Performance comparison of machine learning methods in the bus arrival time prediction problem

A A Agafonov[1] **and A S Yumaganov**[1]

[1]Samara National Research University, Moskovskoye shosse, 34, Samara, Russia, 443086

e-mail: ant.agafonov@gmail.com, yumagan@gmail.com

Abstract. The problem of predicting the movement of public transport is one of the most popular problems in the field of transport planning due to its practical significance. Various parametric and non-parametric models are used to solve this problem. In this paper, heterogeneous information affecting the prediction value is used to predict the arrival time of public transport, and a comparison of the main machine learning algorithms for the public transport arrival time forecasting is given: neural networks, support vector regression. An experimental analysis of the algorithms was carried out on real traffic information about bus routes in Samara, Russia.

1. Introduction

Public passenger transport is an important part of the transport system. Efficient use of passenger transport will help to reduce road congestion by reducing the use of personal vehicles, as well as cut down fuel consumption and reduce environmental pollution. To improve the quality of passenger transport service, among other things, it is necessary to provide passengers with information about the exact arrival time of vehicles at stops. This information is important for passengers because it allows them to choose alternative routes and reduce the waiting time for vehicles.

The arrival time of vehicles at stops can be considered as stochastic, since it depends on many factors, including the passing time of road segments, the time spent at stops and the delay time at intersections. Furthermore, such factors as traffic congestion, incidents and weather conditions must be taken into account to predict the arrival time. Thus, the development of prediction model that takes into account various spatial-temporal factors is a difficult task.

Despite the popularity of the above mentioned problem, many papers consider only individual factors (for example, speed of the vehicle on the current and previous road segments) to predict the arrival time at stop. Moreover, the comparison of algorithms in those papers is carried out on different sets of data that often include information about only one or a few routes.

In this paper, a comparison of different public transport arrival time prediction models including artificial neural networks, support vector regression and linear regression is made. Heterogeneous information describing the transport situation is used for prediction. Comparison of algorithms is carried out on the traffic data of bus network in Samara, Russia.

2. Related works

There are a large number of studies devoted to the problem of public transport arrival prediction. All existing works can be divided into several categories according to the type of used models and algorithms: parametric and non-parametric regression models, Kalman filters based models, artificial neural networks, the support vector machine, hybrid models.

Linear regression models [1, 2] are constructed as regression functions from a set of independent variables. The applicability of these models to transport systems is limited due to the strong correlation of the variables of the regression function. Nonparametric regression, in particular, the k-nearest-neighbor method, was used to solve the prediction problem in the papers [3, 4, 5]. However, the requirement of a large sample size imposes a restriction on the use of this method in real time. In [6], a clustering algorithm was used to determine the distribution of the travel time of the road segment.

Models based on the Kalman filter [7] allow to estimate the future values of the dependent variables based on the recursive procedure, taking into account the stochastic nature of the process and the noise of the measurements. Models of artificial neural networks (ANN) [8, 9] are the most commonly used approaches for predicting arrival time. Prediction model presented in [8] combines two models of neural networks trained using two sets of data respectively: travel times dataset and arrival time at stops dataset. Authors of [9] used the Bayesian approach to combine several neural networks to build a prediction.

The support vector regression (SVR) is a set of similar learning algorithms with a teacher used for classification and regression analysis problems [11, 12]. In [12], the travel time of the current and next road segments was used for prediction. In [11], the authors used a genetic algorithm to select SVR parameters. The authors of [13] used a prediction model that combines two SVR models.

Hybrid models are also used to reduce the forecast error [14, 15, 16]. These models combine several heterogeneous methods and algorithms. The travel time prediction problem is necessary to solve other complex problems, such as reliable path finding [17] or autonomous vehicles routing [18].

The results of a comparison of several regression models and machine learning methods are presented in [19], the best result was shown by the SVR model. Inverse results were obtained in [20], the best results were shown by the neural network model.

In most works, the best results of the public transport arrival time prediction were shown using machine learning methods: neural network models and SVR. However, the choice of a particular model depends on the used input data.

3. Basic notation and problem formulation

A transport network is considered as a directed graph, the vertices of which correspond to the stops and the edges denotes segments of the transport network between the stops.

Let's s denotes a bus stop from set S; w_{ij} denotes the segment of the transport network between the stops $i \in S$ and $j \in S$ with length $|w_{ij}|$; r denotes public transport route from set R; R_{ij} denotes the set of routes passing through segment w_{ij}; n denotes a vehicle from set N; N_r denotes a set of vehicles with route $r \in R$.

