Traffic Flow Forecasting Method based on Gradient Boosting Decision Tree

XIA Ying[a], CHEN Jungang[b]

Research Center of Spatial Information System, Chongqing University of Posts and Telecommunications, Chongqing, China
[a]xiaying@cqupt.edu.cn, [b]391108961@qq.com

Keywords: Traffic Flow Forecasting, Sliding Time Window, Temporal Correlation Search, Feature Extension, Gradient Boosting Decision Tree

Abstract. Accurate traffic flow forecasting is very important for intelligent transportation system. This paper proposes a traffic flow forecasting method based on gradient boosting decision tree. In the preprocess phase, sliding time window and feature extension of traffic data are designed on the basis of time series analysis and temporal correlation search is introduced to find prediction training set. In the prediction phase, gradient boosting decision tree is used to predict the traffic flow. Experimental results show that the traffic flow forecasting method based on gradient boosting decision tree is effective and can obtain higher prediction accuracy.

Introduction

Traffic flow data has the characteristics of nonlinearity and randomness. There are many traffic flow forecasting methods are proposed based on traditional prediction models such as historical trend model[1], neural network model[2], time series model[3], Kalman filter model[4], support vector machine model[5] and so on. These methods can predict short-term traffic flow effectively, but cannot adapt well to large scale data processing.

Combining machine learning algorithm with traditional prediction model has become a trend that can be used for medium and long term prediction. The gradient boosting decision tree (GBDT) is an integrated learning method that is suitable for dealing with nonlinear data and adopted by many scholars to apply traffic flow forecasting. For example, Friedman proposed gradient boosting method[6], Tianshu Wu presented an online boosting approach for traffic flow forecasting under abnormal conditions[7], Dong Tian gave an improved AGBDT algorithm[8], Xia Li proposed GBRT model for freight vehicle travel time prediction[9], Amr Abdullatif proposed layered ensemble model for short-term traffic flow forecasting[10].

In this paper, a traffic flow forecasting method based on gradient boosting decision tree is proposed for further improvement of the prediction accuracy.

Design of Sliding Time Window

Given a time series *X(t)*, its multiple dimensional element $X(t_i)$ describes the value at time t_i. For easily description, the following definitions are given.

Time window: A time interval for traversing a fixed length of the time series *X(t)*, called a time window w of *X(t)*.

Window length: The number of time series *X(t)* falling in the time window w, called the window length of w.

Prediction window: Window for prediction of time series *X(t)*, called prediction window.

Historical series: The feature or attribute in the time window, called a historical series.

Current series: The class or label in the time window, called the current series.

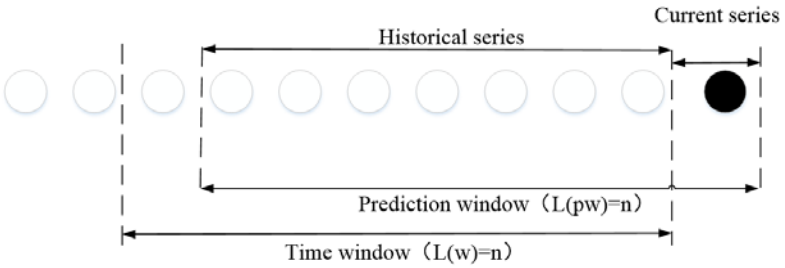

Fig.1. Sliding Time Window

Figure 1 gives the meaning of each definition. For example, according to the cyclical change of traffic flow per week, the time window length can be set as 8, that is, considering the 7-day historical series as attribute or feature, the current series as class or label. The traffic flow of each day can be counted by given time interval. We set the last predicted series as prediction window, one slide forward one day until it slides to the first day of the historical series, so that forming a sliding time window. By sliding the time window, the time series is transformed into a new data set, that is, several time windows.

Correlation Search of Sliding Time Windows

In order to prepare prediction training set, it is necessary to find the correlation between time window and prediction window. In this paper, cosine similarity is used to measure the correlation of each time window and the prediction window. For the n-dimensional space vector x and y, the cosine similarity is the cosine of the two vectors in the space angle, the similarity value is between [-1,1], the higher the value, the higher the similarity[3].

Based on the cosine similarity measuring, the top-k time windows with the highest similarity are selected as integrated learning training set for subsequent prediction.

Extension of Temporal Features

The more abundant the characteristics of traffic flow and time variation, the better the prediction accuracy. Besides the original attributes, for the integrated learning training set and prediction window, three features are constructed to expand the sample attributes, that is, time difference, time trend and time deviation.

The temporal feature extension algorithm is as follows:

Input: Dataset $D=\{x_i^1, x_i^2, x_i^3, \cdots x_i^j\}, i=1,2,\cdots N, j=1,2,\cdots N, N \in R$.

Output: Dataset $D^{'}=\left\{x_i^1, x_i^2, \cdots, x_i^j, cx_i^j, jx_i^j, fx_i^j\right\}, i=1,2,\cdots N, j=1,2,\cdots N, N \in R$.

(1) Initialization, $cx_i^j=0$, $jx_i^j=0$, $fx_i^j=0$

(2) for $i=1,2,\cdots N$ //N is the number of time series.

for $j=1,2,\cdots N-1$

$x_i^{j+1}=x_i^{j+1}$

$cx_i^j=x_i^{j+1}-x_i^j$ //reflect changes in adjacent time differences.

$jx_i^j=\dfrac{x_i^1+x_i^2+\cdots x_i^{j+1}}{j+1}$ //reflect the time trend of data set.

$fx_i^j=\dfrac{1}{j}\left(x_i^{j+1}-jx_i^j\right)^2$ //reflect the degree of time deviation.

