2011 14th International IEEE Conference on
Intelligent Transportation Systems
Washington, DC, USA. October 5-7, 2011

Traffic Speed Forecasting by Mixture of Experts

Vladimir Coric, Zhuang Wang, Slobodan Vucetic

Abstract— **Traffic speed is one of the most important quantities for travel information systems. Accurate speed forecasting can help in trip planning by allowing travelers to avoid the congested routes, either by choosing the alternative routes or by changing the departure time. It is also helpful for traffic monitoring, control, and planning. An important feature of traffic is that it consists of free flow and congested regimes, which have significantly different properties. Training a single traffic speed predictor for both regimes typically results in suboptimal accuracy. To address this problem, a mixture of experts algorithm which consists of two regime-specific linear predictors and a decision tree gating function was developed. A generalized expectation maximization algorithm was used to train the linear predictors and the decision tree. The proposed algorithm was evaluated on a 5-mile stretch of I35 highway in Minneapolis containing 10 single loop detector stations, with prediction horizons ranging from 5 minutes to one hour ahead. Experimental results showed that mixture of experts approach outperforms several popular benchmark approaches.**

I. Introduction

DESPITE the significant investments over the last few decades to enhance and improve road infrastructure worldwide, the capacity of road networks has not kept pace with the ever increasing growth in demand. As a result, congestion has become endemic to many highways and city streets. As an alternative to costly and sometimes infeasible construction of new roads, transportation departments are increasingly looking at ways to improve traffic flow over the existing infrastructure. The biggest challenge in accomplishing this goal is the ability to monitor traffic, estimate its current state, and forecast its future behavior. Having this ability, more efficient strategies for real-time traffic control and management could be developed. Moreover, informing travelers about the current and future traffic can motivate them to modify travel plans during congested periods and, in doing so, relieve the congestion.

Due to importance of traffic forecasting, transportation research community developed and evaluated numerous statistical and machine learning methods for predicting traffic quantities such as traffic volume (number of cars per hour), speed, and travel time. In this paper, we focus on short-term traffic forecasting [10], which refers to prediction horizons ranging from few minutes to an hour ahead. The baseline approaches for short-term traffic forecasting are random walk and historical predictions. The random walk method predicts the future traffic to be equal to the current traffic and is quite accurate for prediction horizons of up to few minutes ahead. The historical prediction uses the average historical traffic under the same conditions, such as location, day of the week, and time of the day, and is quite competitive for time horizons of over one hour ahead.

As a generalization of historical prediction, *k*-nearest neighbor method has been used [12] to search for the most similar traffic patterns in historical data to the current traffic state. Parametric models that include linear regression [5,7,16] and autoregressive integrating moving average (ARIMA) [8,9] have also been popular thanks to their simplicity and reasonable accuracy. Machine learning models such as neural networks [13] and support vector machines [15] have also been used with success, indicating that the nonlinearities in traffic behavior could be exploited.

Traffic flow can often be characterized as being in one of the several regimes. Typical two regimes are *free flow*, in which density of cars is low enough to allow uninterrupted flow of traffic near the speed limit, and *congestion*, in which high density of cars causes a significant drop in traffic speed. Learning a single predictor on data with regimes could require powerful nonlinear methods and result in an overly complicated model prone to overfitting [11]. As an alternative, it might be more reasonable to train simpler predictors on each regime separately. This idea can be implemented through the mixture of experts framework [6] which is illustrated in Figure 1. The figure depicts two *experts* that produce predictions y_1 and y_2 for the given input x and a *gating function* that decides how much to trust each expert at any given time. The objective of this paper is to propose a new mixture of experts architecture that is appropriate for traffic forecasting.

There are two major classes of mixture of experts models for time series forecasting, depending on the functional form of gating function. In first, inputs to the gating function are the same as inputs to the experts. If, in addition, both the experts and the gating function are feedforward neural networks, the whole mixture of experts model can be represented as a feedforward neural network [14]. In second, the gating function is a Markov chain that models transitions between regimes probabilistically, and is leads to the regime switching model that has been popular in time series analysis [3]. The problem with the first approach is that it can still

Manuscript received May 15, 2011. This work was funded in part by the National Science Foundation Grant IIS-0546155.

V. Coric and S. Vucetic are with the Temple University, Philadelphia, PA 19122 USA (e-mails: vladimir.coric@temple.edu, vucetic@temple.edu).

