Spatio-Temporal Analysis of Vehicular Traffic Data

Akash Agarwal

Under the supervision of

Prof. Soumya Kanti Ghosh

Department of Computer Science & Engineering

Indian Institute of Technology, Kharagpur

May, 2016

Overview

Introduction

- Nowadays huge volume of historical vehicular traffic data is available on the internet
- Source of data is either from GPS devices, sensors present on road, etc
- Data available for differernt road segments and for different time intervals can be used for analysis
- The available incident data can be used to predict the incidents in future which can help in decision making in Cyber Physical System(CPS)

Motivation and Objectives of the Present Work

Motivation

- Traffic incidents kill around 1.24 million people worldwide [1]
- Results in enormous cost/loss to the society
- Increasing need of efficient methodologies for identifying the risk factors of the accidents

Objective

- Identification of feature set (predictor features and target features)
- Pre-processing of data
- Implementing classification algorithms like CART, Random Tree, Bayesian Network etc.., for classification
- Generate classification rules between traffic features and frequency of occurance of incidents and draw inferences
- Evaluating the performances of different classifiers using training, testing, previously unseen data

[1] http://www.progressive-economy.org/trade_facts/

Data Collection and Area of Study

- **Data Source**:
  - California Department of Transportation (Caltrans) Performance Measurement System
  - Microsoft Fetch Climate Explorer
- **Study Area**: Incidents reported in district 11, freeway I-5 only for the MainLine type of Roads(both north and south bound) over the year 2014 and 2015

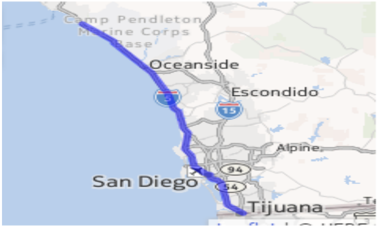
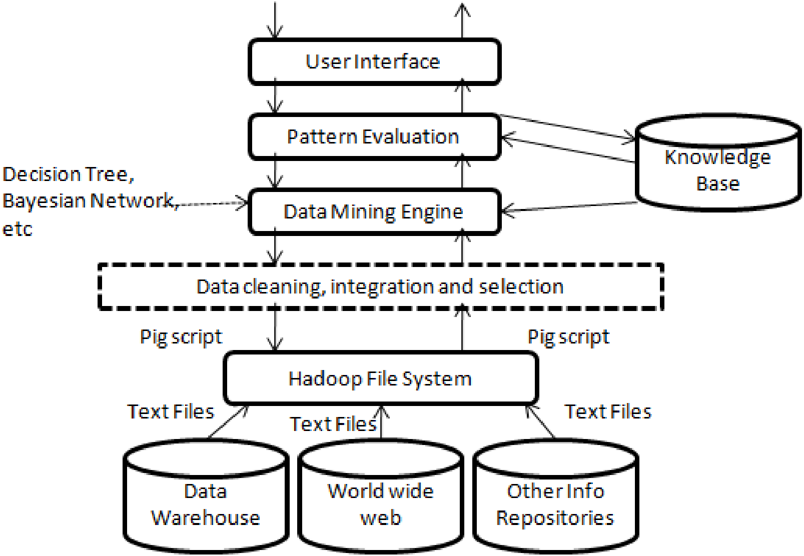


Figure: Freeway I-5 of District 11, California

Generation and Analysis of Prediction Model

Overall Architecture



Variable Description

- Timestamp
- Station ID
- Freeway Number
- Freeway Direction
- Direction of Travel
- Total Flow
- Average Occupancy
- Average Speed
- Absolute Postmile
- No of Lanes

Data Preprocessing

- Data sets used for attribute/feature selection from PeMS website
  - Station hourly data
  - Station metadata
  - Incident data
- Data sets used for attribute/feature selection from PeMS website
  - Precipitation rate
  - Sunshine fraction
- Approximately 24 million records of station hourly data and 2 milion records for incident data is available for analysis for the year 2014 and 2015

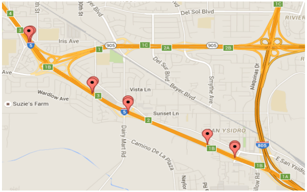


Figure: Location of few stations

Data Preprocessing (contd...)