Prediction based on Gradient Boosting Decision Tree

On the basis of the gradient boosting decision tree, several time windows with the highest similarity are selected as training set. Setting the current time series of the prediction window as the prediction

sequence, and the mean value of all prediction sequences is taken as the final forecast sequence.

We refer the prediction algorithm based on gradient boosting decision tree[6] after correlation search of sliding time windows and feature extension as C-GBDT. The prediction algorithm is as follows:

Input: training data set $D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$, $x_i\in R^n$, $y\in R$;

Loss function $L(y,f(x))$.

Output: regression tree $f_M(x)$

(1) initialization, $f_0(x)=\arg\min_c\sum_{i=1}^{N}L(y_i,c)$

(2) for $m=1,2,\cdots M$ //M is the number of regression trees

(a) for $i=1,2,\cdots N$, //N is the number of samples

$$r_{mi}=-\left[\frac{\partial L(y_i,f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{m-1}(x)}=y_i-f_{m-1}(x_i)$$ // r_{mi} is the residual

(b) fit a regression tree to r_{mi}, get the leaf node area R_{mj} of the m-th tree, $j=1,2,\cdots,J$

(c) for $j=1,2,\cdots,J$ //J is dimension numbers of x

$$c_{mj}=\arg\min_c\sum_{x_i\in R_{mj}}L(y_i,f_{m-1}(x)+c)$$

(d) update $f_m(x)=f_{m-1}(x)+\sum_{j=1}^{J}c_{mj}I(x\in R_{mj})$

(3) get the regression tree model

$$f_M(x)=\sum_{m=1}^{M}\sum_{j=1}^{J}c_{mj}I(x\in R_{mj})$$

Experiments and Result Analysis

The experimental environment is Intel (R) Core (TM) i3-3220 @ 3.3GHZ CPU, 8GB RAM, 64-bit windows 7, programming by PyCharm and Python. Experiment data is the microwave detection data of Huangke junction in Hefei of China (http://www.openits.cn/openData/710.jhtml). We select 9 detectors, 9 sections, 6 nodes, two and half month traffic flow as test data set. The traffic flow data of 2016/06/24 00:00 to 2016/08/30 23:59 are used as training data, the time interval is 1 minute, we predict the traffic flow from 2016/08/31 00:00 to 2016/08/31 23:59.

For each intersection, traffic flow is counted per 10 minutes. The length of the sliding window is 8, and the most relevant 15 time windows are determined according to the correlation search. To predict traffic flow of August 31, time windows are selected as current time series include 8/3-8/10, 7/13-7/20 and 8/10-8/17.

The mean absolute percentage error (MAPE), mean absolute error (MAE) and mean square error (MSE) are used to evaluate the prediction performance. The smaller the value is, the more accurate the prediction results are.

$$MAPE=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i-y_i'}{y_i}\right|,\quad MAE=\frac{1}{n}\sum_{i=1}^{n}|y_i-y_i'|,\quad MSE=\frac{1}{n}\sum_{i=1}^{n}(y_i-y_i')$$

In which, y_i is actual value, y_i' is prediction value, n is sample numbers.

In order to verify the performance of C-GBDT, we compare it with basic GBDT model and validate them by prediction window. As shown in Table 1, for traffic flow prediction from 2016/08/31 00:00 to 2016/08/31 23:59, C-GBDT model which combines correlation search of sliding time windows and feature extension gains higher accuracy of traffic flow forecasting.

Table 1. Comparison of Prediction Accuracy

	MAPE	MAE	MSE
GBDT	0.097498214	7.260624354	187.5345606
C-GBD	0.083437162	5.918425477	137.7322268

Summary

Accurate traffic flow prediction is very important for intelligent transportation system. In this paper, based on the analysis of traditional short-term traffic flow forecasting model, considering ensemble learning method, a gradient boosting decision tree model combines correlation search of sliding time windows and feature extension is proposed and it can improve the prediction accuracy effectively.

References:

[1] Y.Tang, et al: Application of Improved Time Series Model in Short - term Traffic Flow Forecast of Expressway. Computer Application Research, (2015). 32(1): p. 146-149.

[2] H.Huang,T.Tang: Short-term Traffic Flow Forecasting Based on ARIMA-ANN. IEEE International Conference on Control and Automation. (2007).

[3] Y.Zhao,Y.Chen and W.Guan: ETC Short - term Traffic Flow Forecasting Model Based on Multidimensional Time Series. Journal of Transportation Systems Engineering and Information, (2016). 16(4): p. 191-198.

[4] Ms.Hang, Xg. Yang and Gx. Peng: Dynamic Prediction of High - speed Road Stroke Based on Kalman Filter. Journal of Tongji University Natural Science Edition, (2002). 30(9): p. 1068-1072.

[5] G.Fu,Gq.Han,F.Lu and Zx.Xu: Short-term Traffic Flow Forecasting Model Based on Support Vector Machine Regression. Journal of South China University of Technology (Natural Science Edition), (2013). 41(9): p. 71-76.

[6] Friedman,J.H: Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics, (2001). 29(5): p. 1189-1232.

[7] T.Wu,et al: A online boosting approach for traffic flow forecasting under abnormal conditions. The 9th International Conference on Fuzzy Systems and Knowledge Discovery. (2012).

[8] D.Tian,et al: An accurate eye pupil localization approach based on adaptive gradient boosting decision tree. Visual Communications and Image Processing (VCIP). (2016).

[9] X.Li and R.Bai: Freight Vehicle Travel Time Prediction Using Gradient Boosting Regression Tree. The 15th International Conference on Machine Learning and Applications (ICMLA). (2016).

[10] A.Abdullatif, et al :Layered ensemble model for short-term traffic flow forecasting with outlier detection. The 2nd International Forum on Research and Technologies for Society and Industry Leveraging a better tomorrow (RTSI). (2016).