Z. Wang is with the Siemens Corporate Research, Princeton, NJ 08540 USA (e-mail: zhuang.wang@siemens.com).

result in an overly complex model that is prone to overfitting and difficult to interpret. The problem with the second approach is that Markov chain can be too slow to adapt to changing traffic conditions. In our proposed design experts are linear regression models and gating function is a decision tree. Such design alleviates the overfitting problem, retains modeling flexibility, and allows easy interpretation of the resulting model.

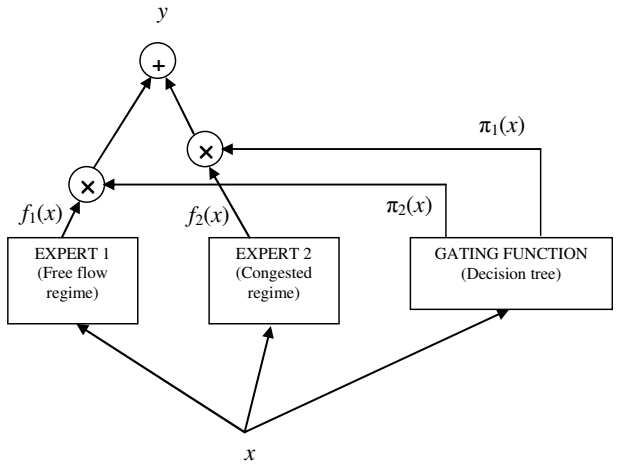

Figure 1. Mixture of experts architecture

It is worth mentioning that two related approaches were studied in the traffic forecasting literature. One accounts for regime change by detecting shifts in the process mean and updating the intercept term of an ARIMA model [1]. This approach can be treated as a simplified heuristic version of the mixture of experts approach. An approach that has been popular in traffic forecasting consists of fitting a separate linear predictor at different times of a day and ensuring the smooth transition of predictor weights during the day [16]. This approach can be treated as an extreme version of mixture of experts model where there are many experts and gating function is a deterministic function of the time of day.

In this paper, we consider traffic speed forecasting and we evaluate the proposed and benchmark predictors on real-life traffic data from a highway segment in Minneapolis, MN. It should be noted that our approach can also be applied to other traffic forecasting problems such as prediction of traffic volume and travel time and also to similar time series forecasting problems in other domains.

II. METHODOLOGY

In this section we describe the proposed mixture of experts approach for speed forecasting. Let us denote with x_i a set of attributes at time t_i and with y_i the target variable representing traffic speed at time $t_i+\delta$, where δ is the forecasting horizon. We assume that target variable is generated as,

$$y_i = h_k(x_i) + \varepsilon_i, \quad (1)$$

where h_k is the unknown regression function for k-th regime and ε_i is the target noise. If the noise is Gaussian, $\varepsilon_i \sim N(0, \sigma_k^2)$, the probability of target y_i given x_i can be written as

$$p_k(y_i \mid x_i) = \mathrm{N}(y_i \mid h_k(x_i), \sigma_k^2). \quad (2)$$

Using the mixture model the target probability is

$$p(y_i \mid x_i) = \sum_{k=1}^{K} p(y_i \mid regime_k, x_i) = \sum_{k=1}^{K} \pi_{ik} p_k(y_i \mid x_i), \quad (3)$$

where $\pi_{ik} = p(regime_k \mid x_i)$ is the *prior probability* that i-th example is generated by the k-th regime and $p_k(y_i \mid x_i)$ is the target probability when the i-th example is generated by the k-th regime. The target of i-th example can be calculated as the expected value of the mixture model,

$$\hat{y}(x_i) = E[y_i \mid x_i] = \sum_{k=1}^{K} \pi_{ik} E_k[y_i \mid \mathbf{x}_i]. \quad (4)$$

The learning problem is to determine π_{ik} and $p_k(y_i \mid x_i)$. We assume that they are parametric functions expressed as $\pi_{ik}(\boldsymbol{\theta}_g)$ and $p_k(y_i \mid x_i, \boldsymbol{\theta}_p)$. The mixture model from (3) can now be rewritten as

$$p(y_i \mid x_i, \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_{ik}(\boldsymbol{\theta}_g) p_k(y_i \mid x_i, \boldsymbol{\theta}_p), \quad (5)$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_g, \boldsymbol{\theta}_p)$ are the model parameters.

