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

A Deep Learning Based Approach for Traffic Data Imputation

Yanjie Duan, Yisheng Lv, Wenwen Kang, Yifei Zhao

Abstract— **Traffic data is a fundamental component for applications and researches in transportation systems. However, real traffic data collected from loop detectors or other channels often include missing data which affects the relative applications and researches. This paper proposes an approach based on deep learning to impute the missing traffic data. The proposed approach treats the traffic data including observed data and missing data as a whole data item and restores the complete data with the deep structural network. The deep learning approach can discover the correlations contained in the data structure by a layer-wise pre-training and improve the imputation accuracy by conducting a fine-tuning afterwards. We analyze the imputation patterns that can be realized with the proposed approach and conduct a series of experiments. The results show that the proposed approach can keep a stable error under different traffic data missing rate. Deep learning is promising in the field of traffic data imputation.**

I. INTRODUCTION

Traffic data is a fundamental component for applications and researches in transportation systems. Both route planning for individuals and transportation management and control for researchers and governments need sufficient traffic data [1]. However, real traffic data collected from loop detectors or other channels are often incomplete due to various reasons. These missing data make traffic analysis and other operations difficult in practice [2]. Traffic data imputation aims to fill in these missing data points as accurate as possible. It has been a hot topic [3] and will remain hot as traffic data are getting richer. In this paper, we propose a approach based on deep learning for traffic data imputation. To the best of our knowledge, it is the first time to introduce deep learning to the field of traffic data imputation. The specific architecture we use is denoising stacked autoencoder (DSAE) [4] which is composed of a denoising autoencoder (DAE) and stacked autoencoders (SAE) [5]. Considering the various patterns of traffic data imputation, we analyze all the possible patterns that can be realized with the proposed approach. To evaluate the performance of the approach, a series of experiments are conducted. The results show that our approach is rather competitive both in accuracy and in versatility. Additionally, our method needs little domain knowledge to obtain the imputed data during applications. That is convenient for researchers to get a complete data set in their researches.

This work was supported in part by the National Natural Science Foundation of China under Grants 71232006, 61233001, 61174172, 61104160, 61203079, 61203166.

Yanjie Duan, Wenwen Kang and Yifei Zhao are with Qingdao Academy of Intelligent Industries, Qingdao, Shandong, 266109, China. (e-mail: duanyanjie2012@ia.ac.cn).

Yisheng Lv is with Beijing Engineering Research Center of Intelligent Systems and Technology, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China.

The rest of this paper is organized as follows: Section II reviews the related work in traffic data imputation and gives a brief introduction to deep learning. Section III describes the imputation approach based on DSAE. Section IV shows the experiments and results analysis. Section V makes the conclusions of this paper and gives some points of view in future work.

II. RELATED WORK

A. Traffic Data Imputation

Because of the necessity of traffic data imputation, there have been many researches investigating this problem using a wide range of theories and methods. These methods are mainly based on time series analysis and prediction, non-parameter regression and statistical learning estimation [3]. Time series analysis and prediction method often uses historical data of the location to build a prediction framework and predict the missing data of the same location. The simplest method is to replace the missing data with the data in history usually the previous time period or the previous time interval. Another method is the autoregressive integrated moving-average (ARIMA) method [6] which is in common use. Non-parameter regression method often uses the data of neighboring locations or neighboring states to estimate the missing data of the current location. The missing data is estimated by the average or the weighted average of the neighboring data. A typical example of this method is k-NN method [7] of which the key work is to determine the neighbors by an appropriate distance metric. Statistical learning estimation method often uses the observed data to learn a scheme, then inferences the missing data in a fashion of iteration. A classical method is the Markov Chain Monte Carlo (MCMC) multiple imputation method [8][9]. The basic idea of MCMC multiple imputation method is to treat the missing data as a parameter of interest, draw a series of samples of the parameter and estimate the parameter using the samples. That means the imputation of the missing data is a combination of multiple imputed values instead of only one value. Another method is the neural networks [10] which is promising to obtain more accurate imputations than traditional imputation methods given more observed data. Due to the complex patterns of traffic data and the diversity of application scenarios, different methods are being used in the field of traffic data imputation. However, the existing methods usually treat the missing data separated from the observed data and need extra domain knowledge about the specific data and the method. From a new perspective, this paper considers the imputation process as the recovery of data which consists of missing and observed data.

