Discovering Patterns in Traffic Sensor Data

Farnoush Banaei-Kashani
Computer Science Department
University of Southern California
Los Angeles, CA 90089
banaeika@usc.edu

Cyrus Shahabi
Computer Science Department
University of Southern California
Los Angeles, CA 90089
shahabi@usc.edu

Bei Pan
Computer Science Department
University of Southern California
Los Angeles, CA 90089
beipan@usc.edu

ABSTRACT

We maintain a one of a kind, large-scale and high resolution (both spatially and temporally) traffic sensor dataset collected from the entire Los Angeles County road network. Traffic sensors (installed under the road pavement) are used to measure real-time traffic flows through road segments. In this paper, we exploit this dataset to rigorously verify two popular instinctive understandings about traffic flows on road segments: 1) each road segment has a typical traffic flow (known by local travelers) and one can often categorize road segments based on the similarity of their traffic flows, and 2) the road segments within each category not only have similar traffic flows but also are similar in their other characteristics (such as locality, connectivity). Toward this end, we developed a hypothesis analysis framework based on a variety of clustering and correlation evaluation techniques and leveraged this framework to respectively show the following. First, the set of road segments can indeed be partitioned into a set of distinct subpartitions with similar traffic flows, and there is a limited number of *signature* traffic patterns/labels each of which can accurately represent all traffic flows of a subpartition of the road segments. Second, all segments within each subpartition (represented by one signature) are also highly similar in three other characteristics, namely, direction, connectivity and locality. Our experiments verify our observations with high confidence.

Categories and Subject Descriptors

I.1 [**Computing Methodologies**]: Expressions and Their Representation; I.6 [**Simulation and Modeling**]: Miscellaneous

General Terms

Algorithms

Keywords

Patterns, Temporal and Non-Temporal Road Network Features, Feature Subset Selection

IWGS 2011 Chicago, Illinois, USA

1. INTRODUCTION

With accurate characterization of traffic flows (or traffic rates) through road segments of a transportation network, one can optimize the transportation system to become smarter (better mobility, less congestion, less travel time, and less travel cost) and greener (less waste of fuel and less greenhouse gas production). Such a capability is crucial for the eagerly-anticipated Intelligent Transportation Systems (ITS) [1].

Until recently, in lack of accurate traffic data, traffic flows were estimated based sporadic surveys and/or simplistic simulations/models; hence, inaccurate and sparse traffic flow estimations. More recently, due to widespread use of GPS-enabled mobile devices (e.g., cell phones and PDAs), various projects (such as Millennium Project [2]) are initiated for GPS-based traffic flow analysis. This approach relies on collecting participatory GPS data from many individual travelers of the transportation network and aggregating the travel data to estimate real-time traffic flows. However, GPS-based traffic flow analysis also suffers from inaccuracy due to (often) limited user participation as well as inherent imprecision of the current GPS devices.

In this paper, our study is different with the previous work on traffic flow analysis in two ways. First, we use high resolution (both spatially and temporally) traffic loop-sensor data collected from the entire Los Angeles County road network (see Section 3 for more details). Loop-sensors directly monitor real-time traffic flows. Second, we focus on discovering patterns in the traffic flows (rather than traffic flows themselves). In particular, we would like to verify the following two hypotheses with ted data:

1. "There is a set of *signature traffic flow patterns* that can collectively represent the ongoing traffic flow on all segments of the road network with a fairly accurate approximation." If such a signature set exists, it can be exploited to can label each segment of the road network with a signature pattern (aka speed profiles) that accurately characterizes the typical traffic flow of the segment. In particular, we want to identify the members and cardinality of the signature set.

2. "There is a correlation between the signature traffic flow pattern of a road segment and other features of the road segment (such as locality, connectivity, etc.)." If such correlation exists, we intend to select the feature subset that yields the strongest correlation.

With a limited-sized signature set, among other use-cases, one can represent the typical traffic flow data of a road network with limited storage, can enable various time-aware planning processes (e.g., see [8, 13] for two instances of time-aware route planning processes) that consider typical traffic flow patterns in a road network

(rather than assuming fixed travel-time for each road segment) , and perhaps most importantly, can classify the road segments based on their signature traffic flow for a better understanding of their traffic behavior. Moreover, if we identify a subset of the road segments' features that are highly correlated with the signature traffic patterns of the road segments, we not only can have signature traffic patterns for the road segments with existing traffic sensor data, but also can leverage such a correlation to generate traffic flow data as well as signature traffic patterns for the (often many) segments of the road network for which traffic sensor data is not available (e.g., when data is missing, data is not proprietary and not open to public, and/or where traffic sensors are not installed at all).

