Evaluation of Forecast Scheme Performances Based on Statistical Error Measurements

Ebrahim Kittani†, Mothana†, Jamal Raiyn‡
Department of Computer science
Al-Qasemi Academic college of Education, ISRAEL

Abstract

This paper will review various forecast schemes. Some of the forecast schemes are based on data mining. Forecast schemes are introduced to reduce urban traffic congestion and to manage the travel information. The underlying reason is that the capacity of transportation traffic system is regularly exceeded by traffic demand. We will evaluate and compare forecast schemes performance based on statistical error measurements. There are three basic strategies to relieve congestion; the first strategy is to increase the transportation infrastructure. However this strategy is very expensive and can only be accomplished in the long term. The second strategy is to limit the traffic demand or make traveling more expensive, which will be strongly disapproved by travelers. The third strategy is focus on efficient and intelligent utilization of the existing transportation infrastructure. This strategy is a best trade-off and gains more and more attention. Currently, the Intelligent Transportation System (ITS) is the most promising approach to implementation of the third strategy. Various forecast schemes have been introduced to manage the travel information like clustering, classification schemes. Some of the most often used methods to traffic forecast are limited to expressways during daytime hours. Meanwhile the robustness and accuracy of exponential smoothing forecasting has led to its widespread use in applications where a large number of series necessitates an automated procedure, such as inventory control.

1. Introduction

Conceptually, traffic information [3] may fall into one of the three categories as follows; Historical information [11], real-time information, and Predictive information. The historical data is a collection of past observations of the system. Historical data describe the traffic states of a transportation system during previous time periods. It is mainly used to classify daily graphs or special events. Real-time information is most up-to-date and can be calculated, e.g., by on-line simulations. Predictive information, like traffic forecasts, can help to change the travel behavior of road users by providing information about the future state of the network. The real-time information achieved to update the historical adaptive information, special in the case that the real-time information does not matching the historical information. The historical information is necessary to carry out a plausible forecast in real-time [18]. To develop a robust forecast model, it is needed to collect accurate travel information; firstly it is necessary to optimize the resource allocation in cellular systems. The optimization of the resource allocation in the cellular system consider various issues like repeated handoff, radio spectrum coverage, call blocking probabilities, delay and interference. Our study and analysis has been shown that the mentioned issues are strongly influence the quality of the collected travel flow information. Nowadays the phenomenon of traffic congestion has become predominant due to the rapid increase in the number of vehicles. Traffic congestion appears when too many vehicles attempt to use a common transportation infrastructure with limited capacity. Traffic congestion appears when too many vehicles attempt to use a common transportation infrastructure with limited capacity. To reduce the traffic congestion several methods have been proposed. General, the traffic information collected based on sensors or cellular systems. Conceptually, traffic information [5] may fall into one of the three categories as illustrated in Figure 1:

- Historical Information
2. Information collection strategies

There are two major strategies for the travel data collections; at the present time, the most and widely used technology is the traditional strategy that is based on the magnetic loop detectors, installed under the roadway surface as illustrates figure 2. The lack of this technology expressed in the high cost of the installation and maintenance of the local detectors, they are typically installed only on a relatively small area of the roadway system, thus providing limited coverage of the entire transportation network. Furthermore in an urban environment there are many traffic interruptions. These interruptions cause delays that are not easily depicted by measuring speeds at any point along the road. Measuring travel times has traditionally relied recognition cameras capturing the progress of vehicles travelling along a pre-defined route as illustrates figure 3. Such systems also have the benefit of being able to count passing traffic and have become a vital tool in dealing with congestion and pollution in many cities. Such cameras are relatively costly to install and maintain. Now, with the rise of the mobile phone, in particular the smartphone, hands-free kits in cars and tablet computers, it is becoming increasingly possible to monitor traffic flows using Bluetooth and Wi-Fi signals. While not all vehicles contain mobile phones emitting Bluetooth or Wi-Fi signals, the proportion that do is now high enough that, according to system suppliers, meaningful travel time data can be obtained by tracking signals from such devices. To avoid equipment costs along the road network, we introduce a modern strategy that is based on the cellular phone service. The Information collection based on cellular systems can be gathered on millisecond compared to the traffic data collection based on detectors [19].
of the vehicle in which the phone is located. If a large enough sample is obtained, estimates of average travel times on the roadway segment being monitored could then be developed.

