2014 IEEE 17th International Conference on
Intelligent Transportation Systems (ITSC)
October 8-11, 2014. Qingdao, China

Historical Data based Real Time Prediction of Vehicle Arrival Time

Santa Maiti, Arpan Pal, Arindam Pal, T Chattopadhyay and Arijit Mukherjee[1]

Abstract— **In recent times, most of the industries provide transportation facility for their employees from scheduled pick-up and drop points. In order to reduce longer waiting time, it is important to accurately predict the vehicle arrival in real time. This paper proposes a simple, lightweight yet powerful historical data based vehicle arrival time prediction model. Unlike previous work, the proposed model uses very limited input features namely vehicle trajectory and timestamp considering the scarcity and unavailability of data in the developing countries regarding traffic congestion, weather, scheduled arrival time, leg time, dwell time etc. Our proposed model is evaluated against standard Artificial Neural Network (ANN) and Support Vector Machine (SVM) regression models using real bus data of an industry campus at Siruseri, Chennai collected over four months of time period. The result shows that proposed historical data based model can predict two and half (approx.) times faster than ANN model and two (approx.) times faster than SVM model while it also achieves a comparable accuracy (75.56%) with respect to ANN model (76%) and SVM model (71.3%). Hence, the proposed historical data based model is capable of providing a real time system by balancing the trade-off between prediction time and prediction accuracy.**

I. INTRODUCTION

In last few decades it has been observed that most of the industries were established in the outskirts of city or locality. In these industrial areas employees face a major challenge in communication due to inadequacy in public transport. Also, in case of industries situated in central areas or well-connected areas, employees face difficulties to reach their destination due to the facts like increasing traffic, randomness and uncertainty of public transport. Therefore, now-a-days most of the industries provide pickup and drop facility to maintain the timeliness and to comfort their employees. While pickup and drop service solves transportation related problem, it is vital to predict vehicle's arrival time in real time at pickup and drop points to avoid longer waiting time.

A very similar case is identified in our organization campus located at Siruseri, Chennai, India. The campus is spread over 70 acres and designed to accommodate around 40000 employees. The distance of the campus from the main localities is around 50-70 Km. To provide transportation for employees, 200 buses run throughout a day in six different routes in fifteen minutes interval. Though the starting time of the buses are fixed, the arrival time of bus at pickup or drop points varies due to unpredictable traffic, weather change, variation in road and bus condition, availing alternative route etc. Inconsistency in bus arrival time puts the employee in dilemma of waiting for the bus or to avail some alternative means of transportation. Continuous bus tracking and update can be a solution. But, the approach is cumbersome and non-scalable. Alternatively, the bus movement data can be recorded, analyzed and modeled to predict arrival time at pickup and drop points. The prediction model should be accurate, fast, robust and scalable enough to handle thousands of prediction requests.

From the literature survey we observed that most of the bus arrival time prediction model use enriched feature set to predict bus arrival time. But, the main difficulties for building such system in the developing countries are unavailability of data and inaccuracy in reported data regarding traffic, weather, road etc. The proposed models also require longer training time to enhance the prediction accuracy. Whereas, in a large scale real world scenario new trip data need to be added into the database in daily or hourly basis. Therefore, the objective of our work is to provide a lightweight real time prediction model, retaining prediction accuracy while considering limited feature set.

After an extensive analysis we propose a simple historical data based (HD) model which considers vehicle trajectory and timestamp as input features. In the proposed methodology we have divided the bus route into bus stop wise segments and the travel time into fifteen minutes equal intervals. Segment wise bus speed and time slot wise bus speed are used to build the model. To substantiate the efficacy of the proposed approach, we have compared our model with ANN regression model and SVM regression model based approaches considering time and location as feature set. Result shows the proposed HD model performs approximately two and half times and two times faster than ANN model and SVM model respectively with comparable prediction accuracy and therefore capable to provide real time prediction system.

The organization of the paper is as follows. In Section II, we discuss the state-of-the-art. Section III talks about the proposed methodology of building arrival time prediction model. The experimental result is discussed in Section IV. Section V concludes the paper.

