, pp. 708-711, 2014.

[14]   Y. He, S. Blandin, L. Wynter and B. Trager, "Analysis and Real-Time Prediction of Local Incident Impact on Transportation Networks," in *Data Mining Workshop (ICDMW), 2014 IEEE International Conference on*, 2014.

[15]   X. Zhang and J. A. Rice, "Short-term travel time prediction," *Transportation Research Part C: Emerging Technologies,* vol. 11, no. 3–4, pp. 187-210, 2003.

[16]   T. Rashed and C. Jurgens, Remote Sensing of Urban and Suburban Areas, vol. 10, T. Rashed and C. Jurgens, Eds., Springer, 2010, pp. 181-192.

[17]   U. Lee and M. Gerla, "A survey of urban vehicular sensing platforms," *Computer Networks,* vol. 54, no. 4, pp. 527-544, March 2010.

[18]   M. Boban, J. Barros and O. Tonguz, "Geometry-Based Vehicle-to-Vehicle Channel Modeling for Large-Scale Simulation," *IEEE Transactions on Vehicular Technology,* vol. 63, no. 9, pp. 4146-4164, Nov 2014.

[19]   M. Nozari Zarmehri and C. Soares, "Improving Data Mining Results by taking Advantage of the Data Warehouse Dimensions: A Case Study in Outlier Detection," in *Encontro Nacional de Inteligencia Artificial e Computacional*, Sao Carlos, Brazil, 2014.

[20]   M. Nozari Zarmehri and C. Soares, "Using Data Hierarchies to Support the Development of Personalized Data Mining Models: a Case Study in Error Detection in Foreign Trade Transactions," *Submitted to International Journal of Data Warehousing and Mining,* 2016.

[21]   D. H. Wolpert and W. G. Macready, "No free lunch theorems for optimization," *IEEE Transactions on Evolutionary Computation,* vol. 1, no. 1, pp. 67-82, April 1997.

[22]   C. Giraud-Carrier, R. Vilalta and P. Brazdil, "Introduction to the Special Issue on Meta-Learning," *Machine Learning,* vol. 54, no. 3, pp. 187-193, 2004.

[23]   P. Brazdil, C. Giraud-carrier, C. Soares and R. Vilalta, Metalearning: Applications to Data Mining, C. Sammut and G. I. Webb, Eds., Springer, 2009, pp. 662-666.

[24]   Cmuportugal.org, *DRIVE-IN: Distributed Routing and Infotainment through Vehicular Inter-Networking,* 2014.

[25]   M. A. Hearst, S. T. Dumais, E. Osman, J. Platt and B. Scholkopf, "Support vector machines," *Intelligent Systems and their Applications, IEEE,* vol. 13, no. 4, pp. 18-28, 1998.

[26]   B. Scholkopf and A. Smola, "Support Vector Machines," *Encyclopedia of Biostatistics,* 1998.

[27]   I. Steinwart and A. Christmann, Support vector machines, Springer Science and Business Media, 2008.

[28]   Y. Amit and D. Geman, "Shape Quantization and Recognition with Randomized Trees," *Neural Computation, Massachusetts Institute of Technology,* vol. 9, no. 7, pp. 1545-1588, October 1997.

[29]   L. Breiman, "Random Forests," *Machine Learning,* vol. 45, no. 1, pp. 5-32.

[30]   A. Liaw and M. Wiener, "Classification and Regression by randomForest," *R News,* vol. 2, no. 3, pp. 18-22, 2002.

[31]   L. B. J. F. R. Olshen, C. J. Stone and others, "Classification and regression trees," *Wadsworth International Group,* vol. 93, no. 99, p. 101, 1984.

[32]   S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE Transactions on Systems, Man, and Cybernetics,* vol. 21, no. 3, pp. 660-674, 1991.

[33]   B. Ripley, "tree: Classification and regression trees," 2014.

[34]   G. A. F. Seber and A. J. Lee, Linear regression analysis, vol. 936, John Wiley and Sons, 2012.

[35]   D. C. Montgomery, E. A. Peck and G. G. Vining, Introduction to linear regression analysis, vol. 821, John Wiley and Sons, 2012.

[36]   M. Nozari Zarmehri and C. Soares, "Advances in Intelligent Data Analysis XIV: 14th International Symposium, IDA 2015, Saint Etienne. France, October 22 -24, 2015. Proceedings," in *Advances in Intelligent Data Analysis XIV*, E. Fromont, T. De Bie and M. van Leeuwen, Eds., Ch, a: Springer International Publishing, 2015, pp. 205-216.

[37]   M. Zambrano-Bigiarini, "hydroGOF: Goodness-of-fit functions for comparison of simulated and observed hydrological time series," 2014.

[38]   R. C. Team, "R: A Language and Environment for Statistical Computing," Vienna, 2015.