Data Mining to Improve Traffic Forecasting by Recognizing Anomalies

James Howard
Colorado School of Mines
1500 Illinois St
Golden, Colorado 80401
jahoward@mines.edu

William Hoff
Colorado School of Mines
1500 Illinois St
Golden, Colrado 80401
whoff@mines.edu

ABSTRACT

Accurate traffic forecasting is of great interest for commercial, security, and efficiency applications. By traffic, we mean the movement of vehicles along a road network, the movement of people in a building, or similar data derived from the actions of a group of people. Traditional forecasting methods use statistical models learned from historical data. However, the accuracy of these models fail during the presence of anomalies. In such cases, the forecast can deviate significantly from historical averages. If anomalies can be observed multiple times, however, then a system can be trained to recognize the anomaly when it is happening and generate a more accurate forecast.

In this paper, we present a system that automatically discovers anomalies in time series data, and can recognize their occurrence in subsequent data. We then incorporate these modeled anomalies with common forecasting models to significantly improve short-term forecasting accuracy. We demonstrate improved short term forecasting accuracy on three datasets: a publicly available vehicle traffic dataset, and two building occupancy datasets, derived from a sensor network of motion sensors.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

General Terms

Traffic Forecasting

Keywords

ACM proceedings

1. INTRODUCTION

Human controlled traffic systems are everywhere from the occupancy and motion of people throughout a building to the speed and load of vehicles on a freeway. With the worlds traffic systems becoming connected, the possibility to dynamically control elements with in that system is at hand. DISCUSS BRIEFLY THE NEED TO FORECAST OCCUPANCY

What effects can accurate forecasting algorithms have on a system? According to the United States Department of Transportation, optimal timing of traffic lights on major roadways across the United States could account for approximately a 22% reduction in emissions along with a 10% reduction in fuel consumption [**?**]. Similarly, the United Stated Department of Energy [**?**] estimates that energy for heating and cooling accounts for approximately 35-45% of a building's total energy expenditure. Fully automated building control systems which optimize based on building usage and occupancy can greatly reduce this energy usage.

EMPHASIZE AGAIN THE NEED TO FORECAST ACCURATELY. PERHAPS DISCUSS THE CURRENT ALGORITHM ACCURACY AND JUSTIFY THE NEED FOR ACCURACY DURING THE TIMES THAT THE SYSTEM IS PERFORMING AT ITS "WORST."

OUTLINE FOR THE REST OF THE PAPER.

2. RELATED WORK

Work related to traffic forecasting Work on building occupancy forecasting Work on Anomaly Detection for time series. Work on Activity recognition

3. NOTATION AND METRICS

3.1 Notation

We define a time series dataset used within as $\{x_t^{(m)}\}$. Each x_t^m is an aggregate of the readings from sensor m reading at time block t. In total the number of time blocks are represented by N.

Forecasts for a given model k from the set of all models K are represented by

$$y_{t+1}^{k,m} = f(x_t, ..., x_1; \theta_k). \qquad (1)$$

Thus the forecast of x_{t+1} is a function of all past data and some trained parameterization θ_k for that model. For this work we forecast a model for each individual sensor and for convenience often drop the m and k from our forecast notation. Also, in this work we need to forecast more than one time step into the future. Future forecasts are performed through iterative one step ahead forecasts. An example of a

Knowledge Discovery and Data Mining 2014, New York, New York USA

forecast two time steps ahead of current time t is given by

$$y_{t+2} = f(y_{t+1}, x_t, ..., x_1; \theta_k). \tag{2}$$

Such a forecast is simply the forecast for one time step into the future but now with the forecasted value of y_{t+1} used as the most recent datapoint to forecast y_{t+2}. Forecasting in this nature allows for forecasts any number of time steps into the future.

Another useful time series used in this work is the residual dataset defined as

$$r_{t,\delta} = x_{t+\delta} - y_{t+\delta}. \tag{3}$$

This set of residuals is the difference between the raw data and a forecasting function operating δ time steps into the future.