Toward verifying the aforementioned hypotheses, first we applied an efficient unsupervised clustering technique and identified a set of eleven distinct signature patterns that accurately represent the typical traffic flows on all segments of the Los Angeles County road network with a maximum of 5mph error at each time instant. We use the set of signature traffic patterns to label each segment of the road network with the corresponding (best representative) traffic signature. Second, we defined a variety of (non-temporal) features (namely, length, direction, capacity, connectivity, density, and locality) for each road segment, and thereafter, used a feature subset selection technique to determine a subset of the features that demonstrate highest correlation with the corresponding traffic signature labels of the segments. In particular, we found that the combination of direction, connectivity and locality of a road segment can best predict the traffic signature of the segment with 0.695 positive correlation.

This study complements our prior work [6]. With [6], we presumed the two popular instinctive understandings and *assuming* the existence of a limited number of signature traffic patterns as well as a strong correlation between the signature patterns and some selected road segment features, we develop a traffic data generation framework. There, we leveraged these assumptions to estimate traffic flows for the road segments for which traffic data is not available. In contrast, in this paper we rigorously study these assumptions and demonstrate their validity.

The remainder of this paper is organized as follows. Section 2 reviews the related work. In Section 3, we describe our two real datasets (i.e., Los Angeles County road network and sensor data), and subsequently in Section 4, we establish the corresponding data models for these datasets. In Section 5, we discuss our approach toward discovering the signature traffic patterns. Next, in Section 6 we define the set of (non-temporal) features for road segments, and in Section 7 we discuss our approach toward studying their correlation with the traffic signatures of the road segments. Section 8 presents the results of our experimental analysis that verify the two hypotheses we posed above. Finally, in Section 9 we conclude and discuss our future work.

2. RELATED WORK

Recently, there has been considerable research on extending traditional data mining techniques to the context of traffic data analysis. Since we apply the Feature Subset Selection (FSS) technique to analyze traffic data, we categorize the related work into two parts: mining traffic data, and FSS techniques.

2.1 Mining Traffic Data

Two types of traffic dataset on road network have been studied: traffic sensor data, and trajectory data.

For the sensor network traffic dataset, Shekhar [18] employ various data mining techniques, such as time series relationship analysis techniques and clustering methods to explore the Twin-Cities traffic dataset, and visualize the discoveries on a highway map. In particular, they focus on the problems of detecting traffic flow outliers [19], and predicting sequential patterns [10]. This work presents several case studies on prediction or exploration of traffic flow and relevant to our first hypothesis, whereas few work has been done to investigate the road network features and the associated correlations between them and the sensor traffic data which is relevant to our second hypothesis.

For the trajectory dataset, microscopic simulation models, the so called activity-based models [4, 16, 11] are applied to detect the correlation between individual's activity and general traffic flow. In particular, Giannotti [11] apply spatiotemporal clustering methods in the context of trajectory traffic modeling to explore sequences of spatial areas of interests with similar traffic flow. These activity-based models, although relevant to our second hypothesis, always require the availability of the trajectories of the activities or sequences of behavior that happen within the network. These trajectories are in forms of GPS data with too many parameters which are hard to collect and often inaccurate.

2.2 Feature Subset Selection

Another line of work relevant to our study is feature subset selection (FSS) approaches. In general, FSS is one of the techniques to identify a relevant subset of original features from a given dataset by removing irrelevant and redundant features. Since the goal of FSS is to eliminate useless knowledge, it has a large variety of applications in the data mining field. In [21], Yoon apply feature selection techniques for hand gesture applications. Vainer et al[20] also present a feature selection approach for learning models that obtain accurate classification of large scale data of Voltage-Sensitive Dye Imaging. Zhao[22] propose a framework of feature extraction for classification of hand movement imagery based on both temporal and spatial features of Single-Trial EEG. However, these studies focus on feature subset selection as a way to preprocess the data to improve the prediction or classification performance, while our focus is to find a better way to detect underlying feature correlations among road network features and traffic signature patterns.

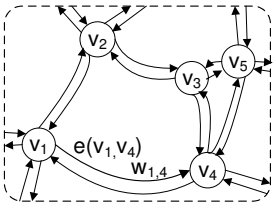
3. DATASETS

Here, we provide a detailed description of our Los Angeles County dataset. This dataset includes two types of data:

- **Road Network Data**: the detailed presentation of the road network . We obtain this data from Navteq [15], which includes the network skeleton and points of interests and the information along the road. Each network segment is represented in the vector data format and described by more than 20 attributes such as direction, speed limit, zip code, location, lane category, etc.