To facilitate model optimization, we consider the regime assignment as the unobserved data and introduce a latent binary indicator variable z_{ik}, where

$$z_{ik} = \begin{cases} 1, & \text{if } x_i \in \text{regime}_k \\ 0, & \text{otherwise,} \end{cases} \quad (6)$$

Denoting $\mathbf{z}_i = [z_{i1} \ldots z_{iK}]$, the complete probability for x_i is

$$p(y_i, \mathbf{z}_i \mid x_i, \boldsymbol{\theta}) = \prod_{k=1}^{K} (p(y_i, z_{ik} = 1 \mid x_i, \boldsymbol{\theta}))^{z_{ik}}. \quad (7)$$

By denoting $X = \{x_i, i = 1 \ldots N\}$, $Y = \{y_i, i = 1 \ldots N\}$ and $Z = \{\mathbf{z}_i, i = 1 \ldots N\}$, the log-likelihood of the complete data $D_{complete} = \{(x_i, y_i, \mathbf{z}_i), i = 1 \ldots N\}$ is given by

$$\ln p(Y, Z \mid X, \boldsymbol{\theta}) = \sum_{i=1}^{N} \sum_{k=1}^{K} z_{ik} \ln(\pi_{ik}(\boldsymbol{\theta}_g) p(y_i \mid x_i, \boldsymbol{\theta}_p)) \quad (8)$$

To find $\boldsymbol{\theta}$ that increases the complete log-likelihood (8), the expectation-maximization (EM) algorithm is used. EM starts with an initial guess of $\boldsymbol{\theta}$ and updates it by alternating between expectation (E) and maximization (M) steps until convergence.

In the **E-step**, the algorithm evaluates the expected value of the log-likelihood, with respect to the current estimate of the posterior probability of Z given X and Y. By denoting the current parameter estimate as $\boldsymbol{\theta}^c$, the expectation can be expressed as

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^c) = \sum_{i=1}^{N} \sum_{k=1}^{K} \gamma_{ik}(\boldsymbol{\theta}^c)(\ln \pi_{ik}(\boldsymbol{\theta}_g) + \ln p(y_i \mid x_i, \boldsymbol{\theta}_p)) \quad (9)$$

where

$$\gamma_{ik}(\boldsymbol{\theta}^c) \equiv p(z_{ik} = 1 \mid x_i, y_i, \boldsymbol{\theta}^c) = \frac{\pi_{ik}(\boldsymbol{\theta}_g^c) p_k(y_i \mid x_i, \boldsymbol{\theta}_p^c)}{\sum_{j=1}^{K} \pi_{ij}(\boldsymbol{\theta}_g^c) p_j(y_i \mid x_i, \boldsymbol{\theta}_p^c)} \quad (10)$$

is the *posterior* that *i*-th example is from *k*-th regime.

In the **M-step**, the algorithm updates the model parameters $\boldsymbol{\theta}$ to maximize Q,

$$\boldsymbol{\theta}^{(new)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(old)}). \quad (11)$$

To optimize (11) we have to define the parametric functions $\pi_{ik}(\boldsymbol{\theta}_g)$ and $p_k(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}_p)$. This will be discussed in the following section.

A. Model optimization

Let us discuss how to solve the optimization problem (11) depending on how $\pi_{ik}(\boldsymbol{\theta}_g)$ and $p_k(y_i \mid x_i, \boldsymbol{\theta}_p)$ are defined. Starting from the assumption (2), we can define

$$\ln p_k(y_i \mid x_i, \boldsymbol{\theta}_p) = \frac{(y_i - f_k(x_i, \mathbf{w}_k))^2}{\delta_k^2} + \ln 2\pi\delta_k{}^2. \quad (12)$$

where $f_k(x_i, \mathbf{w}_k)$ is a predictor for *k*-th regime and the parameter set is $\boldsymbol{\theta}_p = \{(\mathbf{w}_k, \delta_k), k = 1...K\}$. In this work, we will use linear regression functions $f_k(x, \mathbf{w}_k) = x_i{}^T \mathbf{w}_k$.

Instead of using a parametric gating function $\pi_{ij}(\boldsymbol{\theta}_g)$ we will use decision tree. Since decision tree is nonparametric, we will use notation π_{ij} instead of $\pi_{ij}(\boldsymbol{\theta}_g)$. As a result, direct maximization (11) of Q from (9) by gradient descent approaches is not possible. Instead, we will use a generalized expectation maximization procedure that is guaranteed to increase value of Q at each M step, instead of maximizing it. We accomplish the M step in two stages.

