Multiobjective Selection of Input Sensors for SVR Applied to Road Traffic Prediction

Jiri Petrlik, Otto Fucik, and Lukas Sekanina

Brno University of Technology, Faculty of Information Technology, IT4I Centre, Czech Republic
{ipetrlik,fucik,sekanina}@fit.vutbr.cz

Abstract. Modern traffic sensors can measure various road traffic variables such as the traffic flow and average speed. However, some measurements can lead to incorrect data which cannot further be used in subsequent processing tasks such as traffic prediction or intelligent control. In this paper, we propose a method selecting a subset of input sensors for a support vector regression (SVR) model which is used for traffic prediction. The method is based on a multimodal and multiobjective NSGA-II algorithm. The multiobjective approach allowed us to find a good trade-off between the prediction error and the number of sensors in real-world situations when many traffic data measurements are unavailable.

Keywords: Road traffic forecasting, multiobjective feature selection, multiobjective genetic algorithms.

1 Introduction

Modern traffic sensors (induction loop detectors, radars and camera detectors) can measure various road traffic variables such as the *traffic flow* representing the number of vehicles passing a given road segment per time interval, the *occupancy* which is a dimensionless traffic variable describing the fraction of a time interval in which the current place is occupied by a vehicle, and the *arithmetic mean speed* of vehicles passing the current place. The detectors usually aggregate these data from intervals between 20 s and 5 min [1]. The information provided by traffic sensors is used in modern intelligent traffic systems (ITS), for traffic system planning and other purposes.

As the traffic sensors are not one-hundred percent reliable, the problem of estimation of missing values was identified. Various solutions to this problem have been proposed by means of modern soft computing methods which, in addition to the estimation, can also be employed to predict the future values on desired sensors. The predicted values can be utilized in ITS to control, for example, traffic lights and variable message signs. One of the most promising machine learning methods of short-term traffic flow forecasting is *support vector regression* (SVR). Previous methods based on SVR have not considered the selection of proper inputs (sensors). However, a proper selection of these sensors can significantly influence the quality of prediction.

T. Bartz-Beielstein et al. (Eds.): PPSN XIII 2014, LNCS 8672, pp. 802–811, 2014.

In this paper, we propose a new multiobjective optimization method based on a genetic algorithm for selection of a subset of inputs for SVR. The proposed solution can be used for short-term traffic forecasting or for estimation of unmeasured values from broken sensors. Dealing with the missing values is important, because if the value from an input sensor is unavailable, the SVR method does not work at all! The proposed method is constructed as multiobjective because there is a natural trade-of between the error of prediction, the number of input sensors for SVR model and the data unavailability rate (a time fraction in which the SVR model can not be used because of missing data). The proposed method is evaluated using publicly available data and compared with a single objective optimization scenario.

The rest of the paper is organized as follows. Section 2 introduces the short-term traffic forecasting problem and methods based on SVR which can be used to solve this problem. Section 3 deals with multiobjective evolutionary algorithms. In Section 4, the proposed method is described. Experimental evaluation is performed in Section 5. Section 6 concludes the paper.

2 Road Traffic Forecasting Using SVR

Artificial neural networks and SVR were applied to solve the road traffic forecasting problem [2,4,5,6]. However, SVR becomes a more popular method in this task. SVR is a variant of support vector machine (SVM). While SVMs are usually used for classification problems, SVR is designed for regression and prediction problems [3]. The original SVM algorithm could work only as a linear classifier. In order to deal with non-linear problems, SVM/SVR was extended to support nonlinear kernel functions such as polynomial kernels, Gaussian radial basis kernels, and hyperbolic tangent kernels.

A SVR modification called Online-SVR (OL-SVR) [4] was previously used for short-term prediction of traffic behavior. The data from seven randomly selected highways were used to evaluate this method under typical and atypical traffic conditions. The atypical traffic conditions can appear, for example, during holidays or traffic incidents. The method was compared with neural networks (multilayer perceptron), Gaussian maximum likelihood (GML) and Holt's exponential smoothing. For typical traffic conditions the OL-SVR model outperformed the multilayer preceptron and Holt's exponential smoothing, but GML model provided better results. For atypical traffic conditions the OL-SVR model outperformed all other methods [4].