The problem of arrival time prediction for the vehicle $n \in N$ with route $r \in R$ at the stop $j \in S$ can be formulated as:

$$t_j^{arr,n} = t_i^{dep,n} + T_{ij}^{travel,n}, \tag{1}$$

where $t_j^{arr,n}$ denotes the arrival time at the stop j, $t_i^{dep,n}$ denotes the departure time from the stop i, $T_{ij}^{travel,n}$ denotes the travel time between stops i and j.

Then the problem of the arrival time prediction is reduced to the problem of travel time prediction $T_{ij}^{travel,n}$ or, equivalently, problem of vehicle's speed v_{ij}^{n} prediction.

The problem can be formulated as follows:
using the transport network graph, as well as statistical and real-time data, predict a speed $\hat{v}_{ij}^{n}(t_c, t)$ *at the time* t*, considering that the prediction is calculated at time* t_c*.*

4. Proposed model

4.1. Factors of prediction

In order to obtain a speed prediction $\hat{v}_{ij}^{n}$ of a vehicle $n \in N$ running the route $r \in R$, various factors affecting the predicted value can be taken into account. In contrast to the works known to the authors, this article proposes the use of heterogeneous information describing the transport situation. This information defined as follows:

- The speed v_{ij}^{n} of the vehicle $n \in N$ on the segment w_{ij};
- The weighted average speed $v_{ij}^{route,r}$ of vehicles running the route $r \in R$ on the segment w_{ij}:

$$v_{ij}^{route,r}(t) = \frac{\sum_{k \in N_r} \omega\left(t - t_i^{dep,k}\right) v_{ij}^{k}}{\sum_{k \in N_r} \omega\left(t - t_i^{dep,k}\right)},$$

  where $\omega(t)$ is a kernel

$$\omega(t) = \begin{cases} \exp(-\alpha t), & t \le \Delta_{max}, \\ 0, & t > \Delta_{max}; \end{cases}$$

  Δ_{max} is a time interval for which estimates of speed are considered.
- The weighted average speed v_{ij}^{all} of vehicles with any route on the segment w_{ij}:

$$v_{ij}^{all}(t) = \frac{\sum_{r \in R_{ij}} \sum_{k \in N_r} \omega\left(t - t_i^{dep,k}\right) v_{ij}^{k}}{\sum_{r \in R_{ij}} \sum_{k \in N_r} \omega\left(t - t_i^{dep,k}\right)};$$

- The average hourly traffic flow speed v^{hour};
- The average daily traffic flow speed v^{day};
- The historical average speed $v_{ij}^{stat}(t)$ of vehicles with any route on the segment w_{ij} at time interval t;
- The average traffic flow speed $v_{ij}^{flow}(t)$ on the segment w_{ij} at the time point t;
- The traffic flow speed v_{ij}^{fNow} on the segment w_{ij} at the current time.

It is assumed that the average hourly and average daily speeds reflect the current seasonal and weather situation indirectly, the average speed of the traffic flow reflects the changes in the traffic situation and the occurrence of congestion.

4.2. The basic model of an artificial neural network

In [20], the neural network model with one hidden layer containing 5 neurons was used as a prediction model. Three factors were used to predict the travel time of the vehicle $n \in N$ with the route $r \in R$ on the road segment w_{ij}:

- the weighted speed of vehicle with the same route on the road segment $v_{ij}^{route,r}(t)$;
- the weighted speed of vehicle with any route on the road segment $v_{ij}^{all}(t)$;
- the vehicle speed on the previous segment $v_{i-1,i}^{n}$.

We denote this model as $\mathrm{ANN}^{3,5,1}$.

4.3. Support vector regression model

The support vector regression (SVR) method is a special class of algorithms characterized by the use of kernels. The most common kernels are linear, polynomial, radial basis function, sigmoid. In this work a radial basis function is used in the following form:

$$k(\mathbf{x}, \mathbf{x}') = \exp(-\gamma \|\mathbf{x} - \mathbf{x}'\|^2),$$

where $\gamma > 0$ is a model parameter, $\mathbf{x}$ and $\mathbf{x}'$ are the input data of the model. The three above mentioned factors are used as an input data.

4.4. Extended model of artificial neural network

We proposed to use an extended model of the neural network to predict the speed $\hat{v}_{ij}^n(t_c, t)$ of a vehicle $n \in N$, running the route $r \in R$. The input data includes all the factors described in Section 4.1, and it can be written as a vector:

$$\mathbf{V} = \Big(v_{i-1,i}^n, v_{ij}^{n1}, v_{ij}^{n2}, v_{ij}^{route,r}(t), v_{ij}^{all}(t), v_{ij}^{stat}(t_c), v_{ij}^{stat}(t), \\ v_{ij}^{flow}(t_c), v_{ij}^{flow}(t), v^{hour}(t), v^{day}(t), v_{ij}^{fNow}\Big).$$

where $n1$ is a preceding vehicle of the route r which passed the transport segment w_{ij}, $n2$ is a preceding vehicle of any route which passed the road segment w_{ij}.