B. Deep Learning

Deep learning has been developing fast since 2006 when deep neural networks constructed of Restricted Boltzmann Machines (RBMS) [11] was proposed. A kind of deep learning method trains a multilayer neural network with a small number of nodes in the central layer and reconstructs the input data in the output layer. The training process includes a layer-wise pretraining step and a fine-tuning step instead of training the whole network directly. In the layer-wise pretraining step, each layer is trained in an RBM [11][12] or in an autoencoder [5]. In an RBM shown in Fig. 1, there is a single hidden layer connected with the visible layer and the connection is undirected, symmetrical. While

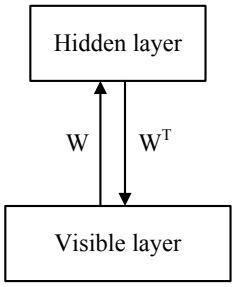

Fig. 1. The structure of an RBM

in an autoencoder shown in Fig. 2, there is also a single hidden layer apart from an input layer and an output layer. The hidden layer is connected with the input layer and the output layer in a directed, unsymmetrical manner. Usually, the output target is the same with the input thus the hidden layer is a code of the input data and can be decoded into the output data. In a way, an autoencoder is a shallow neural network in which the output target is the input data itself. Whether each layer is trained in an RBM or an autoencoder,

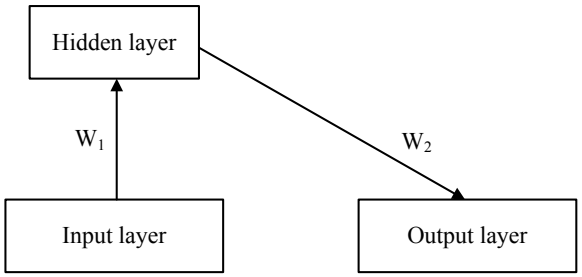

Fig. 2. The structure of an autoencoder

the hidden layer is treated as an input layer of the next RBM or autoencoder. Thus the multilayer neural network can be pretrained layer by layer only using the input data. After pretraining, the fine-tuning step adjusts the weights of the whole network using the output data in a supervised learning style.

These deep neural networks as introduced above reduce the dimensions of the input data in the central layer and can restore the input data in the output layer. The central layer can be seen as a representation of the input data. Considering the robustness of the representations, DAE was proposed in [4]. DAE aims to be robust to partial destruction of the input data. In a DAE shown in Fig. 3, the input data is the partially destroyed data of the raw input data. The error used in the training process is the difference between the raw input data and the output data. That means DAE can obtain almost the same representation of the destroyed input data and can restore the raw input data in the output layer. In the next section, we will introduce the DSAE based imputation architecture triggered by DAE and constructed by multiple autoencoders.

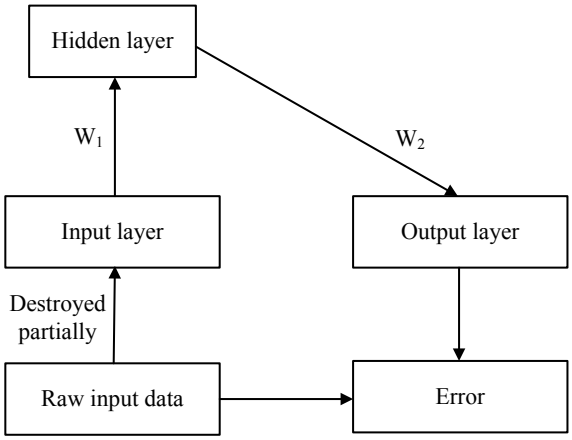

Fig. 3. The structure of a DAE

III. IMPUTATION BASED ON DSAE

In this section, we first describe the proposed deep learning architecture for traffic data imputation and the corresponding training process, then analyze traffic data imputation patterns that can be realized with the deep learning approach.

A. DSAE based imputation architecture

Traditional traffic data imputation methods often treat the missing data separated from the observed data. This paper treats the traffic data including missing data and observed data as a whole data item and try to restore the complete data from the incomplete data. In Fig. 3 assuming that the complete traffic data as the raw input data exists, the incomplete traffic data can be seen as the partially destroyed raw input data. Therefore DAE can be used for traffic data imputation. The proposed deep learning method for traffic data imputation is based on DSAE. DSAE shown in Fig. 4 can be seen as a DAE filled with multiple autoencoders in the middle.