SVR requires to correctly set various meta-parameters such as the kernel type and regularization parameter. For example, chaotic simulated annealing was successfully used for parameter tuning. The results showed that SVR with optimized meta-parameters is as good as other techniques like seasonal autoregressive integrated moving average (SARIMA), seasonal Holt-Winter's model and back-propagation neural networks [5]. In another approach, a modified version of a particle swarm optimization (PSO) was utilized to find the optimal settings of SVR meta-parameters. The results proved that SVR model with

meta-parameters set by PSO can outperform the back propagation neural networks and ARMA model [6].

However, all the methods assumed that all sensors work nearly all the time, which is an unrealistic assumption. Hence the main objective of this paper is finding a solution which will work with unreliable sensors, i.e. missing data.

3 Multiobjective Genetic Algorithms

The most important objective in the prediction tasks is minimizing the error of prediction which is usually calculated by some error metrics such as the root mean squared error (RMSE). However, in real-world scenarios, other objectives have to be considered, for example, the number of data streams (sensors) has to be minimized because of their cost, maintainability, and reliability. In the context of this paper, the goal of the multiobjective scenario is to find the smallest subset of input sensors for which the number of missing values is minimal and the RMSE is minimal. In general, the multiobjective optimization problem can be defined in the following form:

$$\begin{aligned} &\text{minimize: } f_m(\boldsymbol{x}), && m = 1, 2, ..., M \\ &\text{subject to: } g_j(\boldsymbol{x}) \geq 0 && j = 1, 2, ..., J \\ &\qquad\qquad h_k(\boldsymbol{x}) = 0 && k = 1, 2, ..., K \end{aligned} \tag{1}$$

where $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ is a vector representing the solution consisting of n decision variables. The objective functions are denoted $f_1, \ldots, f_M$. These functions are to be minimized. Functions $g_j(\boldsymbol{x})$ and $h_k(\boldsymbol{x})$ define the optimization constrains. In order to compare two solutions, Pareto-dominance relations were established [7]: Solution $\boldsymbol{x^{(1)}}$ dominates another solution $\boldsymbol{x^{(2)}}$ if the following conditions are satisfied: (1) The solution $\boldsymbol{x^{(1)}}$ is no worse than $\boldsymbol{x^{(2)}}$ in all objectives. (2) The solution $\boldsymbol{x^{(1)}}$ is strictly better than $\boldsymbol{x^{(2)}}$ in at least one objective.

In the set of solutions P, the non-dominated subset of solutions P' contains those solutions that are not dominated by any member of P. The non-dominated subset of all possible solutions is called Pareto-optimal set. The goal of multiobjective optimization is to find solutions of the Pareto-optimal set.

In the past, many variants of multiobjective genetic algorithms were proposed, for example, Vector Evaluated Genetic Algorithm (VEGA), Strength Pareto Evolutionary Algorithm (SPEA), and non-dominated sorting genetic algorithm (NSGA-II) [8]. A modification of NSGA-II algorithm called the multimodal NSGA-II was previously successfully used to solve the feature selection problem – identifying a minimal subset of genes for cancer classification [9]. Multimodal algorithms are utilized in the case that many different but equally good solutions exist and it is important to find many of them.

4 Method

The proposed method can be used to either predict the traffic flow or estimate missing values for a broken sensor. In the first phase, the SVR model is trained

using historical data (train set) in the supervised learning scenario [10]. Trained SVR model then describes mathematical dependencies among the values of the sensor for which predictions are desired and other sensors in the area. Other historical data, unseen during the learning phase (test set), are used to validate the resulting model. The multiobjective multimodal NSGA-II algorithm is employed to find the proper subset of input sensors for the SVR model.