- **Traffic Sensor Data**: the dynamic traffic data on the road network. This sensor dataset is collected the from 1592 sensors (i.e.,loop detector) on the freeways during the period from October 2008 to June 2009 through RIITS [17]. A sensor turns on and off as cars pass over them. The sampling rate of the sensor data is 1 reading/sensor/min. The main traffic parameters measured by each sensor are *occupancy* and *volume*. The first parameter, occupancy O, is defined as the percentage of time a point on network segment is occupied by vehicles. The other parameter, volume V, is defined as the number of vehicles flowing past a point during a time interval. Based on these two parameters, we derive a third

(a) Road Network With Sensors



(b) Road Network Model

Figure 1: A partial snapshot of the Los Angeles Road Network

parameter, $Speed(S)$, from the occupancy and volume readings using the formula introduced in [3] $S = \frac{C*V}{O}$ where C is a constant proportional to the average length of a car.

4. DATA MODELS

In this section, we define our models corresponding to each dataset in turns.

4.1 Road Network Model

We model a road network as a direct graph $G(V, E, W_T)$, where $V = \{v_i\}$ is a set of nodes representing the intersections and terminal points, and $E \subseteq (V \times V)$ is a set of edges representing road segments between two nodes. Each edge e is denoted by $e_{i,j}$ where i and j indicate the index of starting and ending nodes of e respectively, and $i \neq j$. For every edge $e_{i,j} \in E$, we associate a weight $w_{i,j} \in W_T$, whose value is a function of the time variable t. This association indicates that over time , due to traffic congestion, we observe different costs to go through each edge. (Figure 1(a) illustrates a real road network with sensors and Figure 1(b) shows the corresponding road network model constructed)

4.2 Traffic Data Model

We model traffic flow F_s at a sensor s during a time interval $[t_s, t_e]$ as a time series $F_s = \{(S_i, t_i)\}$ where $t_i = t_{i-1} + T$ (T is a constant sensor sampling interval) and S_i is the estimated average speed during $[t_{i-1}, t_i]$.

Accordingly, we define traffic flow F_{uv} for a road segment e_{uv} from $G(V, E, W_T)$ as follows:

$$F_{uv} = \{\frac{1}{n}(\sum_{s \in e_{uv}} S_i), t_i\} \quad \forall t_s < t_i \leq t_e$$

where n is the number of sensors locates on the edge e_{uv} segment. Figure 2 shows an example of a traffic flow on a segment of I-405 freeway in LA between 6:00 AM and 9:00 PM on a weekday, where the time interval T is 15 minutes. As we could see from this figure, the speed of F_{uv} varies by time variable t and it also inversely reflect of the variation of weight w_{uv} on the segment e_{uv}.

5. MINING TRAFFIC FLOW PATTERNS

To verify the first hypothesis of Section 1, we apply X-means [5] to our traffic dataset to cluster and discover the representative traffic signature patterns.

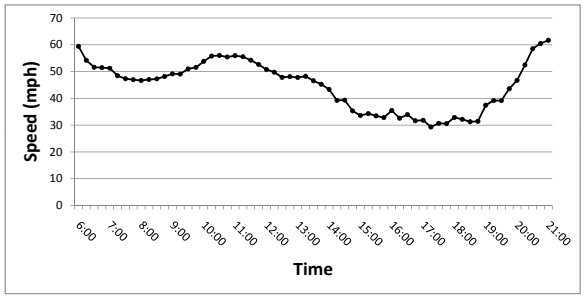


Figure 2: Real Time Traffic Flow Example

X-means is a clustering algorithm extending K-means [14] with efficient estimation of K (i.e., the number of clusters). This algorithm requires user only to specify a range in which the true K reasonably lies instead of providing the exact value of K in the K-means algorithms. The output is a set of centroids, one centroid per cluster, and the value of K. In essence, the algorithm starts by setting K to the lower bound of the user-provided range and continues to add centroids where they are needed until the upper bound is reached. During this process, the centroid set that achieves the best evaluation score is recorded and becomes the final output.