The input of the DSAE based imputation architecture is the traffic data partially destroyed, the output target is the complete traffic data as the raw input data and the central layers are abstract representations of the traffic data. Define traffic data as $X = \{x_{ij}|i = 1, 2, ..., p, j = 1, 2, ..., q\}$, the raw input data as $X^r = \{x^r_{ij}|i = 1, 2, ..., p, j = 1, 2, ..., q\}$, the output data as $Y = \{y_{ij}|i = 1, 2, ..., p, j = 1, 2, ..., q\}$, where p is the total number of the data items and q is the dimension of one data item. Thus the (input, target) pairs are

$\{(x_i, y_i)|i = 1, 2, ..., p\}$. The recovery error for the whole imputation architecture is denoted as L.

B. Training process

The training process of the deep architecture proposed above includes the pre-training step and the fine-training step. In the pre-training step, each layer is trained by minimizing the reconstruction error $L(X, Y)$ of the current autoencoder. Assuming the parameters are θ, then

$$\theta = \arg\min_{\theta} L(X, Y) = \arg\min_{\theta} \frac{1}{2} \sum_{i=1}^{p} \|(x_i - y_i)\|^2,$$

where θ includes the weights and biases of the current autoencoder, X, Y is the input and the output of the current autoencoder. The hidden layer of the current autoencoder is the input layer of the next autoencoder. Thus the deep architecture can be pre-trained layer by layer. After pre-training, fine-tune all the parameters by minimizing the recovery error, then

$$\theta = \arg\min_{\theta} L(X^r, Y) = \arg\min_{\theta} \frac{1}{2} \sum_{i=1}^{p} \|(x_i^r - y_i)\|^2.$$

Until now, the deep architecture has been trained completely. The above training process is based on an assumption that the meta parameters such as the number of layers and the number of nodes in each layer have been set. In practise, we usually choose the proper meta parameters through experiences and experiments.

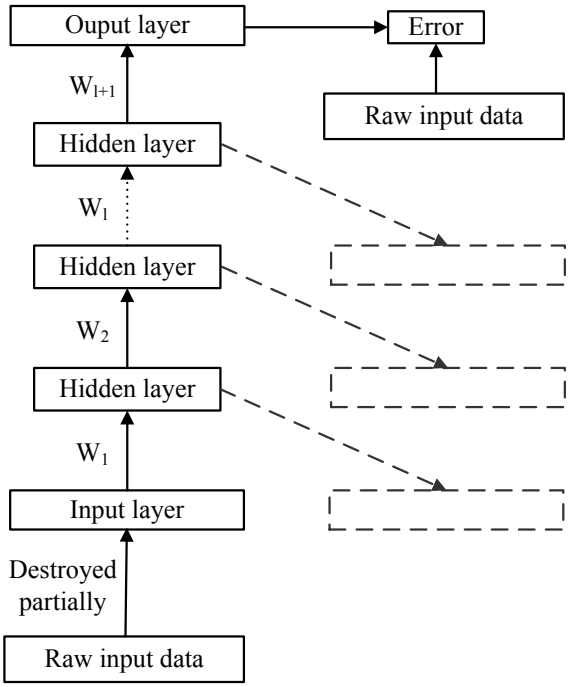

Fig. 4. The DSAE based imputation architecture

C. Imputation Patterns

In this paper, we focus on the imputation process of traffic data assuming all missing data have been recognized. In real world, the range of traffic data can be very wide including data from detectors and data from surveys. In terms of traffic data from detectors deployed in different locations, the proposed deep learning based imputation approach can use the data in different patterns. Traffic data from detectors naturally has a period (or cycle) of one day or one week. Assuming the obtained data are collected in N even time intervals during every period and contain D periods and M locations, then the domain of the traffic data can be described as a cube in Fig. 5.