Traffic data are usually available as a set of time series $s_1, \ldots, s_n$; one time series for each variable measured by a traffic sensor. In order to train the SVR model, it is necessary to convert these data into training samples (Fig. 1). By means of a sliding window, the current value ($s_i^{(0)}$) and a few (h) previous values ($s_i^{(-1)}, \ldots, s_i^{(-h)}$) from each series are taken into a training sample. In the case of estimating the current value of a broken sensor (Fig. 1, left), the current value $f^{(0)}$ is included into the training sample as a dependent variable. In the case of traffic forecasting in the place of sensor, the future value $f^{(+l)}$ is included into the training sample (Fig. 1, right), where l represents the prediction horizon.

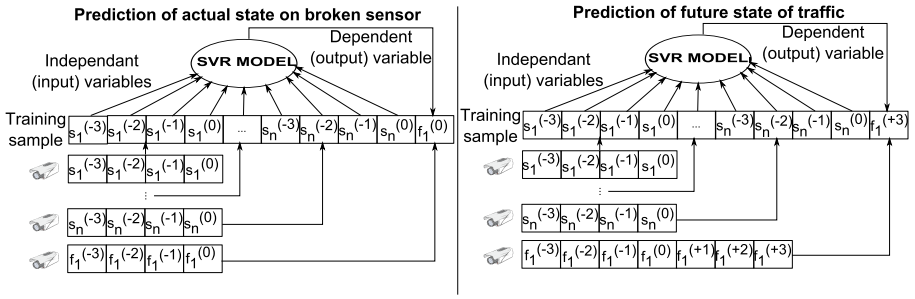

Fig. 1. Composition of training samples for SVR: prediction of a current value (left) and prediction of a future value (right) of a sensor producing f

We employed the multiobjective multimodal NSGA-II operating over binary strings. Each gene represents one input sensor, where 1 denotes including and 0 excluding of a particular sensor from the input vector fed to SVR (Fig. 2).

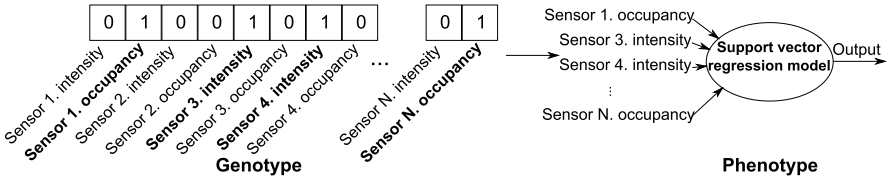

Fig. 2. Chromosome encoding and a corresponding phenotype (SVR model)

Three objectives are considered (all to be minimized) – the number of sensors used as inputs for SVR, the rate of missing samples for prediction and the

prediction error. The rate of missing samples is portion of time for which the concrete model can't be used because of missing data. All objectives are evaluated using the test set. Two well-known metrics can be used: root mean squared error (RMSE) and relative squared error (RSE) defined as

$$RMSE = \sqrt{\frac{\sum_{i=1}^{d}(y_i - y_i')^2}{d}} \qquad RSE = \frac{\sum_{i=1}^{d}(y_i - y_i')^2}{\sum_{i=1}^{d}(y_i - \overline{y})^2}, \tag{2}$$

where d represents the number of regression samples, y^i is the desired value for ith regression sample and y_i' is predicted by current model. The value $\overline{y}$ denotes the mean value predicted by a naive regression model. In the further evaluation of our method we will use RMSE as the error metric.

5 Experimental Results

5.1 Data Sets

We used publicly available data from traffic sensors in Seattle [11]. Sensors are placed on 23 intersections in the city and measure the traffic flow, occupancy and average speed. The rough data are aggregated in data tables into 1 minute intervals for a period starting on May 1^{st} and ending on October 31^{st} 2011. Among other information in the data tables, each row contains the traffic flow, occupancy and average speed for one sensor, and a flag indicating correctness of the measured data. All our experiments are performed using the data coming from one subarea of Seattle. Incorrect records were removed from the data tables and the remaining data were aggregated into 5 minute intervals.