II. BACKGROUND

In order to predict bus arrival time, researchers have proposed several prediction models with different feature sets. In [1] Jeong and Rilett compared historical data based model, regression model and artificial neural network (ANN) model where arrival time of bus in previous bus stop, traffic congestion (schedule adherence time) and dwell time at a bus stop are considered as feature set. For prediction

[1]Authors belong to Innovation Lab, Tata Consultancy Services, Kolkata, India. Email: {santa.maiti, arpan.pal, arindam.pal1, t.chattopadhyay, mukherjee.arijit}@tcs.com

they also considered four separate time periods - weekend, weekday peak hours, weekday non-peak hours and weekday evening. The work established that historical data based model outperformed the regression model and artificial neural network (ANN) model outperformed the historical data based model for AVL (Automatic Vehicle Location) data collected in Houston, Texas Metro area. Historical data based model comparatively performed better during weekday peak hours due to less variation in bus speed. The reason behind the poor performance of regression model is its inability to determine the non-linear relationship between bus arrival time and schedule adherence time. Though ANN performed best in above mentioned test bed, it has been observed that ANN requires a large amount of training data to estimate the distribution of input pattern [2]. It is also difficult to generalize the results because of its over-fitting nature. Additionally, performance of ANN solely depends on selection of control parameters including relevant input variables, hidden layer size, learning rate, and momentum [3], [4], [5], [6].

In [7], a dynamic model consisting of two primary components (ANN model for predicting bus travel time between time points and Kalman filter-based dynamic algorithm to adjust the arrival-time) is proposed considering four feature sets - day, time, weather and segment. This approach is experimented on bus data (scheduled arrival time, date, time, bus door open and close time, stop sequence, latitude - longitude, number of passengers travel distance, dwell time, door open and close time, leg time etc.) of Essex county, Union county and Middlesex county in New Jersey.

Arrival time prediction using SVM model [2], proposed by Bin et al. claimed that SVM model outperformed the ANN model though its training time scales somewhere between quadratic and cubic. Four patterns - peak time and sunny day, off-peak time and sunny day, peak time and rainy day, off-peak time and rainy day are selected as study patterns and segment, travel time of current segment, and the latest travel time of next segment are used as the feature sets.

In a very different way, [8] estimated bus arrival time based on historical model by crowdsourcing mobile phone data (cell tower signals, movement statuses, audio recordings etc.). Difficulty of this model lies behind correctly classifying bus path in case of overlapping routes and the model fails in case of bus with no passenger.

We have implemented the state of art namely historical data based model, ANN regression model and SVM regression model in the limited data environment. We tuned the parameters to obtain best result of each model and compared the results. The proposed historical data based model only considered average travel time of a vehicle which can not capture the relationship of vehicle speed on time of a day. We enhance the model by considering vehicle speed dependency on time of the day.

III. PROPOSED METHODOLOGY

In this work we have proposed a historical data based model considering vehicle location and timestamp as the input feature set. The model predicts arrival time of a vehicle on a particular day for given stop based on the arrival time of the vehicle at the previous stop and the stored historical data. Our proposed methodology has three separate modules namely, data collection, data analysis and pre-processing and model development. An overview of our approach is shown in Fig. 1.

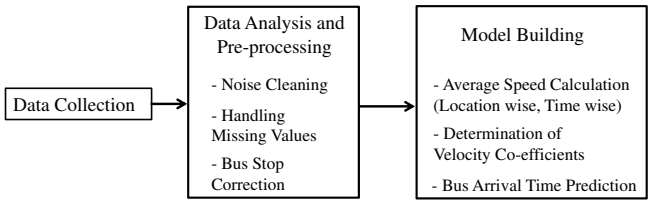

Fig. 1. Block diagram of proposed model

A. Data Collection

In order to train and test our model, we have chosen the 73.8 Km long bus route (Fig. 2) from Manali to Siruseri Chennai, India via Chennai bypass road (AH 45). Connect-Port X5 R devices are used for sensing which are installed in buses, running along the route to collect the trip related information. During each trip, the sensor sends six attribute values specifically, sensing timestamp, bus location (latitude, longitude), speed, associates card swiping timestamp and number of associates boarding and getting down from bus. We have only used sensing timestamp and the bus location data to build our model. The bus data is collected over a four-month time period. The collected data is thoroughly analyzed to identify the factors that can directly or indirectly impact on bus arrival time.

B. Data Analysis and Pre-processing

During analysis, we have realized that the raw sensor data contain missing values, erroneous readings and the data are in incompatible (unequally sampled, reverse order latitude-longitude reading) format. Therefore, the data need to be pre-processed before feeding it into the proposed model.