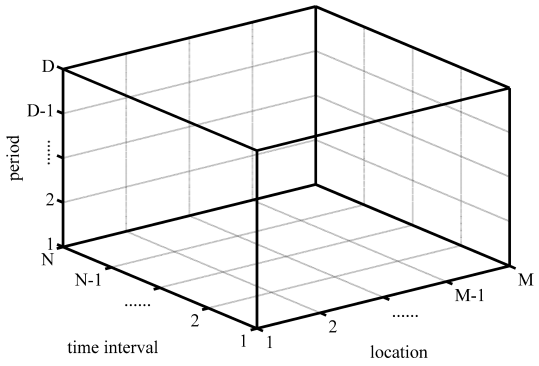

Fig. 5. The traffic data cube

According to the data structures adopted, there are various imputation patterns of the proposed approach. The simplest data structure is that one data item contains data of one period, single location thus the imputation process can be seen as a line recovery in the perspective of data cube. While the most complex data structure is that one data item contains data of multiple periods, multiple locations thus the imputation process can be seen as a 3-D recovery. All possible imputation patterns are listed in Table I. The proposed deep learning based imputation approach can be applied to realize any of these patterns according to actual demand without too much work of selecting features artificially.

TABLE I
IMPUTATION PATTERNS

	Period	Location	Pattern
1	Single	Single	1-D
2	Single	Multiple	2-D
3	Multiple	Single	2-D
4	Multiple	Multiple	3-D

IV. EXPERIMENTS AND RESULTS

A. Data description

The proposed deep learning based approach for traffic data imputation is experimented on the data set collected from the Caltrans Performance Measurement System (PeMS). The

system consists of more than 15000 detector stations and collects traffic data every 30 seconds. The raw 30-second dataset includes gaps due to various reasons. PeMS uses comprehensive algorithms to fill these gaps and aggregates the data into 5-minute increments. In this paper, we take the 5-minute flow data of one single detector station as the dataset of the experiments. We randomly choose the detector station 500010092 as an example and use its data of weekdays in the year 2013. The imputation pattern experimented in this paper is 1-D, since one data item contains the flow data of single period and single location. Set the data period to be one day, then there are 288 time intervals in one period and 250 (the number of weekdays in the year 2013 apart from September 17) periods in total. Thus the dimension of the input and output is $q = 288$ and the number of data items is 250. We divide the whole 250 data items into training set and test set with a ratio of 3 to 2. The partially destroyed process is implemented by randomly setting some data missing according to the missing rate.

B. Criterions

To evaluate the imputation approach, we adopt three criterions to measure the error of the imputed data. They are root mean square error (RMSE),

$$RMSE = [\frac{1}{n}\sum_{i=1}^{n}(x_i^r - y_i)^2]^{\frac{1}{2}},$$

mean absolute error (MAE),

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i^r - y_i|,$$

and mean relative error (MRE),

$$MRE = \frac{1}{n}\sum_{i=1}^{n}\frac{|x_i^r - y_i|}{x_i^r},$$

where n is the total number of the missing data, y_i is the ith imputed data and x_i^r is the corresponding raw data.

C. Structure setting of DSAE

The structure of DSAE includes the number of layers, the number of nodes in each layer and the activation functions of each layer. The input layer and output layer have been decided to contain $q = 288$ nodes. A typical way of setting hidden layers for the purpose of restoring data is to choose decreasing number of nodes then increasing number of nodes in symmetry. Therefore a reasonable set of the architecture of hidden layers is three layers with 144, 72, 144 nodes respectively. We choose the sigmoid function as the activation function of each layer.

D. Results and Analysis

We divide the dataset into training set and test set. Then we conduct series of experiments with the missing rate ranging from 0.01 to 0.90 and obtain the imputation results of the test set. Apart from the deep learning approach, we also conduct experiments with artificial neural networks with the same set of layers and nodes for contrast. All the experiments are conducted on a computer with Core i5 CPU and 4G RAM. Each experiment with a certain missing rate and using a certain network costs less than 1 second. The RMSE, MAE, MRE of the imputed data under different missing rates are shown in Fig. 6, 7, 8 respectively. The RMSE of our deep learning based approach ranges from 16.9 to 20.3 veh/5-minute while the MAE ranges from 11.3 to 13.8 veh/5-minute and the MRE ranges from 0.24 to 0.35 under all the experimented missing rates. Both the RMSE and the MAE of our approach are smaller than the neural network method with most missing rates. Additionally, the error fluctuation of our approach is smaller than the neural network which can be seen obviously from the contrast of MRE.

