# Using a Random Subspace Predictor to Integrate Spatial and Temporal Information for Traffic Flow Forecasting

Shiliang Sun and Changshui Zhang

State Key Laboratory of Intelligent Technology and Systems,
Department of Automation, Tsinghua University, Beijing, China, 100084
sunsl02@mails.tsinghua.edu.cn, zcs@mail.tsinghua.edu.cn

**Abstract.** Traffic flow forecasting is an important issue for the application of Intelligent Transportation Systems. Due to practical limitations, traffic flow records may be partially missing or substantially contaminated by noise. In this paper, a robust traffic flow predictor, termed random subspace predictor, is developed integrating the entire spatial and temporal information in a transportation network to cope with this case. Experimental results demonstrate the effectiveness and robustness of the random subspace predictor.

## 1 Introduction

In recent years utilizing signal processing and machine learning techniques for traffic flow forecasting has drawn more and more attention [1][2][3]. In this paper, we concentrate on using the ideology of random subspace to deal with the issue of traffic flow forecasting with incomplete data.

Up to the present, some approaches ranging from simple to complex are proposed for traffic flow forecasting, such as random walk, historical average, time series models, Kalman filter theory, neural network approaches, non-parametric methods, simulation models, fuzzy-neural approach, and Markov chain model [1][4]∼[7]. Although these methods have alleviated difficulties in traffic modelling and forecasting to some extent, most of them have not made good use of spatial information from the viewpoint of networks to analyze the trends of the object site. Besides, the existing methods hardly work when data used for forecasting is incomplete, i.e. partially missing or substantially contaminated by noise, while this situation often occurs in practice.

The main contribution of this paper is that we present a robust random subspace predictor to carry out traffic flow forecasting with incomplete data. Encouraging experimental results with real-world data show that our approach is reliable, accurate and robust for traffic flow modelling and forecasting with incomplete data.

L. Wang, K. Chen, and Y.S. Ong (Eds.): ICNC 2005, LNCS 3611, pp. 652–655, 2005.

## 2 Random Subspace Predictor

In a transportation network, there are usually a lot of sites (road links) related or informative to the traffic flow forecasting of the current site. However, using all the related links as input variables would involve much irrelevance, redundancy and would be prohibitive for computation. Consequently, a variable selection procedure is of great demand. Up to date many variable selection algorithms include variable ranking as a principal or auxiliary selection mechanism because of its simplicity, scalability, and good empirical success [8].

In this article, we use the norm of Pearson correlation coefficient $|R(i)|$ as a variable ranking criterion [8]. After the variable ranking stage, we can obtain $M$ input variables making up of the input space $S$ which is informative for the prediction of the object flow. However, using the entire high dimensional input space and small training data set to forecast traffic flow directly would arouse over-adaptation. Therefore, we adopt the random subspace ideology [9] and generate $K$ random subspaces $\{S_i\}_{i=1}^{K}$. Every time we randomly select a subspace of $m$ dimensions with replacement from the input space $S$ and use them as an input to forecast the object traffic flow. The dimension $m$ of random subspace is determined by the training set to make the forecasting results stable.

Given the selected input variables and the output variable, we utilize the Gaussian Mixture Model (GMM) to approximate their joint probability distribution whose parameters are estimated through the Competitive Expectation Maximization (CEM) algorithm. Then we can obtain the optimum prediction formulation as an analytic solution under the M.M.S.E. criterion. For details about the GMM, CEM algorithm and the prediction formulation, please refer to our previous articles [2][10]. In succession, we can repeat the above operation $K$ times and would obtain $K$ forecasting results $\{F_k\}_{k=1}^{K}$ for the current object flow. The outputs are combined using the fusion methodology of averaging of them and the average is taken as the final forecasting result $F(S)$ of our random subspace predictor (RSP), i.e. $F(S) = \frac{1}{K}\sum_{k=1}^{K} F_k(S_k)$ .

The RSP is very robust and can still work well when the input data are incomplete. If some incomplete data appear in the subspace, we can just remove this subspace and generate a new one till there are no incomplete data involved in the used subspace, and this does not influence the final fusion result much. We will discuss this matter in the following section with a specific instance.

## 3 Experiments

The field data analyzed is the vehicle flow rates recorded every 15 minutes along many road links by the UTC/SCOOT system in Traffic Management Bureau of Beijing, whose unit is vehicles per hour (veh/hr). We select a representative traffic patch to verify the proposed approach, which is given in Fig. 1. An arrow shows the direction of traffic flow, which reaches the corresponding downstream link from its upstream link. The raw data for use are of 25 days and totally 2400 sample points taken from March, 2002. To validate our approach objectively,

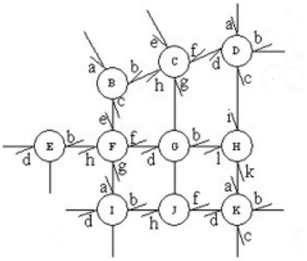

**Fig. 1.** A representative transportation network taken from the urban traffic map of Beijing

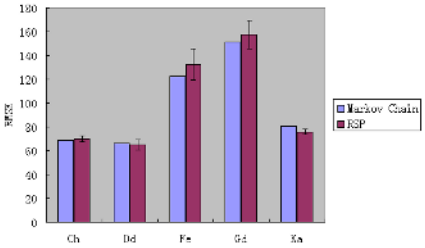

**Fig. 2.** A performance comparison of two methods for short-term traffic flow forecasting at five different road links

the first 2112 points (training set) of them are employed to learn parameters of GMM and the rest (test set) are employed to test the forecasting performance. To evaluate our presented approach, we utilize the Markov chain method as a base line [6]. The dimension of subspace in our approach is taken as 4 (the same with the Markov chain method in [6]) for each object site. The only difference between our proposed method and the Markov Chain method is that we utilize the whole spatial and temporal information to forecast while the latter only uses the temporal information of the object site.