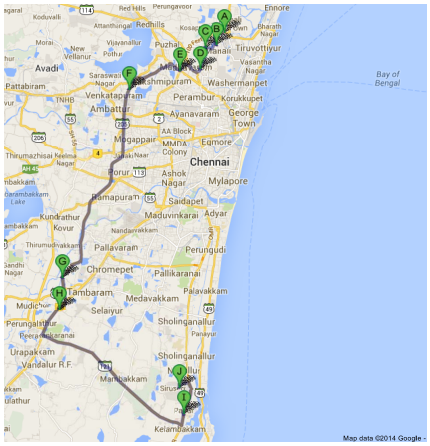

Fig. 2. Route of the bus

1) Inaccurate and missing values: We have plotted the latitude and longitude of each day bus data to track the movement of the bus. We have noticed that for some of the days, the sensor data is totally wrong which may be due to some sensor malfunctioning or the bus following a totally different route. To eliminate the faulty days we have considered small grids along the route and checked whether the bus locations are within the grids or not. If more than 10% locations of a day are outside of the grids, we have considered it as a faulty day. The grids are also used to identify the missing data. If for a particular day there is no entry for consecutive grids then the day has missing values. We have discarded a day, if there is no data for more than two consecutive grids. Otherwise we have interpolated the missing values.

2) Correction of Bus Stop Location: In the test dataset there are ten bus stops along the route. Though the locations of bus stops are given, we have identified that all the stops are not exactly situated on the bus route. We have plotted location wise bus speed and found that the bus is skipping few bus stops and stopping at nearby places on the route. We analyzed the fact and found that the stop locations are shifted from the designated locations as per employee demands. Primarily, we have used Google map[1] to locate nearby bus stops with respect to given stop location. For two bus stops, the stop locations cannot be identified as the given locations are apart from the route; according to the Google map there is no nearby bus stop and the speed of the bus on the nearest location of the given stop location, on the route is quite high. To solve this issue we have considered the stop's nearest grids positioned on the route and studied the speed variation. The middle point of the grid having less average speed and less speed variance is considered as the bus stop. In Fig. 3 black solid circle shows the given locations of the bus stops. Google map wise identified bus stops are marked using blue stars. Solid green box shows the grid wise identified bus stops.

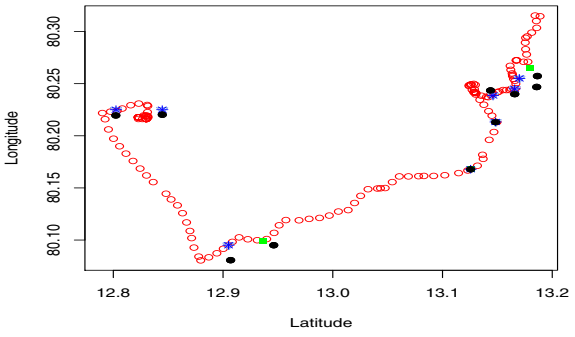

Fig. 3. Correction of bus stop location

C. Model Building

We have plotted four months speed data and found that the speed value follows a similar, periodical pattern with respect to time and location. Therefore, we have decided to analyze the speed pattern with respect to time and location. In our test case, the sensors send data throughout a day, before starting from source and also after reaching at destination. To avoid the unwanted data, we have limited timewise (8am-11am and 7:30pm-11pm) and loationwise (start to destination) sensor reading.

1) Location wise Speed Calculation: We have also identified that the sensor data (location) are unequally sampled which means there is no guarantee to get data from same location for each day. Hence, we need to group the readings zone wise. Based on the identified ten bus stops, we have divided the total route into nine segments. Length of each segment (distance between two consecutive bus stops) is estimated using Google map by mentioning intermediate locations to avoid multi path options. Next, we have identified nearest timestamp reading for each bus stop and calculated average speed for each segment. The speed variation for each segment over four month time period is shown in Fig. 4 using boxplot [9] where outliers are identified as values outside inner fences numerical boundaries[2]. Note that, there are few outliers in the boxplot. After analyzing the data we have understood that in some cases the nearest sensor reading is a bit far from the bus stop location which results the outliers. The median speed value $(VL_{i,j})$ is calculated using Equ. 1 for each segment and it is considered as the representative speed for that segment.

$$VL_{i,j} = median_N \frac{d_{i,j}}{t_{i,j}} \tag{1}$$

Where i varies from 1 to 9 and $j = i+1$. $d_{i,j}$ is the distance between bus stop i and j. $t_{i,j}$ is the time interval between bus stop i and j. N is the total number of days.