The neural network model of the following form is used for prediction: one input layer (12 neurons), one hidden layer (13 neurons) and one output layer (1 neuron). The Adam [21] method was used as the optimization method.

4.5. Experiments

Experimental studies of models were carried out on traffic data of bus routes in the transport network of Samara, Russia, for two months, from August 1, 2018 to September 30, 2018. The forecast was performed for 837 vehicles on 176 routes.

The comparison of the linear regression model LR, basic neural network model $\text{ANN}^{3,5,1}$, support vector regression model SVR and the extended neural network model ANN^{ext} was made.

In order to evaluate the prediction quality of each prediction model, two standard metrics were used: mean absolute percentage error (MAPE) and mean absolute error (MAE).

$$\text{MAPE} = \frac{1}{n} \sum_{t=1}^{n} \frac{|v_t - \hat{v}_t|}{v_t} \times 100\% \tag{2}$$

$$\text{MAE} = \frac{1}{n} \sum_{t=1}^{n} |v_t - \hat{v}_t| \tag{3}$$

where v_t is a real value and $\hat{v}_t$ is a predicted value.

Table 1 shows the comparison of prediction models for one of the routes of the analysed transport network.

Table 1. Comparison of prediction models.

	LR	$\text{ANN}^{3,5,1}$	SVR	ANN^{ext}
MAPE	29.58	29.76	34.75	27.75
MAE	1.76	1.77	2.20	1.60

In this case, the size of the input data used for training and forecasting was limited to the size of selected route's data. Data obtained on a given day were used as a test data, all the rest data were used as a train data. The table shows the average MAE and MAPE values obtained for 7 days. From the obtained results it can be seen that the average value of the prediction error for one road segment is quite high. The best result is demonstrated by the extended model of an artificial neural network.

However, more interesting are the results of predicting the arrival time of vehicles at distant stops. For experimental studies of the dependence of MAPE and MAE on the forecast horizon, the full volume of data on the vehicles movement was used. The studies were carryed out for one day and all routes, while the data obtained for the entire above-mentioned period of time except the selected day were used as archival data. The time spent on training the SVR model amounts to tens of hours for such a significant amount of input data and the results obtained above show the superiority of other models. Thus the SVR model was not used on these experimental studies.The dependence of MAPE and MAE on the forecast horizon are shown in Figure 1.

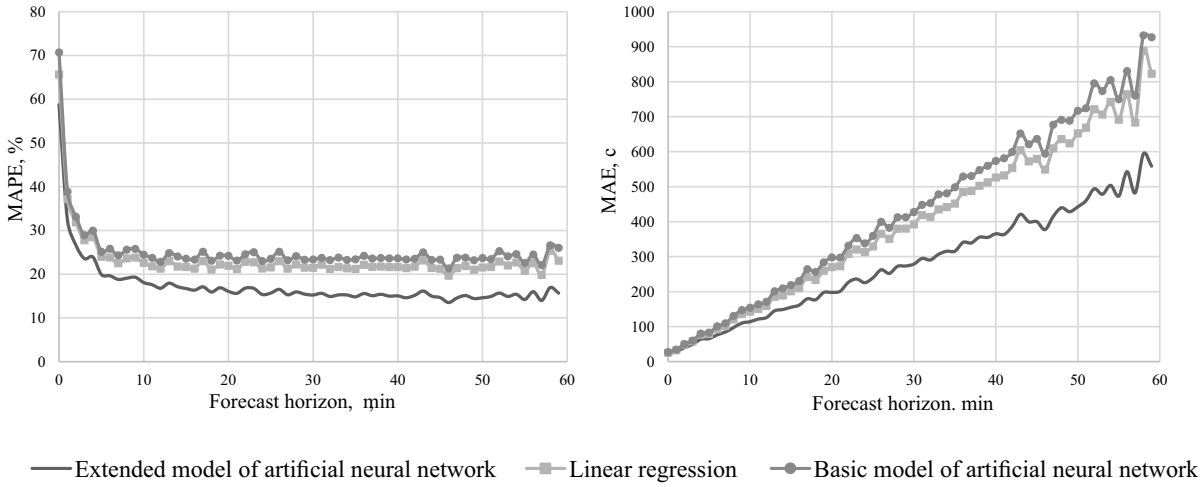

Figure 1. The dependence of MAPE and MAE on the forecast horizon.