The cellular concept [2, 7, 14] is a mobile network architecture composed ideally of hexagonal cells as illustrates figure 4. The cells represent geographic areas. Inside the coverage area, the users, called mobile stations (MS) are able to communicate with the network while moving inside the cellular network. Each cell has a base station (BS), which serves the mobile stations. Base stations are linked to a mobile switching centre (MSC) also called mobile telephone switching office (MTSO) responsible for controlling the calls and acting as a gateway to other networks. The BS allocates network resources for the users within its cell for the communications to take place. When an active user (i.e. a mobile station using a frequency channel) reaches the boundary of the cell, it needs to change its current frequency channel for another belonging to the neighboring cell. This network procedure is known as handoff or handover [6] as shown in Figure 5.

\[ SNR_A^{tgt} > P_r > SNR_B^{tgt} \] (1)

Where \( P_r \) is the received power signal at the MH location, in other words the handoff strategy is designed as follows: At first, we track the change of the SNR of current channel. When the SNR is above threshold \( SNR^{tg} \), handoff will not occur. If the SNR is below threshold \( SNR^{tg} \), we will begin handoff execution immediately. Otherwise, when the SNR is between the two thresholds, we will use our handoff initiation algorithm to make a decision.

As illustrates Figure 6, the handoff carried out when the mobile host (MH) receives a weak SNR from BS-A and at the same time a strong signal from the neighbor BS, i.e.

---

**Figure 4: Information collection based cellular systems**

---

**Figure 5: Handoff process**

---

**2. Forecast methods**

In this section we introduce prediction methods [1][5][9] based on local prediction strategy [8]. The local prediction strategy uses local information only. Figure 7 illustrates different prediction methods e.g. mathematical, distance, logic, and modern heuristic.
2.1 Regression Analysis Method
Most of the conventional forecasting techniques may be classified as regression based. This method necessitates the identification of relevant variables that are strongly correlated to travel time, such as flow, occupancy, travel time in the neighboring links, etc. Kwon et al. (2000) presented an approach to predict travel time using linear regression with a stepwise variable selection method using flow and occupancy data from loop detectors and historical travel time information. Application of a linear regression model for short-term travel time prediction was also discussed by Rice and van Zwet (2002) [17] and Zhang and Rice (2003). They proposed a method to predict freeway travel time using a linear model in which the coefficients vary as smooth function of the departure time.

2.1.1 Regression tree model
Regression tree model [13] is constructed through binary recursive partitioning by which the data are consecutively split along the explanatory variables. Each explanatory variable is evaluated sequentially, and the variable which results in the largest decrease of the deviance in the response variable is selected. Deviance is calculated based on a threshold value in the explanatory variable and this threshold value generates two mean values for the response variable: one mean above the threshold and the other below the threshold. Splitting continues until no further reduction in deviance can be obtained or the data points are too sparse. The data set used to split to construct the regression tree model is called test data, while the data used to feed in the regression tree model for prediction purpose is called validation data.

2.2 Classification pattern
Wild Dieter (1997) [15] has used prediction method based on classification of historical patterns. It classified the historical pattern on a day groups and it takes into consideration the whole available environmental conditions (such as weather condition, events at local places like fair-grounds, gymnasiums, sports fields, theatre) and pattern matching criterion as illustrated figure 8.

2.3 Clustering
Fuzzy C-means (FCM) [10] is a data clustering technique wherein each data point belongs to a cluster to some degree that is specified by a membership grade. This technique was originally introduced by Jim Bezdek in 1981 as an improvement on earlier clustering methods. It provides a method that shows how to group data points that populate some multidimensional space into a specific number of different clusters. Lee H.S. et al., has generated patterns for the high, middle and low speed level for calculating the link speed using FCM. The final link speed is calculated by means of smoothing using the center value of each cluster after
confirming of cluster membership of collected data. Figure 9 shows structure of system for calculating the link speed. The data collection interval is of 5 minute durations.

Therefore, in this research has been proposed a novel speed calculation method for getting speed near to the actual speed. The data is classified into same time intervals of 5 minute durations [12]. And the classified data is used to calculate three pattern (high, middle, low speed) using fuzzy c-mean.

2.4 Moving Average

There are three main types of Moving Average Forecast Model [20]. The moving average is known as “smoothing” process, is an alternative way to summarize the past data. The “smoothing” (i.e., some form of averaging) process is continued by advancing one period and calculating the next average of three numbers, dropping the first number. The general expression for the moving average in simple form is

$$ EMA = (P \times \alpha) + (\text{Previous EMA} \times (1 - \alpha)) $$

$$ P = \text{Current Price} $$

$$ \alpha = \text{Smoothing Factor} = \frac{2}{1 + N} $$

$$ N = \text{Number of Time Periods} $$

OR  $$ n = (2 - \alpha)/\alpha $$

When using the formula to calculate the first point of the EMA, you may notice that there is no value available to use as the previous EMA. This small problem can be solved by starting the calculation with a simple moving average and continuing on with the above formula from there.

Notice: For most business data an Alpha parameter smaller than 0.40 is often effective. However, one may perform a grid search of the parameter space, with $\alpha = 0.1$ to $\alpha = 0.9$, with increments of 0.1. Then the best alpha has the smallest Mean Absolute Error (MA Error).

3. Methodology

Figure 6 describes the forecast scheme process. Travel time information provided into Matlab data structure. The data structure has been divided in training set and test set. We then began a search through the data set to identify cells whose data appeared clean and to eliminate cells with problems.