3.2 Metrics of Forecasting Accuracy

We use mean absolute scale error (MASE) [**?**] and root mean squared error (RMSE) as cost functions to compare with other previously implemented techniques. These forecasting cost functions are commonly used in forecasting literature and provide a baseline by which to compare our novel forecasting approach to classic forecasting models.

RMSE
One of the most common error functions used to determine the quality of a set of forecasts. It measures the average difference between two time series. In this case, the input series x and the forecast series y.

$$RMSE_\delta = \sqrt{\frac{\sum_{t=1}^{N-\delta}(x_{t+\delta} - y_{t+\delta})^2}{N-\delta}} \tag{4}$$

MASE
Mean absolute scaled error was developed to be a generally applicable measurement of forecast accuracy. The metric is especially useful for datasets with intermittent demand unlike the common closely related measurement of mean absolute percentage error (MAPE). The reason it works well with intermittent time is that it will not return undefined or infinite values unless all dataset is equal [**?**].

$$MASE_\delta = \frac{\sum_{t=1}^{N-\delta} r_{t,\delta}}{\frac{N}{N-1}\sum_{i=2}^{N}|y_i - y_{i-1}|} \tag{5}$$

3.3 SQEONAN

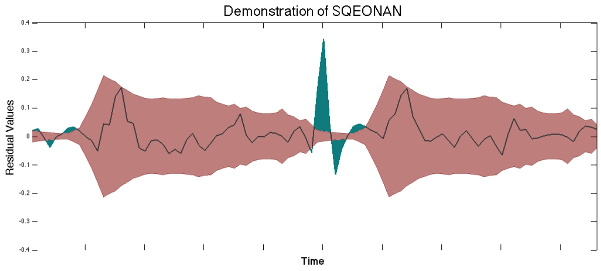

Figure 1: A demonstration of the SQEONAN metric. The sum of the area of all solid teal regions make up the metric.

TODO WE MAY HAVE TO TALK ABOUT σ being a vector and not a scalar

Due to our focus on removing the effects of anomalous events on our forecaster, we can not rely solely on traditional forecasting techniques. Anomalous events will occur infrequently and thus measures which rely on overall forecasting accuracy will likely not demonstrate much improvement if only the effects of anomalies are removed from our dataset. We will demonstrate this further at a later time.

To better measure the effect of our approach, we introduce a new measure of forecasting accuracy which we call Sum of Squared Error Outside of Noise Against a Naive Forecaster. SQENAN measures a forecast's accuracy during its worst case scenarios. This is performed by measuring the sum of squared errors of forecasts outside a prescribed boundary. We compute this boundary from the forecasting accuracy of a naive forecaster. For our work, we consider the naive forecaster to be the historical average of the data for a given time.

The results of this work deal with boundaries set by one and three standard deviations of the residual dataset formed but he historic average (naive) model. SQEONAN

$$SQEONAN_{\delta,\sigma} = \sum_{t=1}^{N-\delta} A(r_{t,\delta}; k\sigma)^2. \tag{6}$$

Where $A(r_\delta; \sigma)$ is

$$A(r_{t,\delta}; \sigma_t) = \begin{cases} r_{t,\delta} & \text{If } r_{t,\delta} \geq \sigma_t \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

σ_t is the standard deviation of the data at that time step for that given day. For certain comparisons, it is useful to use $k\sigma$ with values of $k > 1$.

TODO: Make a note how this is just a sum and thus different length datasets will have different SQEONAN values. Thus it is difficult to compare performance across datasets. Instead we should just use this metric to compare the results of various techniques to a given dataset.

An example demonstrating the regions summed by SQEONAN is given by 1. In this image, only the area of the teal regions are summed towards the SQEONAN metric. All other regions are zero. The large salmon colored region is the area that is one standard deviation for all days at that time for the residual dataset.

4. EXISTING ALGORITHMS

5. CLUSTER IDENTIFICATION

6. ALGORITHM

7. DATASETS

7.1 Freeway Traffic

7.2 Building Traffic

8. RESULTS

9. CONCLUSIONS

10. ACKNOWLEDGMENTS

We acknowledge the support of the National Science Foundation (grant CNS-0931748) as well as Northrop Grumman Corp and Lockheed Martin Corp.

APPENDIX