The X-means clustering algorithm partition the traffic data into finite groups of similar patterns. Although the centroids of these groups lose certain fine details as compared to each traffic flow, they could still be considered as good representatives for the whole traffic flow dataset. In addition, in our problem setting, the number of the signature patterns is unknown, unlike some other clustering methods such as K-means, X-means could also help us to identify the clusters without specifying the number of clusters. Owing to the above two functionality by X-means clustering algorithm, we apply this algorithm on our traffic sensor dataset. Specifically, our input includes the set of traffic flows F_{ij} generated by the traffic data model on all road segments from 6:00 AM to 9:00 PM with 15 minutes time interval as well as the range for the value K. We set the lower bound of K as 1, and the upper bound as the number of instances (i.e., road segments) participating in the clustering, since the number of representative signature patterns should be less than the total size of the dataset. The outputs are the cluster centroids which are in the same form as input instances. Therefore, we could use them as the traffic signature patterns to verify our first hypothesis. We present the result of this study in the Section 8.

6. NON-TEMPORAL ROAD NETWORK FEATURES

As we discussed above, the traffic flow signature is derived from the traffic sensor data on road segments, thus could be considered as the temporal feature of the road segments. In this section, we define a set of non-temporal features for each road segment derived from the structure of road network. In Section 6.1, we will use a feature subset selection technique to identify a subset of these non-temporal features that can best characterize the corresponding traffic flow signature (i.e., the traffic patterns) for each road segment. Table 1 provides a brief description of all the features.

6.1 Length (L)

Let e_{ij} be a directed edge in G, starting at v_i and ending at v_j. We define the Euclidean distance L_{ij} between v_i and v_j as the length of this segment. Trivially, in real road network, the seg-

Table 1: Summery of Spatial Road Network Properties

Terminology	Symbol	Description
Length	L	The Euclidean distance between two endpoint
Direction	D	The segment direction in the Navteq direction set
Spatial Capacity	C	The number of lanes
Connectivity	$F_{in}\&F_{out}$	Fan-in & Fan-Out
Density	D^*	The number of substitute lines in vicinity area
Locality	L^*	The locality label indicating the travel demand

ments are not necessary straight lines between two intersections. Therefore, this parameter is an estimation of the actual length of road segments. By incorporating the speed value of traffic flow F_{ij}, we could derive the travel time of edge e_{ij}. For a particular speed value, the length indicates how much time is required to pass through this segment. The longer the segment, the more time required to pass through it.

6.2 Direction (D)

Let e_{ij} be a directed edge in road network, D_{ij} denotes the direction to from start point v_i to the end point v_j. It is important to note that although in real networks, the segments are not necessarily along cardinal directions, we could approximate the direction of each segment using one of the four cardinal directions, i.e., $D_{ij} \in \{North, East, West, South\}$. The traffic may display various patterns along different directions. For example, in LA county, beaches are on the west of the downtown, therefore, during morning rush hour, the traffic near the beaches from west to east is always congested unlike opposite direction. Note that there might be other ways to elaborate the direction definition in the road network. However, the direction definition here is in accordance with the direction attribute standard from [15]. Since we use the data provided by [15] as the experiments data, we propose such definition here.

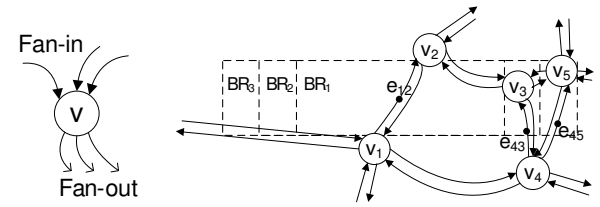
6.3 Spatial Capacity (C)

Let e_{ij} be a directed edge in road network, the spatial capacity $C(e_{ij})$ of e_{ij} is the number of lanes in e_{ij}. The capacity of a segment determines the amount of traffic load that can pass through the segment within particular time period. Based on this parameter, we could investigate the traffic load support capability of the segments.

So far we have analyzed the independent features of each road segment, now let us define the types of topological features of road segments.

6.4 Connectivity ($F_{in}\&F_{out}$)

Let e_{ij} be a directed edge in the road network, starting at v_i and ending at v_j. We define the fan-in $F_{in}(e_{ij})$ for e_{ij} as the number of directed edges ending at node v_i and the fan-out $F_{out}(e_{ij})$ for e_{ij} as the number of directed edges starting at the node v_j. Figure 3(a) illustrates the $F_{in}\&F_{out}$ edges connected to a node representing a road segment. Often, the incoming and outgoing traffic of a segment is positively correlated with the corresponding connectivity of the segment. We observe that if $F_{in} > F_{out}$, there is a higher chance of congestion building up at the segment particularly during the rush hour. In the reverse case, if $F_{out} > F_{in}$, the congestion is less likely to occur at the segment.