The performance of the forecast scheme will be evaluated to improve the forecast scheme.
4. Evaluation of forecast performance

In this section we illustrate the performance analysis of the implemented various forecast models based on historical information. The historical information collected by cellular mobile services. We have introduced and calculated several statistical measures of forecast to study and to analyze the performance of the forecast models. The EMA forecast models based on the historical information that we have used till now presented an impressive forecast performance compared to the actual information as illustrates figure 10. General the moving average offered good results. We distinguish between three main kinds of moving average that we have discussed in previous section, simple moving average, weighted moving average and exponential moving average. A simple moving average is the unweighted mean of the previous data points in the time series. A weighted moving average is a weighted mean of the previous data points in the time series. A weighted moving average is more responsive to recent movements than a simple moving average. An exponentially weighted moving average (EMA) is an exponentially weighted mean of previous data points, the parameter $l$ of an EMA can be expressed as a proportional percentage. The above Figures compare the weight factors for an exponentially smoothed four months moving average with a simple moving average that weights every day equally, and weighted moving average. In the second year we will deal with the real-time forecasting that takes into consideration the different events like weather condition, accident, holidays and special happening.

4.1 Implementation

In this section we illustrate the simulation results and analysis of the implementation of the measured traffic data. The information of the dual magnet loop detectors will be compared to the information that is provided from cellular phone services. Various forecast schemes has been implemented in WEAK as illustrates figure 11 and figure 12. Based on the WEKA platform we have carried out analysis and comparison of different Prediction schemes. WEKA (Waikato Environment for Knowledge Analysis) is a collection of machine learning algorithms for data mining tasks. WEKA contains tools for data pre-processing, classification, regression, clustering, association rules and visualization.
4.2 Statistical Error Measurement

The forecasting performance of the various models and the measures of the predictive effectiveness was evaluated using various summary statistics:

<table>
<thead>
<tr>
<th>Statistical Measurements</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Forecast error</strong></td>
<td>( e_i = x_i - \overline{x} )</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td>( \overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} )</td>
</tr>
<tr>
<td><strong>Variance</strong></td>
<td>( S^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n} )</td>
</tr>
<tr>
<td><strong>MAE</strong></td>
<td>( \frac{\sum_{i=1}^{n}</td>
</tr>
<tr>
<td><strong>RMSE</strong></td>
<td>( \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n}} )</td>
</tr>
<tr>
<td><strong>RAE</strong></td>
<td>( \frac{\sum_{i=1}^{n} (x_{ij} - x_i)}{\sum_{i=1}^{n}</td>
</tr>
<tr>
<td><strong>RRSE</strong></td>
<td>( \sqrt{\frac{\sum_{i=1}^{n} (x_{ij} - \overline{x})^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2}} )</td>
</tr>
<tr>
<td><strong>Theil’s Coeff.</strong></td>
<td>( \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n} + \sqrt{\frac{\sum_{i=1}^{n} (\overline{x})^2}{n}}} )</td>
</tr>
</tbody>
</table>

Table 5: Statistical measurements Error
4.3 Numerical Analysis

Based on the WEKA platform we have carry out analysis and comparison of different Prediction schemes as illustrates Table 6. WEKA (Waikato Environment for Knowledge Analysis) is a collection of machine learning algorithms for data mining tasks. WEKA contains tools for data preprocessing, classification, regression, clustering, association rules and visualization. We have used the Weka to make comparison between the following schemes:

i. Smoothed linear models
ii. Tree Decision Stump
iii. Rule Decision Table
iv. IBk= k-nearest- neighbor classifier
v. KStar-Nearest neighbor with generalized distance function

By review the prediction schemes for historical information, we have implemented and studying various cluster and classification schemes based on WEKA platform. Based on our analysis we found out that these schemes offer a little resolution compared to statistical schemes. Furthermore the statistical methods are more suitable for short term forecasting. In the next phase we will introduce a novel scheme that mixed between statistical and data mining algorithm (cognitive scheme).

<table>
<thead>
<tr>
<th>Forecast Schemes in Comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoothed linear models</td>
</tr>
<tr>
<td>Correlation coefficient</td>
</tr>
<tr>
<td>Relative absolute error</td>
</tr>
<tr>
<td>Root relative squared error</td>
</tr>
</tbody>
</table>

Figure 13: SMA, WMA, EMA in Comparison

Figure 14: Theil’s Coefficient
Conclusion
Various prediction schemes have been proposed to manage the travel traffic flow. In order to select the fit prediction scheme we have carried out analysis and comparison between different forecast schemes. In this paper we have introduced various forecast schemes based on the historical data. Furthermore in this paper we discuss and summarize some prediction methods based on their performance analysis. We conclude that the exponential moving average is the most accurate method, when few data are available. To handle the growth of massively multi-users of UE’s, self-optimized cognitive radio resource allocation scheme based on environment sensing is necessary.

Reference


http://www.springerlink.com/content/733270344I46