In the first stage, for a fixed decision tree from the previous M step, we find $\boldsymbol{\theta}_p$ that maximizes Q from (9). Since optimization of $\mathbf{w}_k$ does not depend on δ_k, we first optimize $\mathbf{w}_k$ while treating δ_k as constant. For regime-specific function f_k, the resulting problem is equivalent to minimizing the weighted squared error,

$$E_k = \sum_{i=1}^{N} \gamma_{ik}(\boldsymbol{\theta}^c)(y_i - f_k(x_i, \mathbf{w}_k))^2. \quad (13)$$

To minimize E_k, by remembering that f_k is a linear regression function, we can obtain $\boldsymbol{w_k}$ in a closed-form by solving the weighted regression problem,

$$w_k = (X^T \Gamma_k X)^{-1} X^T \Gamma_k Y, \quad (14)$$

where Γ_k is a diagonal matrix with entries $\{\gamma_{ik}(\boldsymbol{\theta}^c), i = 1\ldots N\}$. Given $\mathbf{w}_k$, the remaining step is to optimize δ_k. By setting the derivative of Q with respect to δ_k to zero, the optimal δ_k is obtained in the closed-form as

$$\delta_k{}^2 = \sum_{i=1}^{N} \gamma_{ik}(\boldsymbol{\theta}^c)(y_i - f_k(\mathbf{x}_i, \mathbf{w}_k))^2 / \sum_{i=1}^{N} \gamma_{ik}(\boldsymbol{\theta}^c) \quad (15)$$

In the second stage, we train a decision tree with probabilistic outputs to predict γ_{ik} defined as

$$\gamma_{ik} = \frac{\pi_{ik}{}^c p_k(y_i \mid x_i, \boldsymbol{\theta}_p{}^c)}{\sum_{j=1}^{K} \pi_{ij}{}^c p_j(y_i \mid x_i, \boldsymbol{\theta}_p{}^c)}, \quad (16)$$

where $\pi_{ik}{}^c$ is the decision tree from the previous M step, and $\boldsymbol{\theta}_p{}^c$ are newly learned parameters from (14) and (15). The justification for the second stage comes from the first term on the right hand side of (9), which is maximized by approximating the posterior γ_{ik}. Both stages guarantee decrease in the objective function Q and, thus, the convergence of the generalized EM procedure to a local maximum.

Table 1. Outline of the proposed mixture of experts algorithm

1. Partition the data set *D* into *K* candidate regimes
2. Train one predictor f_k on each regime
3. Assign all prior values π_{ik} to be a constant
4. repeat
   a. Estimate noise variance δ_k from (15)
   b. Calculate posteriors γ_{ik} (from (10), where p_k is defined in (2))
   c. Train a decision tree to learn the prior π_{ik}
   d. Train regime predictors f_k (using (14))
5. until convergence
6. Predict using (17)

Let us briefly discuss technical details of training the decision tree with probabilistic outputs. The idea is to treat all examples assigned to *k*-th regime as class *k*. Since the assignment of examples to regimes in the EM approach is probabilistic (depending on γ_{ik} values), we sample with replacement training set of size *N* from the original training set based on the probabilities γ_{ik}. Probabilistic outputs π_{ik} are obtained simply as a fraction of all examples from class *k* in the leaf node. To improve robustness, we use the Laplace correction in estimation of π_{ik}.

Following (4) and given the trained linear experts f_k and decision tree that provides π_{ik} values, label of *i*-th example is predicted as

$$\hat{y}(x_i) = \sum_{j=1}^{K} \pi_{ik} x_i{}^T \mathbf{w}_k. \quad (17)$$

Outline of the proposed mixture of experts algorithm is summarized in Table 1.

III. Experimental Setup

A. Data description

To evaluate the proposed mixture of experts algorithm and compare it with alternatives, we used traffic data collected over a 5 mile stretch of I-35W highway in Minneapolis, MN as shown in Figure 2. This part of Minneapolis highway network is located near the city center on which congestion periods are regularly occurring during both morning and afternoon rush hours (see Figure 3). This very congested segment contains 10 traffic measurement stations in each direction, with spacing of about half a mile. Each station measures traffic at every lane by the single loop detectors that are installed right beneath the pavement. Every single loop detector reports volume (how many cars pass over the sensor) and occupancy (how long the sensor was occupied) during each 30-second interval. The data covered 3 months, from March 1 to Jun 1 of 2003, during periods between 7am and 7pm. In this study we considered only traffic measurements at the second lane from left because it is a