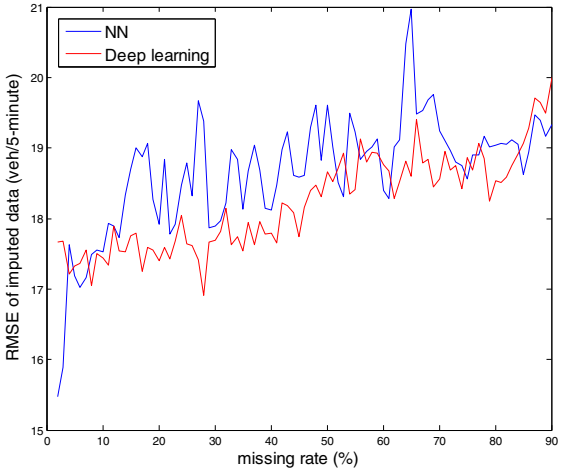

Fig. 6. The RMSE of the imputed data

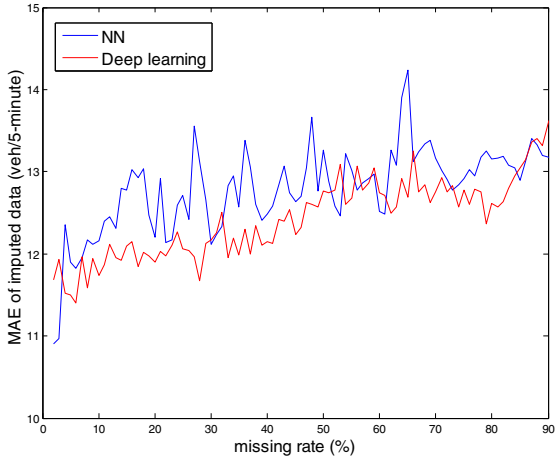

Fig. 7. The MAE of the imputed data

Fig. 9 displays the imputed data of one period with the deep learning based approach under the missing rate of 0.30. From that figure, we can see that the imputed data are quite consistent with the observed data. Considering

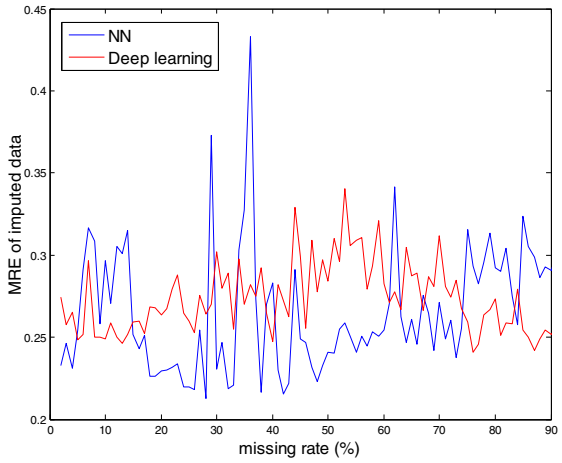

Fig. 8. The MRE of the imputed data

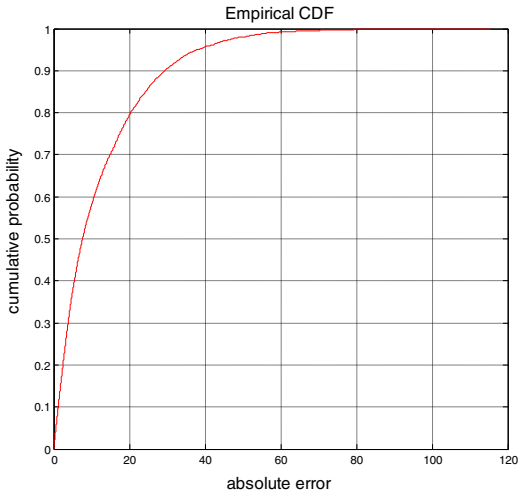

Fig. 10. The empirical cumulative distribution of the absolute imputation error

the statistical properties of the imputed data, we give the distribution of the imputation error under the missing rate of 0.3. Fig. 10 presents the empirical cumulative distribution of the absolute imputation error. 80% of the absolute error are under 20 veh/5-minute and 95% of the absolute error are under 40 veh/5-minute with the maximum flow to be 321 veh/5-minute. That illustrates our approach has a good performance.