Based on the obtained results, it can be concluded that the prediction quality of the extended model of an artificial neural network is higher throughout the forecast horizon than the prediction quality of the other models . The worst result was obtained using the basic model of the artificial neural network. At the same time, the value of MAPE decreases for all considered models with an increase in the forecast horizon value. The prediction quality of the vehicles arrival time at distant stops is significantly higher than the prediction quality for the nearest stops.

5. Conclusion

This paper proposed an extended model of the neural network which takes into account heterogeneous information to predict the arrival time of the public transport. The experiments were carried out on real traffic information about bus routes in the Samara, Russia. The proposed model showed the best results compared to linear regression model, support vector regression model and the basic model of the artificial neural network.

The proposed model can be used to predict the arrival time of public transport in real time.

The possible direction of further research includes the usage of different models for individual routes or periods of the day.

6. References

[1] Agafonov A A, Sergeyev A V and Chernov A V 2012 Forecasting of the motion parameters of city transport by satellite monitoring data *Computer Optics* **36(2)** 453-458
[2] Jeon R and Rilett L 2005 Prediction model of bus arrival time for real-time applications *Transportation Research Record* **1927** 195-204
[3] Chanh H, Park D, Lee S, Lee H and Baek S 2010 Dynamic multi-interval bus travel time prediction using bus transit data *Transportmetrica* **6** 19-38
[4] Smith B, Williams B and Keith Oswald R 2002 Comparison of parametric and nonparametric models for traffic flow forecasting *Transportation Research Part C: Emerging Technologies* **10** 303-321
[5] Agafonov A A, Yumaganov A S and Myasnikov V V 2018 Big data analysis in a geoinformatic problem of short-term traffic flow forecasting based on a K nearest neighbors method *Computer Optics* **42(6)** 1101-1111 DOI: 10.18287/2412-6179-2018-42-6-1101-1111
[6] Xu H and Ying J 2017 Bus arrival time prediction with real-time and historic data *Cluster Computing* **20** 3099-3106
[7] Chen M, Liu X, Xia J and Chien S 2004 A dynamic bus-arrival time prediction model based on APC data *Computer-Aided Civil and Infrastructure Engineering* **19** 364-376
[8] Chien S J, Ding Y and Wei C 2002 Dynamic bus arrival time prediction with artificial neural networks *Journal of Transportation Engineering* **128** 429-438
[9] van Hinsbergen C, van Lint J and van Zuylen H 2009 Bayesian committee of neural networks to predict travel times with confidence intervals *Transportation Research Part C: Emerging Technologies* **17** 498-509
[10] Jeong R and Rilett L 2004 Bus arrival time prediction using artificial neural network model *Proc. of the 7th International IEEE Conference on Intelligent Transportation Systems* **1** 988-993
[11] Yang M, Chen C, Wang L, Yan X and Zhou L 2016 Bus arrival time prediction using support vector machine with genetic algorithm *Neural Network World* **26** 205-217
[12] Bin Y, Zhongzhen Y and Baozhen Y 2006 Bus arrival time prediction using support vector machines *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations* **10** 151-158
[13] Yu B, Yang Z Z and Yu B 2009 Hybrid model for multi-stop arrival time prediction *Neural Network World* **19** 321-332
[14] Agafonov A and Myasnikov V 2015 Traffic flow forecasting algorithm based on combination of adaptive elementary predictors *Communications in Computer and Information Science* **542** 163-174
[15] Yu B, Yang Z Z, Chen K and Yu B 2010 Hybrid model for prediction of bus arrival times at next station *Journal of Advanced Transportation* **44** 193-204
[16] Zheng W, Lee D H and Shi Q 2006 Short-term freeway traffic flow prediction: Bayesian combined neural network approach *Journal of Transportation Engineering* **132** 114-121
[17] Agafonov A A and Myasnikov V V 2016 Method for the reliable shortest path search in time-dependent stochastic networks and its application to GIS-based traffic control *Computer Optics* **40(2)** 275-283 DOI: 10.18287/2412-6179-2016-40-2-275-283
[18] Agafonov A A and Myasnikov V V 2018 Numerical route reservation method in the geoinformatic task of autonomous vehicle routing *Computer Optics* **42(5)** 912-920 DOI: 10.18287/ 2412-6179-2018-42-5-912-920
[19] Yu B, Lam W and Tam M 2011 Bus arrival time prediction at bus stop with multiple routes *Transportation Research Part C: Emerging Technologies* **19** 1157-1170
[20] Yin T, Zhong G, Zhang J, He S and Ran B 2017 A prediction model of bus arrival time at stops with multi-routes *Transportation research procedia* **25** 4627-4640
[21] Kingma D P and Ba J L 2014 Adam: A Method for Stochastic Optimization *Computing Research Repository* **15**

Acknowledgments
This work was supported by the RFBR (research projects N18-29-03135-mk, N 18-07-00605 A).