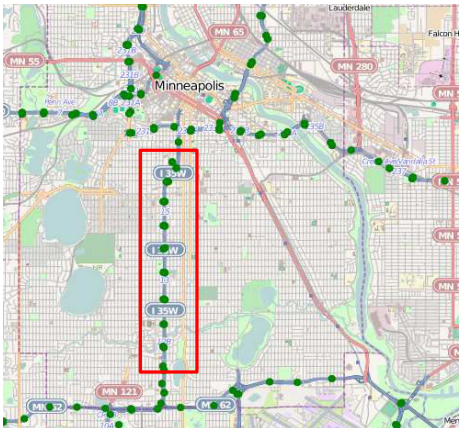

Figure 2. Highway segment in Minneapolis covered by our study

representative of typical traffic conditions.

B. Data preprocessing

For our experiments, we removed holidays and weekends because their traffic behavior is significantly different from weekdays (as seen in Fig. 3). In addition, we also removed several days with a large number of missing or corrupted measurements. For example, all measurements during Tuesday, March 18, are missing, while on Monday, March 3, estimated speed is unusually low during early morning hours. Both days were treated as outliers and were removed from data set. After removal, a total of 58 days remained in our data set.

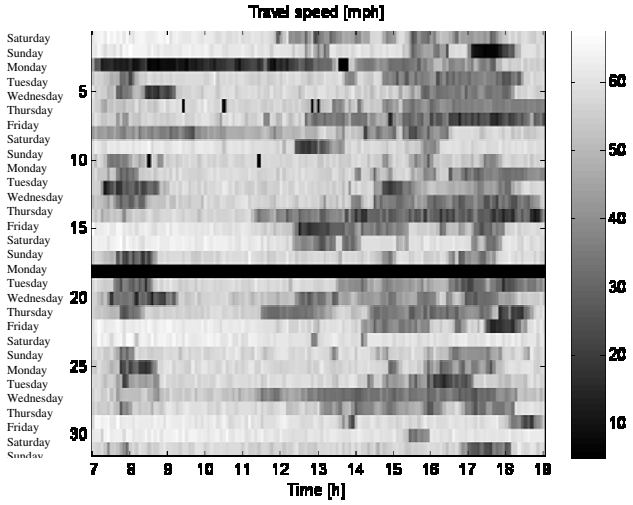

Figure 3. Traffic speed on sensor 8 during March 2003

Starting from the raw 30-second volume and occupancy data, an important step is to estimate traffic speed. This is a nontrivial problem because relationship between speed and occupancy depends on lengths of vehicles passing over the sensor. We used the following standard approach [2] to estimate speed as

$$speed = length \times \frac{volume}{occupancy}, \tag{18}$$

where *length* is the average vehicle length estimated as

$$length = 60\,mph \times median\left(\frac{occupancy}{volume}\right), \tag{19}$$

where median is calculated over all 58 days in our data only during the non-congested time intervals and assuming that (1) free flow speed is 60 miles/hour (equal to the speed limit on this part of highway) and (2) average vehicle length does not change significantly during the day. Because speed estimation using this approach can be quite unreliable, we aggregated speed to 5-minute increments.

C. Experimental setup

The objective of our evaluation was to predict traffic speed in a range from 5 minutes to one hour ahead. The evaluated mixture of experts predictor consisted of two linear regression experts and a decision tree gating function, as described in Section 2. While our approach allows using a larger number of experts, we decided to experiment with two to allow testing a hypothesis that the EM algorithm will be able to discover free flow and congested regimes. Both experts and decision tree shared the same input attributes, although our approach is general enough and allows using a customized set of attributes for each component.

Out of the 58 days in our data set, we used the first 40 as training set and the last 18 as test set. We trained a separate mixture of experts model for each of the 10 sensors and each of the time horizons ranging from 5 minutes to one hour ahead in increments of 5 minutes. Results are reported as Mean Absolute Errors (MAE). For all experiments we used the following 21 attributes: (1) Current speed (at time t) from all 10 sensors along the stretch; (2) Average historical speed at time $t+\delta$ from all 10 detectors on the stretch, where δ is the prediction horizon; (3) Current volume at sensor on which speed is predicted. For linear regression models, we also added the intercept term as the first attribute.