2) Time Slot wise Speed Calculation: Likewise location data, the timestamp sensor reading is also unequally sampled. To identify the speed pattern with respect to time we have divided the bus run time into fifteen minutes time interval and estimated the distance a bus travels in that interval. Fifteen minutes interval is selected as we observed that most of the day the duration of the trip is around two hours so on an average for each segment the bus takes fifteen minutes. However, the interval value needs to be calibrated to decide best interval that can give distinguishable speed value. We have considered this as our future work. Fig. 5 shows the speed variation for each fifteen minutes time interval for pickup trip over four month data. Note that, Fig. 5 has many outliers which can be reduced by choosing a better interval. Similar to the location, median speed $(VT_{i,15})$ for each fifteen minutes interval is considered as the representative speed and calculated using Equ. 2.

$$VT_{i,15} = median_N \frac{d_{i,15}}{t_{15}} \tag{2}$$

Where i varies from 1 to 12 for pickup trip and from 1 to 14 for drop trip. $d_{i,15}$ is the distance covered into i^{th} fifteen

[1]Google map, https://maps.google.com

[2]How to Calculate Outliers, www.wikihow.com/Calculate-Outliers

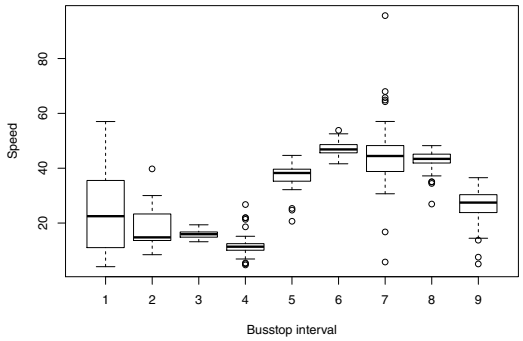

Fig. 4. Segment wise speed variation

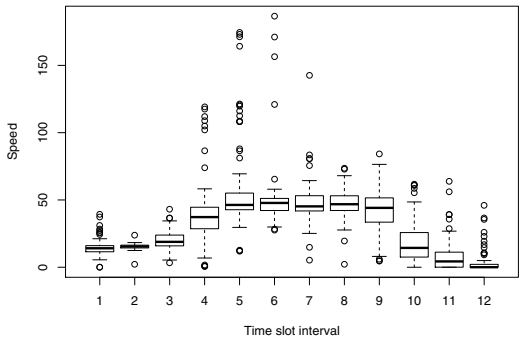

Fig. 5. Time slot wise speed variation

minute time interval and t_{15} is fifteen minutes time interval. N is the total number of days.

3) Determination of Time and Location Coefficient: The analysis shows the speed of the bus varies with respect to both location and time. To decide the weightage of location wise and time wise speed value, we assume a linear relationship between time, location and speed as given in Equ. 3.

$$V_{i,j,k} = \alpha\left(VL_{i,j}\right) + \beta\left(VT_{l,15,k}\right) \tag{3}$$

Where i varies from 1 to 9 and $j = i + 1$. α and β are coefficients for location and time respectively. $V_{i,j,k}$ is k^{th} day actual velocity between bus stop i and j and $VL_{i,j}$ is the representative velocity between bus stop i and j, To evaluate $VT_{l,15,k}$ first we need to find out the fifteen minutes time intervals (let m^{th} interval to n^{th} time interval) between bus stop i and j on k^{th} day and then the average speed in m^{th} to n^{th} intervals are calculated as per Equ. 4

$$VT_{l,15,k} = \frac{\sum_{l=m}^{n} VT_{l,15}}{n - m + 1} \tag{4}$$

The four months bus data are divided into two parts. $9/10^{th}$ data are used for training purpose (to determine α and β) and $1/10^{th}$ data are used for testing purpose. For each day, nine equations are obtained using Equ. 3 for nine segments (ten bus stops). We have used least square approximation [10] to solve these equations. We have observed that the variation of α and β is very less over training time period. Therefore, we have considered average α and average β as the final α and β values respectively.