5.2 SVR Parameters Setting

Although the optimization of SVR metaparameters is not the primary objective of this work, we tried to identify the most suitable setting of basic parameters of SVR which employs radial basis kernels (RBF). Figure 3 shows RSE for various settings of the regularization parameter ($C = \{2^{-5}, 2^{-4}, \ldots, 2^{14}, 2^{15}\}$) and kernel parameter ($\gamma = \{2^{-15}, 2^{-14}, \ldots, 2^{2}, 2^{3}\}$). In the following experiments we will utilize $C = 2^3$ and $\gamma = 2^{-12}$ because a clear minimum of RSE can be seen for them in Fig. 3.

5.3 Evaluation of NSGA-II

The proposed method was evaluated on places 6, 11, 19, 22, and 23 of the area [11]. For each sensor located on these places, four SVR models were created. The first two SVR models are trained to perform a short-term prediction in the horizon of 15 minutes. One of them uses only the actual values measured on the neighbor detectors in the area and the second one uses the actual values and the values measured on these sensors in previous 15 minutes. The other two

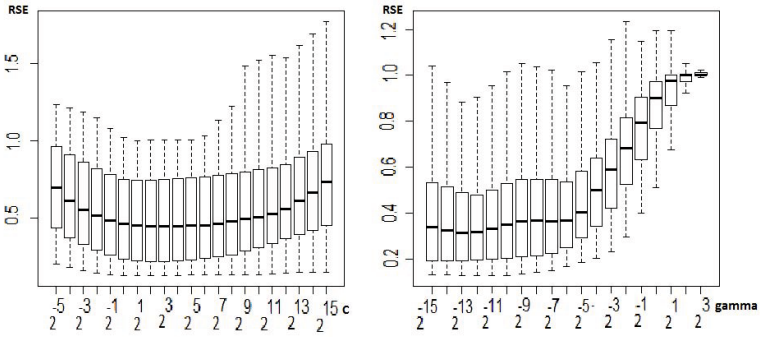

Fig. 3. The effect of setting of the regularization parameter C and kernel parameter γ on the quality of SVR prediction

SVR models are trained to estimate the actual value on the sensor in the case of a sensor error. And again, one of them uses only the actual values measured on the neighbor detectors in the area and the second one uses the actual values and the values measured on these sensors in previous 15 minutes.

The parameters of the NSGA-II genetic algorithm are as follows. The probability of uniform crossover is 70% and the probability of mutation is 5%. Each NSGA-II run, which operates with a 40 member population and 100 generations (4000 fitness evaluation), is repeated 20 times. The prediction error is given as the RMSE. The evolution utilizes approximately 50% of the available data to train SVR model, the remaining data are used to validate the evolved SVR models in the following figures and tables. Experiments were performed on an Anselm supercomputer whose nodes are equipped with two Intel Sandy Bridge E5-2665 chips. These chips contain 8-core processors working at 2.4 GHz. One run takes approximately 10 hours of one processor core. Our software was implemented in the scripting language of the system R for statistical computing [12]. We used publicly available R package e1071 for training of SVR models.

Figure 4 shows the resulting Pareto fronts from a typical NSGA-II run. Numerous non-dominated compromises between RMSE and the number of input sensors (left) and RMSE and the ratio of missing samples (right) are shown. The results were obtained for the future traffic forecasting scenario with the prediction horizon of 15 minutes for sensor number 3 measuring the traffic flow on place 19. The predicted values and correct values for one example solution are shown in Fig. 5.