We compared our approach with random walk, historical average, linear regression, regression tree, time varying coefficient regression (TVC), and Markov switching model.

1. (**RW**) Random walk method predicts traffic speed by using current speed as prediction.
2. (**HIS**) Historical average method calculates average historical speed at given time of the day and uses it for prediction.
3. (**LR**) Standard linear regression predictor is the special case of the mixture of experts approach with a single regime.
4. (**RT**) Regression tree is a standard data mining algorithm. To train the right sized tree, we used a subset of the training data for pruning.
5. (**TVC**) Linear regression with time varying coefficients [4] has been popular in traffic forecasting [16]. The main idea is that regression coefficients w can vary gradually with time (t) and prediction horizon (δ). For given t and δ, TVC is trained by minimizing

$$\sum_{d\in D}\sum_{s\in T}(y_{d,t+\delta+s} - f(x_{d,t}, w_{t,\delta}))^2 K(s) \tag{20}$$

where D are days in training data, T is length of a window centered around $(t+\delta)$, K is the kernel function that imposes smoothness. For our experiments, following [7], $K(s)$ had values 0.3, 0.6, 1, 0.6, 0.3 at $s = -2, -1, 0, 1, 2$, respectively.

6. (**MS**) Markov switching model [3] is similar to the proposed approach, the only difference being that the gating function is modeled as the Markov chain.

IV. Results

Table 2 summarizes accuracies of various predictors for different time horizons. The abbreviation **ME** is used for the proposed mixture of experts model. The MAE results shown are average MAE over the 10 I-35W sensors. We can see that random walk predictor works better than historical predictor for horizons of up to 30 minutes ahead. Interestingly, although the two predictors behave very differently, their overall performance is similar. Both baseline predictors are inferior to the remaining approaches that use current speed and historical speed as attributes. Accuracy of TVC and linear regression method is similar, with TVC being slightly less accurate for short horizons and slightly more accurate for longer horizons. Markov switching model is competitive for shorter horizons, but deteriorates for longer ones because its transition matrix cannot accurately capture regime changes over longer time periods. The proposed mixture of experts approach that uses regression tree as gating function achieves the best overall results. As compared to the Markov switching model, it is evident that regression tree is more appropriate than the Markov chain and that it is better capable of predicting the traffic regimes. It also produces more accurate results than the regression tree, which indicates that using linear experts is more beneficial than using constants at the leafs of a tree.

Table 2. Reported MAE for 12 prediction horizons

Hor.	*HIS*	*RW*	*TVC*	*LR*	*MS*	*RT*	*ME*
5	7.67	4.43	4.24	4.02	4.04	3.95	**3.76**
10	7.67	5.38	5.03	4.89	4.89	4.93	**4.69**
15	7.67	6.12	5.58	5.50	5.63	5.56	**5.33**
20	7.67	6.72	6.00	5.93	6.08	5.92	**5.70**
25	7.67	7.23	6.31	6.27	6.29	6.32	**6.09**
30	7.67	7.74	6.59	6.59	7.07	6.51	**6.30**
35	7.67	8.18	6.78	6.83	7.17	6.72	**6.59**
40	7.67	8.63	6.97	7.04	7.63	6.84	**6.76**
45	7.67	8.98	7.12	7.21	7.87	7.05	**6.91**
50	7.67	9.33	7.25	7.35	7.97	7.13	**7.07**
55	7.67	9.63	7.35	7.45	8.16	7.21	**7.15**
60	7.67	9.94	7.41	7.54	8.58	7.31	**7.24**
Total	7.67	7.69	6.39	6.39	6.78	6.29	**6.13**

In Table 3 we report overall (over all forecasting horizons) MAE for each of the 10 sensors used in our study. Sensors 4, 5, 6 and 7 have larger MAE than the first and last theree sensors. The possible explanation is that there are 3 on and 3 off ramps that are located between sensors 4 and 8 and that their influence causes larger deviations on traffic speed and that they are more difficult to predict.

To get a better insight into the forecasting performance, Figure 4 illustrates true and predicted speed 5 minutes and 1 hour ahead by the mixture of experts method. As expected, 5-minute ahead accuracy is much better. In the 1-hour ahead prediction, we can observe that accuracy during the congested regime is larger than during the free-flow regime. We can also observe a slight delay in recognizing the regime change. However, the mixture of experts approach is more successful in both than the competing predictors.