4) Arrival Time Prediction: After determining α and β values, we can find out the arrival time of a bus to a given bus stop on k^{th} day using Equ. 5.

$$T_{j,k} = T_{i,k} + \frac{d_{i,j}}{V_{i,j,k}} \tag{5}$$

where, $T_{j,k}$ is the arrival time of the bus at j^{th} bus stop. $T_{i,k}$ is the arrival time of the bus at i^{th} bus stop; $j = i + 1$. $d_{i,j}$ is the distance between i^{th} and j^{th} bus stop. $V_{i,j,k}$ is calculated using Equ. 3.

5) Computational Complexity Analysis: The computational complexity of our approach is calculated as follows. Let's assume, n is the number of training data; n' is the number of testing data (number of prediction request); l is the number of segments and t is the number of time slots. Then, location wise average speed calculation requires $O(n.l)$ time and time slot wise average speed calculation requires $O(n.t)$ time. To find bus stop corresponding segment and time slot our approach takes $O(n(l + l.log(t)))$ time. Next, in training phase we have to solve l number of equations with two features (location and time). It takes $O(2^2.l.n)$ time. As, l and t are constants, the training time complexity our proposed approach is $O(n)$. For testing phase the computational complexity is $O(n'.log(t)) \approx O(n')$.

IV. Experimental Results

We have evaluated our proposed model from two perspective - output quality (prediction accuracy) and execution time requirement (training and testing). This section presents a detail description of experiments and the results observed.

A. Experimental Setup

All the experiments and analysis are carried out in Windows environment (Windows 7) with Intel Core i5 (3.1GHz) processor and 4.0 GB memory. All implementation are done in R platform.

For experiment, we have used our organization bus data, collected during pickup trip and drop trip over four month time period. In 73.8 Km long route from Manali to Siruseri Chennai there are ten pickup and drop points. From the recorded data we have identified each day arrival time of the bus at all bus stops. We have also estimated the distances between two consecutive bus stops. $9/10^{th}$ data are used for training purpose and $1/10^{th}$ are data used for testing purpose. For training and testing we have followed k(10) fold cross validation process to avoid data biasness.

B. Experimental Metrics

We have profiled our proposed model to estimate the run time required for training and testing purpose. For measuring prediction accuracy two different metrics - RMSE (Root Mean Square Error)[11], percentage error (PE) are used. RMSE is calculated as per Equ. 6.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(T_{i,j} - T'_{i,j}\right)^2} \tag{6}$$

where N is the number of test samples; $T'_{i,j}$ is the actual arrival time of bus on a sample day i at bus stop j; $T'_{i,j}$ is the predicted arrival time of bus on a sample day i at bus stop j. The RMSEs are calculated for each route segment. The PE can be obtained from Equ. 7.

$$PE = \left(\frac{1}{N} \sum_{i=1}^{N} \frac{|T_{i,j} - T'_{i,j}|}{|T_{i,j} - T_{i,j-1}|} \right) * 100\% \qquad (7)$$

where N is the number of test samples; $T'_{i,j}$ is the actual arrival time of bus on a sample day i at bus stop j; $T'_{i,j}$ is the predicted arrival time of bus on a sample day i at bus stop j; $T'_{i,j}$ is the actual arrival time of bus on a sample day i at bus stop $(j-1)$ (previous bus stop).

C. Experimental Process and Results

From the recoded bus data, first we have calculated time slot wise average velocity and location wise average velocity as described in Sec. III-C. Next, time and location wise velocity coefficients (α and β) are calculated using training data. α and β are determined as 0.004 and 1.028 respectively. It indicates that, with respect to our test data, location wise speed has more impact than time slot wise speed to determine bus arrival time. Variation of time interval may improve the α value. In the testing phase, the bus arrival time at a bus stop is calculated based on location wise average velocity, time wise average velocity, velocity coefficients and arrival time of the bus in previous bus stop.

To compare our methodology with benchmarked arrival time prediction approaches (ANN and SVM regression) we have used same training and testing data. In this experiment we have used resilient backpropagation with weight backtracking ANN model [12] with a single hidden layer as one hidden layer is sufficient for the large majority of problems[3,4]. There are few thumb rules to determine number of neurons in the hidden layer[2,3]. In our case, we find out average RMSE of training dataset for one to ten neurons. Fig. 6 shows RMSE with respect to number of neurons. In Fig. 6, we can notice that for seven neurons the average RMSE and standard deviation of RMSE is less. Therefore, we select single hidden layer with seven neurons for further experiments. Bus arrival time in previous bus stop, location (latitude and longitude) of previous and target bus stops are taken as input parameters and bus arrival time at target bus stop is considered as output parameter during training session.