Another experiment shows that the proposed method, in contrast with a common approach reported in the literature, can provide reasonable results even if many samples are unavailable. The best results obtained from 20 independent runs of NSGA-II are presented as box plots in Fig. 6. Resulting RMSE values are shown for the traffic flow and occupancy ($l = 3$, $h = 3$) when less than 10%, 30%, 50%, and 70% samples are unavailable. The results are given for sensor 3 on place 11. It is important to note that, for example, a value of 70% means that

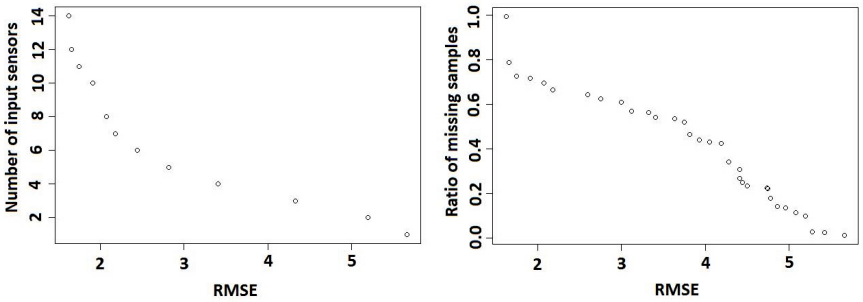

Fig. 4. Non-dominated compromises obtained for the future traffic forecasting scenario with the prediction horizon of 15 minutes for sensor number 3 measuring the traffic flow on place 19

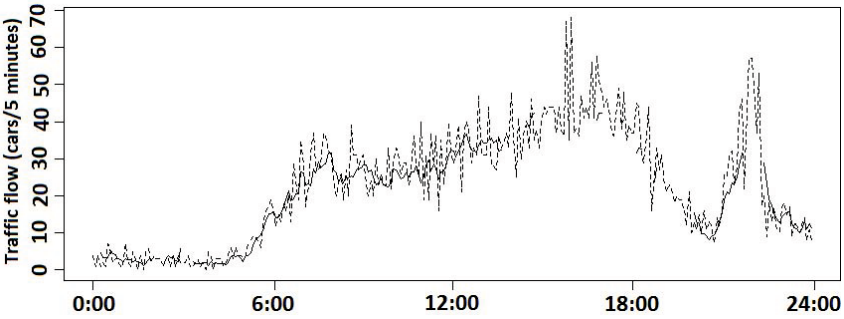

Fig. 5. Predicted values (dashed line) and correct values (normal line) of traffic flow for sensor 3 on place 19 on July 1^{st}, 2011

for a given SVR model the samples from the test set incorect in $24 \cdot 0.7 = 16.8$ hours of a day, i.e. the SVR model will not work for most of the time.

In order to provide results for some other sensors, Fig. 7 summarizes the best RMSE values obtained for places 6, 11, 19, 22, 23. For each place, 3 sensors exhibiting the biggest mean traffic flow and occupancy were chosen. It can be seen that RMSE increases when more samples are available in the test set. The short term traffic prediction scenario with horizon of 15 minutes and 15 minute history is considered in the figure.

And finally, Fig. 8 summarizes the mean RMSE over all sensors on all prediction places in all considered scenarios. The columns are:

- *actual* – the prediction of the actual values on a broken sensor using the actual values on sensors from other places ($l = 0$, $h = 0$)
- *actual (15 min.)* – see actual, but in addition, some historical data are used ($l = 0$, $h = 3$)

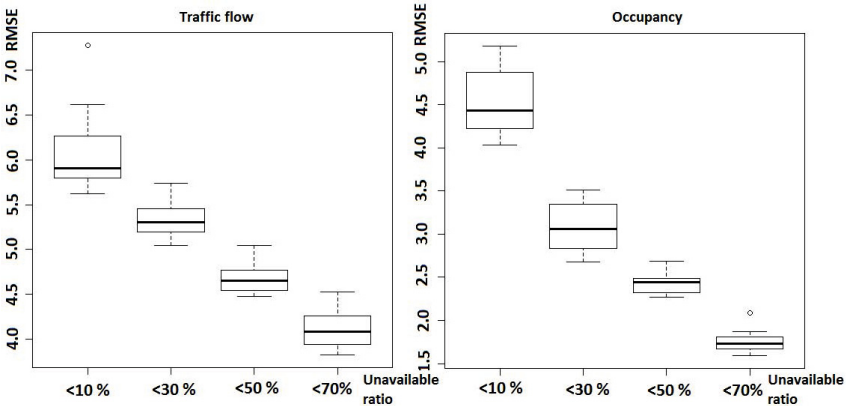

Fig. 6. Prediction error (RMSE) when less than 10%, 30%, 50%, and 70% samples are unavailable from sensor 3 on place 11