Table 3. Reported MAE for all 10 sensors on the highway stretch

Sens.	*HIS*	*RW*	*TVC*	*LR*	*MSE*	*RT*	*ME*
1	6.68	5.93	5.18	5.57	5.66	5.33	**5.18**
2	6.64	6.25	5.46	5.73	5.61	5.42	**5.27**
3	7.38	7.09	6.17	6.11	6.39	5.87	**5.77**
4	8.82	7.78	6.87	7.01	7.26	6.98	**6.79**
5	9.75	8.31	7.30	7.45	7.92	7.03	**6.86**
6	9.00	8.56	7.13	7.08	7.83	6.85	**6.79**
7	8.42	9.19	7.34	7.05	7.82	7.14	**6.99**
8	6.93	7.88	6.09	5.96	6.59	5.99	**5.82**
9	6.04	7.49	5.74	5.55	5.86	5.66	**5.47**
10	7.09	8.44	6.58	6.34	6.88	6.63	**6.39**
Total	7.67	7.69	6.39	6.39	6.78	6.29	**6.13**

In Figure 5, we analyze the resulting two speed forecasting experts on sensor 1 – one specialized for congested and another for free flow regime. Each column in the figure shows importance of each of the 22 attributes (the intercept term, 10 current speeds, 10 historical speeds, and volume) where black dots indicate that the given attribute is significant (absolute value of its t-statistics is above 3). We can notice that the interception term (the first attribute) was important in all experts, other than in the congested expert for 5 minute ahead forecasting, which used the current speed. Congestion regime expert relies on current and historical speed of downstream sensors. The free flow expert found a larger range of attributes to be useful.

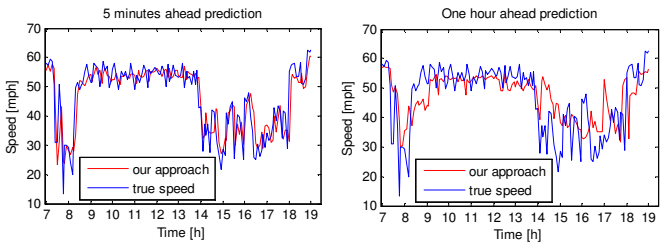

Figure 4. Predictions of mixture of experts for 5 minutes and 1 hour ahead for one of test days

In Figure 6 we show the top portion of the trained regression trees in the mixture of experts model trained for speed prediction at sensor 1. The tree for 5 minute ahead forecasting gives the highest importance to the current speed

at sensor 1 and it also uses current speed at the immediate downstream sensors. On the other hand, the tree for 1 hour ahead forecasting gives higher importance to historical speed (it is the second most important attribute) and to current speed at sensors downstream, which give better information about upcoming regime changes.

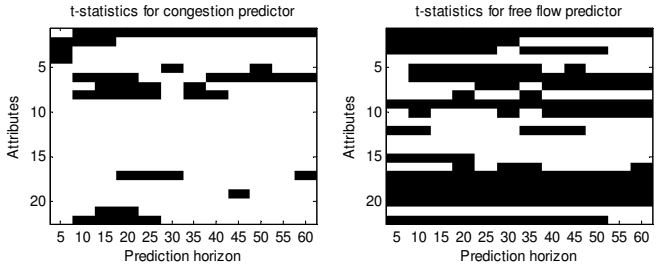

Figure 5. Comparison of t-statistics for different regime predictors

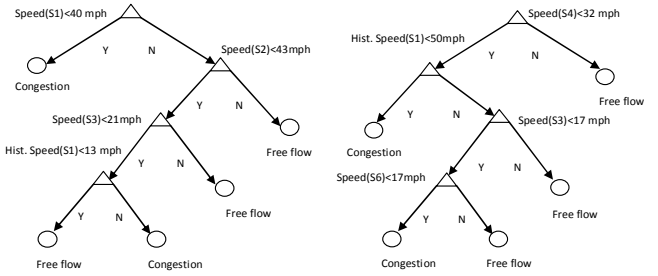

Figure 6. Top of decision trees for sensor 1 for 5 min (left) and 1-hour (right) ahead predictions

Figure 7 shows comparison between actual traffic speed (in miles/hour) at sensor 1 and prior probabilities for expert 1 given by the regression tree gating function. As can be seen, prior probability of expert 1 is tightly related with the actual speed, clearly indicating that expert 1 is specialized for the free-flow regime. This result confirms that the proposed mixture of experts model was successful in uncovering the major two traffic regimes.