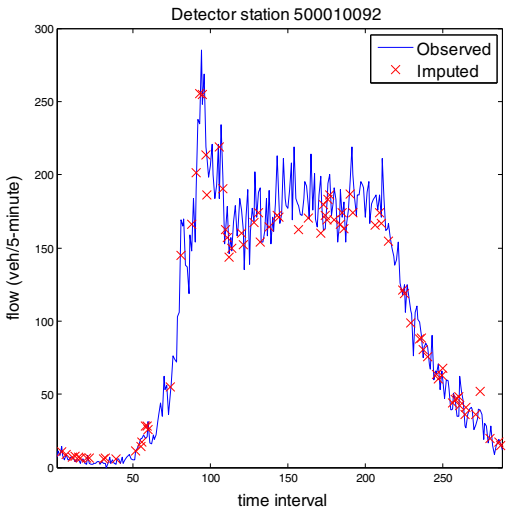

Fig. 9. The imputed traffic data of one period

V. CONCLUSIONS AND FUTURE WORKS

A. Conclusions

This paper proposes a deep learning based approach for traffic data imputation. The imputation model is constructed using a DAE filled with SAE in the middle layers. The approach treats the traffic data including observed data and missing data as a whole data item and restores the complete data with a deep structural network. The deep learning approach can discover the correlations contained in the data structure by a layer-wise pre-training and improve the imputation accuracy by conducting a fine-tuning afterwards. We conduct a series of experiments using the data collected from Caltrans PeMS to evaluate the proposed approach. The results show that our approach is fairly good. Deep learning is promising in the field of traffic data imputation.

B. Future Works

There are still many works to do about the deep learning based traffic data imputation approach. As have been described in the paper, the traffic data structures and imputation patterns can be various in real practise. This paper only tests the performance of the approach in one pattern using one type of traffic data. More experiments are expected to be done in the future. The architecture of the deep network in our approach is relatively simple. It can be more complex and powerful according to the need of applications. Large scale deep networks deserve to be investigated in the field of traffic data imputation.

VI. ACKNOWLEDGMENTS

The authors would like to thank Prof. F.-Y. Wang for his instruction and encouragement.

REFERENCES

[1] 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.

[2] B. L. Smith, W. T. Scherer, J. H. Conklin et al., "Exploring imputation techniques for missing data in transportation management systems," Initiatives in Information Technology and Geospatial Science for Transportation: Planning and Administration, Transportation Research Record 1836, pp. 132-142, Washington: Transportation Research Board Natl Research Council, 2003.

[3] Y. B. Li, Z. H. Li, and L. Li, Missing traffic data: comparison of imputation methods, Iet Intelligent Transport Systems, vol. 8, no. 1, pp. 51-57, Feb, 2014.

[4] P. Vincent, H. Larochelle, Y. Bengio et al., "Extracting and composing robust features with denoising autoencoders." in Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland, pp. 1096-1103, 2008.
[5] Y. Bengio, P. Lamblin, D. Popovici et al., Greedy layer-wise training of deep networks, Advances in neural information processing systems, vol. 19, pp. 153, 2007.
[6] N. L. Nihan, Aid to determining freeway metering rates and detecting loop errors, Journal of Transportation Engineering-Asce, vol. 123, no. 6, pp. 454-458, Nov-Dec, 1997.
[7] Z. B. Liu, S. Sharma, and S. Datla, Imputation of missing traffic data during holiday periods, Transportation Planning and Technology, vol. 31, no. 5, pp. 525-544, 2008.
[8] D. H. Ni, J. D. Leonard, and Trb, "Markov chain Monte Carlo multiple imputation using Bayesian networks for incomplete intelligent transportation systems data," Information Systems and Technology, Transportation Research Record 1935, pp. 57-67, Washington: Transportation Research Board Natl Research Council, 2005.
[9] D. H. Ni, J. D. Leonard, A. Guin et al., Multiple imputation scheme for overcoming the missing values and variability issues in ITS data, Journal of Transportation Engineering-Asce, vol. 131, no. 12, pp. 931-938, Dec, 2005.
[10] M. Zhong, S. Sharma, P. Lingras et al., "Genetically designed models for accurate imputation of missing traffic counts," Information Systems and Technology, Transportation Research Record-Series 1879, pp. 71-79, Washington: Transportation Research Board Natl Research Council, 2004.
[11] G. E. Hinton, and R. R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science, vol. 313, no. 5786, pp. 504-507, Jul, 2006.
[12] G. E. Hinton, S. Osindero, and Y. W. Teh, A fast learning algorithm for deep belief nets, Neural Computation, vol. 18, no. 7, pp. 1527-1554, Jul, 2006.