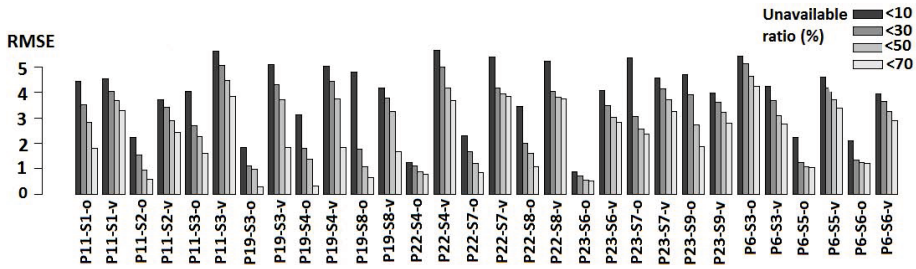

Fig. 7. Prediction error (RMSE) at 5 places (each with 3 sensors) for different amount of unavailable samples (v – traffic flow, o – occupancy)

- *15 min. future* – the traffic prediction in near future with a 15 minute prediction horizon ($l = 3$, $h = 0$). As the input for SVR model the actual values on other sensors were used.
- *15 min. future, 15 min. history* – see the previous one, but the historical data are used ($l = 3$, $h = 3$).

5.4 Comparison with a Single Objective GA

In order to justify the multiobjective approach, we consider a single criterion optimization scenario, in which RMSE is used as the only fitness function. The single-objective GA works with 40 individuals in the population, the probability of crossover is 70%, the probability of mutation is 5%, and 2-individual tournament selection (with elitism) is chosen. Table 1 compares NSGA-II with the single objective GA for several places and sensors (the best values from 20 independent runs are reported). It can be seen that the single objective GA tends to provide solutions with very small RMSE values; however, it opportunistically

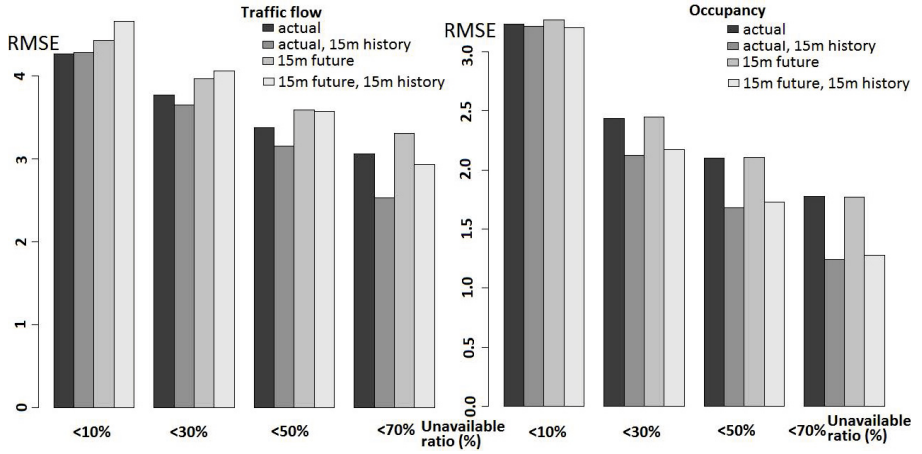

Fig. 8. Mean RMSE over all sensors on all prediction places in all considered scenarios

Table 1. The best RMSE on selected sensors and places for NSGA-II (less than 10%, 30%, 50% and 70% samples unavailable) and a single objective GA