We take road link $Ka$ as an instance to show our approach. $Ka$ represents the vehicle flow from upstream link $H$ to downstream link $K$. All the available traffic flows which may be informative to forecast $Ka$ in the analyzed transportation network includes $\{Ba, Bb, Bc, Ce, Cf, Cg, Ch, Da, Db, Dc, Dd, Eb, Ed, Fe, Ff, Fg, Fh, Gb, Gd, Hi, Hk, Hl, Ia, Ib, Id, Jf, Jh, Ka, Kb, Kc, Kd\}$. Considering the time factor, in order to forecasting the traffic flow $Ka(t)$, we might need judge its relevancy from the above sites with different time indices, such as $\{Ba(t-1), Ba(t-2), ..., Ba(t-d)\}$, etc. In this paper, $d$ is taken as 100 empirically. We retain $M$ ($M = 20$ in this article) most correlated traffic flows constructing the entire input space which are selected with the correlation variable ranking criterion with their corresponding correlation coefficients. These correlation coefficients varies from 0.967 for $Hi(t-1)$ to 0.954 for $Eb(t-3)$. With the selected input space, our RSP generates $K$ ($K = 10$ in this article) random subspaces. The joint probability distribution between the random subspace input and the output $Ka(t)$ are approximated with GMM, and thus we can derive the prediction formulation, carry out traffic flow forecasting on the test data set. Finally we combine the $K$ outputs to form one forecasting output. In addition, we also conducted experiments on four other traffic flows $Ch, Dd, Fe, Gd$. Fig. 2 gives the forecasting results of all the five road links with performances evaluated by Root Mean Square Error (RMSE).

From the experimental results, we can find the effectivity and robustness of our approach which integrates both spatial and temporal information for forecasting. Generally speaking, the RSP has obtained similar accuracy with the Markov chain method. But the PSP can be used in case of incomplete data. In forecasting $Ka(t)$, the Markov chain method would use $Ka(t-1)$, $Ka(t-2), Ka(t-3), Ka(t-4)$ as input. However, if $Ka(t-1)$ is incomplete, then Markov chain method would lose its applicability while the RSP can still work since it can use the other $M-1$ input variables to generate subspace and the final performance would not change much. Further, the RSP can still work stably if multiple input flows are incomplete.

## 4 Conclusions

In this paper, we propose a robust random subspace predictor integrating the whole spatial and temporal information available in a transportation network to carry out traffic flow forecasting. It is simple, effective and still work well when encountering incomplete data. Experimental results demonstrate the applicability of the random subspace predictor.

## Acknowledgements

This work was supported by the National Natural Science Foundation of China under Project 60475001.

## References

1. William, B.M.: Modeling and Forecasting Vehicular Traffic Flow as a Seasonal Stochastic Time Series Process. Doctoral Dissertation. University of Virginia, Charlottesville (1999)
2. Sun, S.L., Zhang C.S., Yu G.Q., Lu, N.J., Xiao F.: Bayesian Network Methods for Traffic Flow Forecasting with Incomplete Data. ECML 2004, Lecture Notes in Artificial Intelligence, Vol. 3201. Springer-Verlag, Berlin Heidelberg (2004) 419-428
3. Yang, L.C., Jia, L., Wang, H.: Wavelet Network with Genetic Algorithm and Its Applications for Traffic Flow Forecasting. Proceedings of the Fifth World Congress on Intelligent Control and Automation, Vol. 6 (2004) 5330 - 5333
4. Chrobok, R., Wahle, J., Schreckenberg, M.: Traffic Forecast Using Simulations of Large Scale Networks. Proceedings of IEEE Intelligent Transportation Systems Conference (2001) 434-439
5. Yin, H.B., Wong, S.C., Xu, J.M., Wong, C.K.: Urban Traffic Flow Prediction Using a Fuzzy-Neural Approach. Transportation Research, Part C, Vol. 10 (2002), 85-98
6. Yu, G.Q., Hu, J.M., Zhang, C.S., Zhuang, L.K., Song J.Y.: Short-Term Traffic Flow Forecasting Based on Markov Chain Model. Proceedings of IEEE Intelligent Vehicles Symposium (2003) 208 - 212
7. Sun, S.L., Yu, G.Q., Zhang, C.S.: Short-Term Traffic Flow Forecasting Using Sampling Markov Chain Method with Incomplete Data. Proceedings of IEEE Intelligent Vehicles Symposium (2004) 437 - 441
8. Guyon, I., Elisseeff, A.: An Introduction to Variable and Feature Selection. Journal of Machine Learning Research, Vol. 3 (2003) 1157-1182
9. Ho, T.K.: The Random Subspace Method for Constructing Decision Forests. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 20 (1998) 832-844
10. Zhang, B.B., Zhang, C.S., Yi, X.: Competitive EM Algorithm for Finite Mixture Models. Pattern Recognition, Vol. 37 (2004) 131-144