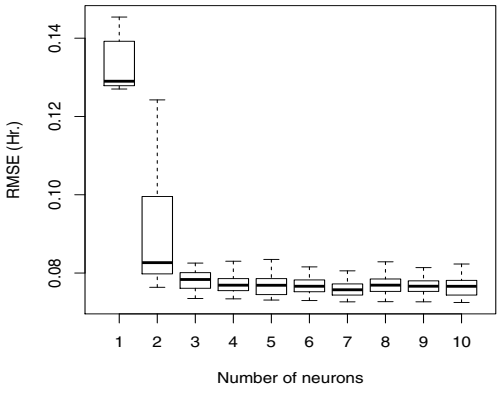

Fig. 6. RMSE with respect to number of neurons

Likewise ANN, same input and output parameters are used to train SVM regression model [13] for predicting bus arrival time. We find out RMSE value of training data with respect to four kernels namely linear, polynomial, radial basis (RBF) and sigmoid as shown in Fig. 7 and tuned corresponding parameters (gamma and cost) to find best suited parameter value. From the figure it is clear that RBF kernel outperforms other three kernels. Hence, SVM model with RBF kernel is used in further comparison.

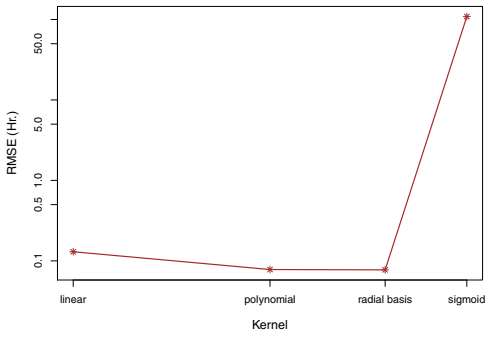

Fig. 7. RMSE with respect to number of kernels

Next, we check the performance of selected ANN, SVM model and proposed HD model on test data. The comparative study of the three models based on RMSE value and runtime are shown in Fig. 8 and Fig. 9.

The comparative PE, training time and testing time for three models are shown in Fig. 10. The figures prove that for the constrained dataset, our proposed HD model performs around 2.5 times faster (testing time) than ANN regression model and around 2 times faster than SVM regression

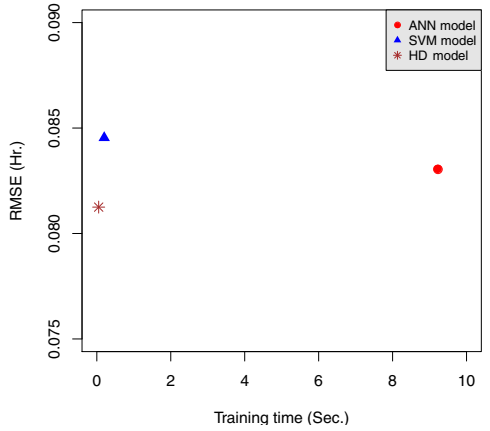

Fig. 8. RMSE Vs training time for ANN, SVM and HD model

[3]Cross Validated, http://stats.stackexchange.com/questions/181/how-to-choose-the-number-of-hidden-layers-and-nodes-in-a-feedforward-neural-netw

[4]Heaton Research, http://www.heatonresearch.com/node/707

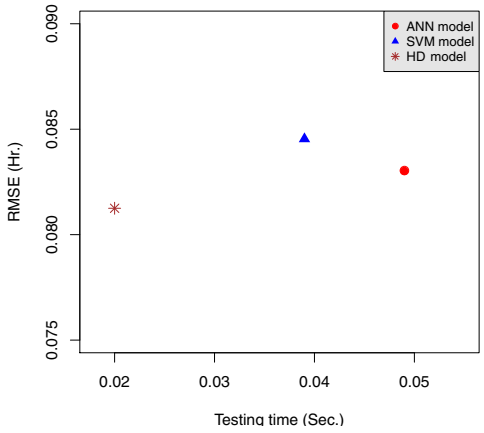

Fig. 9. RMSE Vs testing time for ANN, SVM and HD model

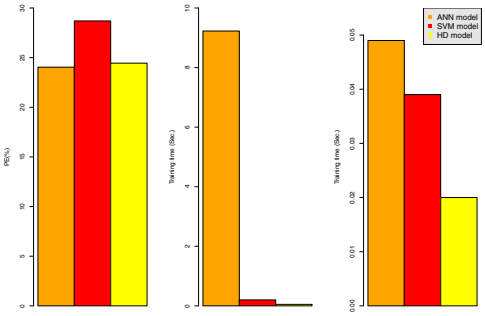

Fig. 10. Percentage error, training time, testing time of ANN, SVM and HD model

model while achieving 75.56% prediction accuracy. Hence the model can be used for bus arrival time prediction in real time.