(a) Connectivity (b) Density

Figure 3: Two Road Network Topology Features

6.5 Density (D^*)

Let $BR(e_{ij})$ be the bounding rectangle covering segment e_{ij}, the density $D^*(e_{ij})$ is the number of directed edges e_{uv} with the following two properties: (1) $D_{ij} = D_{uv}$ (2)the center point of e_{uv} is within $BR(e_{ij})$. The $BR(e_{ij})$ indicates the vicinity area near segment e_{ij}. Often, if traffic on one segment e_{ij} is congested, people may re-route to other road segments nearby as an alternative to arrive their destinations. We consider these segments as substitutes for segment e_{ij}. Towards this end, $D^*(e_{ij})$ approximates the number of potential substitute segments of e_{ij}. It is important to note that different size of BR may result in different values of D^*. As illustrated in Figure 3(b), BR_1, BR_2 and BR_3 are three bounding rectangles of segment e_{12} with different sizes. The points indicate the center of the segments. According to our definition, $D^*(e_{12})$ may takes different values corresponding to the three BRs. In our experiments, for each road segment i, we use the rectangle with L_i in width, $2 \cdot L_i$ in length as the BR_i to calculate the density.

6.6 Locality (L^*)

The travel demand on a road segment depends on the neighborhood of the segment, i.e., the type of area at which the segment locates. Accordingly, the traffic patterns on a segment also depends on its locality. For instance, a road segment from residential area to downtown is often congested during the morning hours, whereas a segment in the opposite direction is usually congested in the afternoon.

To identify the locality for each segment of the Los Angeles County road network, we first use an automatic approach to estimate the locality and then manually adjust the estimated locality. Towards this end, we observe that the locality of a segment is correlated with the types of points of interests ($POIs$) (e.g., shopping center, restaurants, companies, etc) in the neighborhood of the segment. Therefore, we first used a density based clustering algorithm to identify the regions with similar types of nearby POIs. Thereafter, we manually tuned the clusters to label the region with the corresponding locality value. The locality of each segment is the label of the locality of the region in which it locates in (e.g., major residential regions, downtown). Table 2 gives a detailed descriptions of all 11 locality labels.

7. CORRELATIONS BETWEEN TEMPORAL AND NON-TEMPORAL FEATURES

In this section, we study the correlation between temporal features (i.e., traffic signature patterns discussed in Section 5) and the non-temporal features of the transportation road network (i.e., the features defined in Section 6).

One simple way to identify such correlations is to turn to domain experts. However, such domain knowledge is not always available for public access, and the knowledge may not always in agreement with each other. Towards this end, we identify the correlations based on the following intuition: for all segments, if one

Table 2: Locality Label Description

Label	Spatial Information for Freeway Segments
R	Residential Area
D	Downtown Area
B	Business Area
A	Attraction Area
RR	Remote Area, area far from downtown and res.
R2A	From Residential Area to Attraction Area
A2R	From Attraction Area to Residential Area
R2D	From Residential Area to Downtown Area
D2R	From Downtown Area to Residential Area
R2B	From Residential Area to Business Area
B2R	From Business Area to Residential Area

or the combination of multiple non-temporal features could predict the corresponding temporal features, there should be correlations between these two types of features. In Section 6, we provide a comprehensive definition of possible non-temporal features, but some of them probably never holds such correlations with temporal features. To filter them out and identify the optimal set of non-temporal features of road segments that can best characterize the corresponding temporal features of the segments, we use a feature subset selection (FSS) technique. Algorithm 1 shows the details of our FSS technique. We initialize the candidate feature set L with all non-temporal segment features defined in Section 6, and we start removing features from L one at a time and call the evaluation function E to measure the characterizing capability of the new feature set L' (i.e., how well L' could characterize the temporal features of all segments). If removing a feature results in enhanced value of E, the remaining features in L are considered as the current best feature subset and are stored in the $Solution$. At the end of each for loop, the all the features in the current best solution ($Solution$) are considered as the candidate features to be filtered in the next $repeat$ loop. The algorithm terminates when the $Solution$ set is no longer updated.

For the evaluation function E, as noted in the Algorithm 1, it takes in two parameters: the selected feature set a, and the dataset D. To implement E, we reformat the dataset D as follows: for each segment, we consider only the features in a as attributes and its traffic signature discovered in Section 5 as label. Subsequently, we partition reformatted D into training and test datasets. Next, we train a specific classifier with the training dataset on selected features. Finally, we use the test dataset to evaluate the classification accuracy, which is used as the output of E. Intuitively, the optimal classification accuracy based on the selected feature set means the feature set is the most discriminative feature set for the corresponding labels, that is, the selected feature set could best characterize the traffic signatures. Thus, the $Solution$ with maximum E value holds the tightest correlations with the temporal features.