Location			Multiobjective approach RMSE for Unavailable ratio:				Best single objective GA result	
Place	Sensor	Variable	< 10%	< 30%	< 50%	< 70%	RMSE	Unavailable ratio
Current values on sensor.								
11	3	traffic flow	5.27	4.63	4.16	4.01	2.66	96.9
11	3	occupancy	3.81	3.5	3.31	3.31	0.31	99.4
22	4	traffic flow	5.33	4.86	4.31	4.2	1.48	99.4
Prediction horizon 15 min.								
11	3	traffic flow	5.5	4.9	4.37	4.23	2.96	97.2
11	3	occupancy	4.02	3.57	3.41	3.35	0.33	99.4
22	4	traffic flow	5.51	4.89	4.56	4.35	1.84	99
Current values on sensor, 15 min. history.								
11	3	traffic flow	5.2	4.58	3.91	3.34	1.15	99.4
11	3	occupancy	4.04	2.72	2.18	1.5	0.19	99.4
22	4	traffic flow	5.62	4.71	4.09	3.37	1.04	99.4
Prediction horizon 15 min., 15 min. history								
11	3	traffic flow	5.62	5.05	4.48	3.82	1.17	99.4
11	3	occupancy	4.03	2.68	2.27	1.59	0.24	99.4
22	4	traffic flow	5.64	4.98	4.17	3.66	1.15	99.4

exploits the test data containing over 85% missing values (in many cases, over 99%, see the Unavailable ratio column). Such a SVR model will thus be useless in practice, because it will not provide any prediction most of the time. Therefore, the single optimization scenario fails in this task.

6 Conclusions

In this paper, we proposed a new method for multiobjective selection of input sensors for prediction of the traffic flow. The method is based on SVR and multimodal and multiobjective NSGA-II algorithm. Contrasted to a single objective optimization scenario, in which only the prediction error has to be minimized, the multiobjective approach allowed us to find a good trade-of between the prediction error and the number of sensors in real-world situations when many traffic data measurements are not available. One can observe that adding the historical data reduces the prediction error of the occupancy prediction.

Acknowledgments. This work was supported by the IT4Innovations Centre of Excellence CZ.1.05/1.1.00/02.0070, Brno University of Technology under number FIT-S-14-2297, and Technology Agency of the Czech Republic (TACR) project TA02030915.

References

1. Treiber, M., Kesting, A., Thiemann, C.: Traffic Flow Dynamics: Data, Models and Simulation. Springer (2012)
2. Dia, H.: An object-oriented neural network approach to short-term traffic forecasting. European Journal of Operational Research 131(2), 253–261 (2001)
3. Basak, D., Pal, S., Patranabis, D.C.: Support vector regression. Neural Information Processing-Letters and Reviews 11(10), 203–224 (2007)
4. Castro-Neto, M., Jeong, Y.-S., Jeong, M.-K., Han, L.D.: Online-SVR for short-term traffic flow prediction under typical and atypical traffic conditions. Expert Systems with Applications 36(3, pt. 2), 6164–6173 (2009)
5. Hong, W.-C.: Traffic flow forecasting by seasonal svr with chaotic simulated annealing algorithm. Neurocomputing 74(12-13), 2096–2107 (2011)
6. Li, M.-W., Hong, W.-C., Kang, H.-G.: Urban traffic flow forecasting using gausssvr with cat mapping, cloud model and pso hybrid algorithm. Neurocomputing 99(1), 230–240 (2013)
7. Deb, K.: Multi-Objective Optimization using Evolutionary Algorithms. Wiley Interscience Series in Systems and Optimization. Wiley (2001)
8. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Transactions on Evolutionary Computation 6(2), 182–197 (2002)
9. Deb, K., Reddy, A.R.: Reliable classification of two-class cancer data using evolutionary algorithms. Biosystems 72(1-2), 111–129 (2003)
10. Marsland, S.: Machine Learning: An Algorithmic Perspective. CRC (2009)
11. University of Washington Transportation Research Center, Research Data Exchange Website, Seattle Data Environment, datasets: Arterial Travel Times, `www.its-rde.net` (retrieved May 2013)
12. R Core Team: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria (2013)