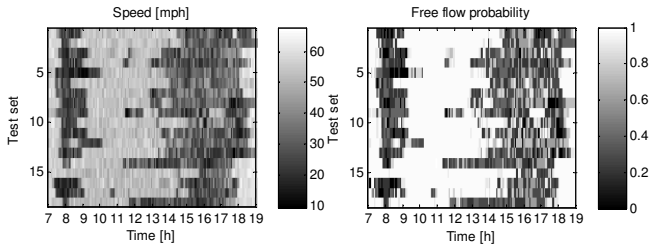

Figure 7. Comparison of actual speed and free flow prior for one of test days

V. CONCLUSION

In this paper we presented a mixture of experts approach that uses linear regression predictors as experts and regression tree as gating network. This relatively simple and computationally efficient approach managed to automatically discover free-flow and congested regimes and was more accurate than several competing algorithms, including regression trees, time varying regression, and Markov switching model. The structure of the proposed model, where both regression tree rules and linear forecasting experts can be easily interpreted by humans, can make it very attractive for traffic engineers. The proposed approach allows use of more than two experts, which could be useful in modeling of other traffic regimes, such as traffic accidents or harsh weather conditions. If human interpretation of the resulting forecasting model is not a primary objective, the proposed approach allows replacing linear predictors with more powerful neural networks.

ACKNOWLEDGMENT

The authors are grateful to Dr. Taek Kwon from University of Minnesota, Duluth, for providing access to Mn/DOT TMC data.

REFERENCES

[1] Cetin, M., Comert, G.: Short-term traffic flow prediction with regime switching models. Transport. Res. Rec. 1965, 23--31 (2006)
[2] Coifman, B.A.: Improved velocity estimation using single loop detectors, Transport. Res. A-Pol, 35, 863--880 (2001)
[3] Hamilton, J. D.: Time series analysis. Princeton University Press, Princeton (1994)
[4] Hastie, T., Tibshirani, R.: Varying coefficient models. J. R. Stat. Soc., 55, 757--796 (1993)
[5] Huang, L., Barth, M.: A novel loglinear model for freeway travel time prediction. Proceedings of the 11th International IEEE Conference on Intelligent Transportation Systems. 210--215 (2008)
[6] Jordan, M. I., Jacobs, R. A.: Hierarchical mixtures of experts and the EM algorithm. Neural. Comput. 6, 181--214 (1994)
[7] Kwon, J., Petty, K.: Travel time prediction algorithm scalable to freeway networks with many nodes with arbitrary travel routes. Trans. Res. B. 1935, 147--153 (2005)
[8] Lee, S., Fambro, D. B.: Application of subset autoregressive integrated moving average model for short-term freeway traffic volume forecasting. Transport. Res Rec. 1678, 179--188 (1999)
[9] Nihan, N.L.: Use of the Box and Jenkins time series technique in traffic forecasting. Transportation. 9, 125--143 (1980)
[10] Nikovski, D., Nishiuma, N., Goto, Y., Kumazawa, H.: Univariate short-term prediction of road travel times. Proceedings of the 8th International IEEE Conference on Intelligent Transportation Systems. 1074--1079 (2005)
[11] Pesaran, M. H., Timmermann, A.: How costly is it to ignore breaks when forecasting the direction of a time series? Int. J. Forecasting. 20, 3, 411--425 (2004)
[12] Robinson, S., Polak, J.: Modeling urban link travel time with inductive loop detector data by using the k-nn method. Trans. Res. B. 1935 47--56 (2005)
[13] Van Lint, J. W. C., Hoogendoorn, S. P., Van Zuylen, H. J.: Freeway travel time prediction with state-space neural networks modeling: State-space dynamics with recurrent neural networks. Trans. Res. Rec. 1811, 30--39 (2002)
[14] Weigend, A. S., Mangeas, A. M., Srivastava, N.: Nonlinear gated experts for time series: discovering regimes and avoiding overfitting. Int. J. Neural Syst. 6, 373--399 (1995)
[15] Wu, C. H., Wei, C. C., Su, D. C., Chang, M. H., Ho, J. M.: Travel time prediction with support vector regression. Proceedings of the 7th International IEEE Conference on Intelligent Transportation Systems. 276--281 (2004)
[16] Zhang, X., Rice, J.A.: Short-term travel time prediction using a time-varying coefficient linear model. Transport. Res. C Emer. 11, 187--210 (2003)