For the specific classifier, we choose Bayesian network classifiers(BNCs). The reason we choose BNCs is many other pattern classification algorithms such as Support Vector Machines (SVMs), Multi-Layer Perceptrons (MLPs), and K-Nearest Neighbors (KNNs) require data to consist of purely numerical values, whereas the feature set in our problem also includes categorical variables, such as direction (D) and locality (L^*). In addition, among current studies [9, 7, 12], BNCs indicate a promising performance for multi-class classification tasks. For the detailed description of BNCs, please refer to [9].

Algorithm 1 Feature Selection Algorithm

Input:
S - the feature set including all feature, $|S|$ = 7;
D - the dataset includes all segments with their S values and labels
$E(a, D)$ - evaluation measure to be maximized based on the selected feature set a, and D
Output:
$Solution$ - optimized feature subset

```
1: let L be the candidate feature set, initialized with S
2: Solution ← L
3: repeat
4:    for all feature element f ∈ L do
5:       select the subset L' ← L\{f}
6:       if E(Solution, D) ≥ E(L', D) then
7:          Solution ← L'
8:       end if
9:    end for
10:   L ← Solution
11: until Solution is not updated in the for Loop
12: Return Solution
```

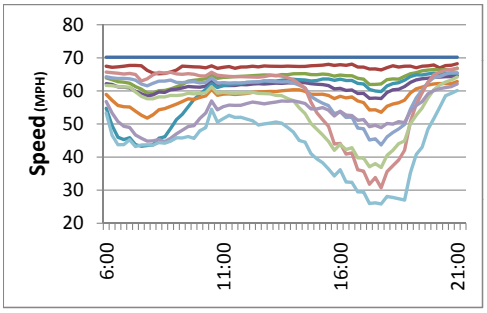
8. PERFORMANCE EVALUATION

In our experiments, we use the datasets described in Section 3. To verify our two hypotheses we conducted two sets of experiments to evaluate the accuracy of the discovered traffic signature patterns and to justify the correlations between the signatures and other road network features.

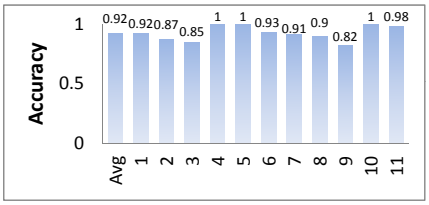
8.1 Evaluation Of Traffic Signatures

As stated in Section 5, we used X-means clustering method to discover traffic signatures. To evaluate the clustering results of X-means, we estimated the distance of cluster members to the cluster center at each time instant. As defined in the model, the speed value of each time instant in traffic flow is measured by mile per hour (MPH). The time interval T is set at 15 minutes and the time period for each traffic flow F_{ij} is from 6am to 9pm. Regarding each traffic flow F_{ij}, if each time instant of the signature pattern (i.e., cluster center) stays within 5MPH variation of the corresponding time instant in F_{ij}, we consider the signature pattern as an accurate representation of F_{ij}. In other words, if there is one time instant in the signature pattern that exceeds 5MPH variation of the same time instant in F_{ij}, the signature pattern is considered as a false representation for F_{ij}. And the accuracy is the ratio between the number of accurately represented traffic flow and the total number of traffic flows in the same clusters. In addition to this accuracy measure, we also apply the root mean square error(MSE) on each time stamp to quantify the speed difference. To calculate MSE, we consider each speed from signature patterns as the estimated value and the corresponding real speed from the traffic flow as the true value.

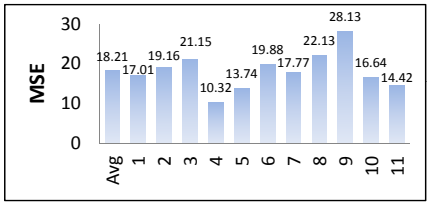
Figure 4(a) displays all the discovered signature patterns from X-means algorithm. There is a total of 11 signature patterns. Figures 4(b) and 4(c) show the corresponding accuracy and MSE within each clusters and the average. From the figure, we observe that on average, 92.87% of traffic data could be accurately represented by the signature patterns, and the average MSE is around 18.21. Based on this result, we could verify the first hypothesis that that there indeed exist 11 different signature patterns that could accurately represent all the possible traffic data in Los Angeles dataset.