V. Conclusion

In this paper we have addressed industry transportation related vehicle arrival time prediction problem, specially applicable for developing countries where availability of adequate and accurate data is a big challenge. Real time bus data are collected on a transit route from Manali to Siruseri Chennai, India. To predict bus arrival time, we have proposed a simple, lightweight historical data based model. Due to the data constraint, very limited information (bus location and timestamp) are considered as input features. Analyzing the experimental result we realize that the historical data based model retains prediction accuracy in limited dataset by considering both location and time component. In the proposed model, location component captures location wise vehicle speed which includes road condition, dwell time and time component captures time wise vehicle speed which includes traffic congestion. We found that, the proposed model outperforms ANN model and SVM model with respect to training and testing time while retaining prediction accuracy. Therefore, the model fits for real time prediction system.

The present model is capable to predict arrival time of a vehicle in a stop when arrival time of the vehicle in previous stop is given. As a future extension, we have planned to provide arrival time of the vehicle in a stop for given arrival time of any previous stop and also predicted arrival time at destination. Though the experimental result is quite promising, we need to consider a longer range of data (throughout a year) to examine the performance of the proposed model over weather change. In this case study we have used ConnectPort X5 R sensor to monitor the vehicle movement. In a cost effective way, GPS enabled mobile phones can also be used.

References

[1] R. Jeong and L. R. Rilett, "Bus Arrival Time Prediction Using Artificial Neural Network Model," in *proceedings of the 7th International IEEE Conference on Intelligent Transportation Systems*, October 2004, pp. 988–993.

[2] Y. Bin, Y. Zhongzhen, and Y. Baozhen, "Bus Arrival Time Prediction Using Support Vector Machines," *Journal of Intelligent Transportation Systems*, vol. 10, no. 4, pp. 151–158, 2006.

[3] S. Lawrence, C. L. Giles, and A. C. Tsoi, "Lessons in Neural Network Training: Overfitting may be Harder than Expected," in *proceedings of the Fourteenth National Conference on Artificial Intelligence, AAAI*, 1997, pp. 540–545.

[4] J. E. Moody, "The Effective Number of Parameters: An Analysis of Generalization and Regularization in Nonlinear Learning Systems," in *proceeding of: Advances in Neural Information Processing Systems*, vol. 4, December 1991, pp. 847–854.

[5] W. S. Sarle, "Stopped Training and Other Remedies for Overfitting," in *proceedings of the twenty-seventh symposium on the interface of computing science and statistics*, 1995, p. 352360.

[6] A. S. Weigend, "On Overfitting and the Effective Number of Hidden Units," in *proceedings of the 1993 Connectionist Models Summer School*, 1994, p. 335342.

[7] M. Chen, X. Liu, J. Xia, and S. I. Chien, "A Dynamic Bus-Arrival Time Prediction Model Based on APC Data," *Computer-Aided Civil and Infrastructure Engineering*, vol. 19, no. 5, pp. 364–376, 2004.

[8] P. Zhou, Y. Zheng, and M. Li, "How Long to Wait?: Predicting Bus Arrival Time with Mobile Phone based Participatory Sensing," *IEEE Transactions on Mobile Computing*, vol. 99, p. 1, 2012.

[9] V. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artificial Intelligence Review*, vol. 22, no. 2, pp. 85–126, 2004. [Online]. Available: http://dx.doi.org/10.1007/s10462-004-4304-y

[10] G. Strang, *Introduction to Linear Algebra*. Wellesley-Cambridge Press, 2003.

[11] R. J. Hyndman and A. B. Koehler, "Another Look at Measures of Forecast Accuracy," *International Journal of Forecasting*, vol. 22, no. 4, pp. 679–688, 2006.

[12] F. Gnther and S. Fritsch, "neuralnet: Training of Neural Networks," *The R Journal*, vol. 2, no. 1, June 2010.

[13] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, F. Leisch, C. Chang, and C. Lin, "Package e1071," Repository CRAN, February 2014.