(a) Discovered Traffic Flow Patterns



(b) Accuracy



(c) MSE

Figure 4: Evaluation of Discovered Traffic Patterns

8.2 Evaluation of Feature Selection

In this set of experiments, for each segment, we consider the 11 signature patterns discovered in Section 8.1 as distinct labels, and the features defined in Section 6 as attributes. As stated in Section 7, the Bayesian Network classifier is applied as the evaluation classifier for our FSS algorithm. The classification accuracy is derived by 10-fold cross validation on our dataset. For the 10-fold cross validation, we partitioned our dataset into 10 sets, and one of the sets is chosen to be test data, while all the other items are considered as training data. This is repeated N times and the average classification accuracy is reported.

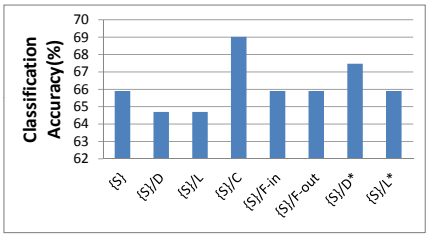
Table 3 depicts the result of the feature subset selection at each iteration of Algorithm 1. As we could see, in the 3rd iteration, the set of remaining features is no longer updated, therefore the algorithm terminates and the selected feature set is {Direction, Length, Fan-in, Fan-out, Locality}. The Spatial Capacity and the Density are filtered in the first two iterations.

More details of the evaluation result in the first and second iterations are shown in Figure 5. The x-axis is the feature set used for classification where $\{S\}$ indicates the set of features remaining in the last iteration. The y-axis is the classification accuracy. As observed from Figure 5(a), the classification accuracy improves significantly when the Spatial Capacity (C) feature is filtered out. From Figure 5(b), the accuracy also has slight improvement when the Density (D^*) feature is dropped. In the third iteration(see Figure 5(c)), the result of dropping any feature is worse than the result at the end of the second iteration, therefore, the algorithm terminates.

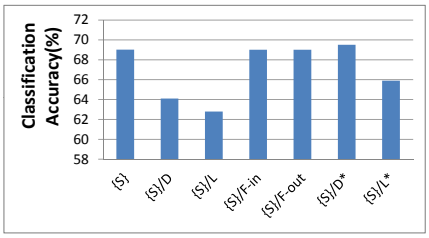
From the result of this experiment, based on the selected features, the classification accuracy could reach almost 70%. Therefore, there are tight correlations between the selected feature set(i.e., {Direction, Length, Fan-in, Fan-out, Locality}) and the temporal labels (i.e., traffic signatures), hence our second hypothesis is verified. Additionally, we also discover the spatial capacity and density features we defined hardly contribute to the correlations with the temporal traffic features. For spatial capacity, such result may indicates the traffic congestions reflected in temporal patterns does not necessary related to the lack of lanes in LA road network. On another hand, the elimination of density feature may due to the application of bounding box in the definition. The concept of bounding box might not correctly captures the detour behavior from people living in LA county.

Table 3: Procedure of feature subset selection

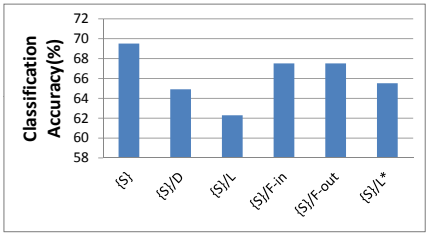
Iter-times	Filtered Features	Remaining Features
0th	N/A	Direction, Length, Spacial Capacity, Fan-in, Fan-out, Density, Locality
1st	Spacial Capacity	Direction, Length, Fan-in, Fan-out, Density, Locality
2nd	Density	Direction, Length, Fan-in, Fan-out, Locality
3rd	$NULL$	Direction, Length, Fan-in, Fan-out, Locality



(a) 1st iteration



(b) 2nd iteration



(c) 3rd iteration

Figure 5: The FSS Evaluation Result at Each Iteration

9. CONCLUSION AND FUTURE DIRECTIONS

In this paper, we analyzed real-world traffic sensor data (namely, the Los Angeles County dataset) and discovered that 1) there is a limited number of signature traffic patterns that can accurately represent the typical traffic flows on *all* segments of the corresponding road network, and 2) given the combination of direction, connectivity and locality of a road segment (and without having access to the actual traffic flow of the segment), one can distinctively determine the corresponding traffic signature of a road segment with high probability. We hypothesize (yet to be verified) that the same observations hold as general rules with other road networks; of course, the size and elements of the signature set as well as the set of selected characteristic road segment features may vary from one road network to another.

We intend to extend this work in two directions. First, we want to develop a generic framework to verify our hypothesis that generalizes our observations to other road networks. Second, we plan to leverage our observations to develop a traffic data generation tool that generates traffic flow data for those road networks for which traffic sensor data is not available.

10. ACKNOWLEDGMENTS

This research has been funded in part by NSF grants CNS- 0831505 (CyberTrust) and IS-1115153, the USC Integrated Media Systems Center (IMSC), and unrestricted cash and equipment gifts from Google and Microsoft. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

11. REFERENCES

[1] Intelligent transportation systems (by Research and Innovative Technology Administration of DOT). http://www.its.dot.gov/, 2010.

[2] The millennium project. http://traffic.berkeley.edu/, 2010.

[3] P. Athol. Interdependence of certain operational characteristics within a moving traffic stream. Technical report, National Research Council, Washington, D.C., 1967.

[4] C. R. Bhat, J. Y. Guo, S. Srinivasan, and A. Sivakumar. A comprehensive econometric micro-simulator for daily activity-travel patterns (cemdap). *Transportation Research Record*, 1894:57–66, 2004.

[5] A. M. Dan Pelleg. X-means: Extending k-means with efficient estimation of the number of clusters. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 727–734, San Francisco, 2000. Morgan Kaufmann.

[6] U. Demiryurek, B. Pan, F. Banaei-Kashani, and C. Shahabi. Towards modeling the traffic data on road networks. In *Proceedings of the Second International Workshop on Computational Transportation Science, IWCTS, ACMGIS*, Seattle, Washington, USA, 2009.

[7] L. Denoyer and P. Gallinari. Bayesian network model for semi-structured document classification. In *Information Processing and Management*, volume 40, page 807Ű827, 2004.

[8] B. Ding, J. X. Yu, and L. Qin. Finding time-dependent shortest paths over large graph. In *EDBT, 11th International Conference on Extending Database Technology*, pages 205–216, Nantes, France, March 2008.

[9] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29:131–163, 1997.

[10] B. George, J. Kang, and S. Shekhar. Stsg: A data model for representation and knowledge discovery in sensor data. In *Proceedings of Workshop on Knowledge from Sensor Data (SensorKDD '07) at ACM-SIGKDD*, San Jose, California, USA, November 2007.

[11] F. Giannotti, M. Nanni, D. Pedreschi, and F. Pinelli. Trajectory pattern mining. In *Proc.13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 330–339, San Jose, California, USA, November 2007.

[12] P. Helman, R. Veroff, S. R. Atlas, and C. Willman. A bayesian network classification methodology for gene expression data. *JOURNAL OF COMPUTATIONAL BIOLOGY*, 11, 2004.

[13] E. Kanoulas, Y. Du, T. Xia, and D. Zhang. Finding fastest paths on a road network with speed patterns. In *Proceedings of the 22nd International Conference on Data Engineering, ICDE*, 2006.

[14] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, page 281Ű297. University of California Press, 1957.

[15] Navteq. http://www.navteq.com, 2007.

[16] V. Peter and B. Mark. Advanced activity-based models in a context of planning decisions. *Journal of the Transportation Research Board, TRB, National Research Council*, pages 34–41, 2005.

[17] RIITS. Regional integration of intelligent transportation systems. http://www.riits.net, 2008.

[18] S. Shekhar, C.-T. Lu, and P. Zhang. Data mining and visualization of twin cities traffic data. Technical Report 15, University of Minnesota, 2001.

[19] S. Shekhar, C.-T. Lu, and P. Zhang. A unified approach to spatial outliers detection. *GeoInformatica*, 7(2), June 2003.

[20] I. Vainer, S. Kraus, G. A. Kaminka, and H. Slovin. Scalable classification in large scale spatiotemporal domains applied to voltage-sensitive dye imaging. In *In Proceedings of the IEEE International Conference on Data Mining(ICDM'09)*, pages 543–551, 2009.

[21] H. Yoon and C. Shahabi. Feature subset selection on multivariate time series with extremely large spatial features. In *Sixth IEEE International Conference on Data Mining - Workshops (ICDMW'06)*, Hong Kong, China, 2006.

[22] Q. Zhao and L. Zhang. Temporal and spatial features of single-trial eeg for brain-computer interface. In *Intell. Neuroscience*, volume 2007, pages 4–4, 2007.