Filtering Dataset

- Filtering of dataset done based on area of study and approx. 4 million records were obtained from the station raw data, and 0.02 million records from incident data
- In station hourly data, the missing values were imputed by linear regression

Data Transformation

- The station data is normalized from per hour to 6 equally spaced time slot distributed over the day.
- This helps in reducing the sparsness of dataset

Data Preprocessing (contd...)

Table: Time of Day and Slot

Time of Day	**Slot**	**Interpretation**
00:00 AM - 03:59 AM	1	Late night
04:00 AM - 07:59 AM	2	Early morning
08:00 AM - 11:59 AM	3	Morning
12:00 PM - 03:59 AM	4	Afternoon
04:00 PM - 07:59 PM	5	Evening
08:00 PM - 11:59 PM	6	Night

- Station raw data processed and the features were accumulated over each slot
- 0.5 million data records were left
- Features are either summed up or their mean taken

Data Preprocessing (contd...)

Integration of Dataset

- Station Metadata and Station Raw Data were merged by the Station ID
- Obtained data set was merged with the incident data
- For large number of rows (90%) count value is 0,
  - Those rows were removed
  - Final dataset contains approx. 10,000 entries

Feature Selection

Table: Feature Selection based on gain ratio

0.03827	timeOfDay
0.02087	Flow
0.02041	dayOfYear
0.02011	Occ
0.01847	Speed
0.00332	noOfLane
0	PPT_rate
0	SS_fraction
0	Direction

Data Preprocessing (contd...)

Finally, the chosen features are

- Spatial Features
  - Flow
  - Occupancy
  - Speed
- Temporal Features
  - Time of Day
  - Day of the year

Feature Analysis

The distribution of incidents over the year(2014) is shown below

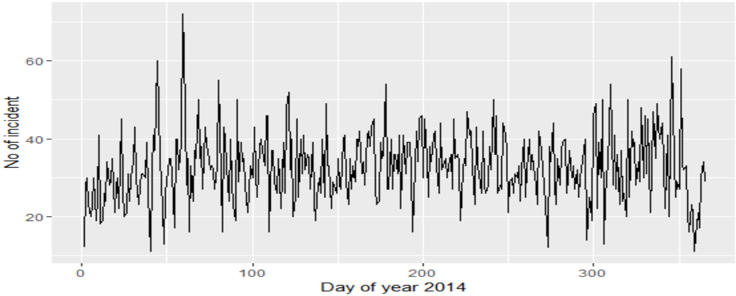


Figure: No of incidents per day

The spread of incidents is more towards the end of February and starting of March

Feature Analysis (contd...)

The distribution of incidents based on time slot is shown in figure below

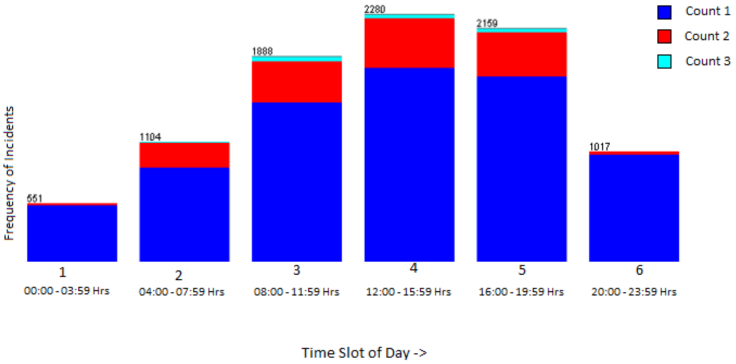


Figure: Variation of number of incidents based on Time of Day

Here the incidents are less during the earlly morning and late night hours

Feature Analysis (contd...)

Table: Analysis of temporal parameters with respect to time of day

Slot of Day	Mean Flow (no of vehicles)	Mean Speed (miles/hour)	Mean Occupancy (%)
1	2603	69.45554	0.01205082
2	11697	66.19565	0.06210145
3	18718	61.95455	0.09318856
4	20519	59.92105	0.11044298
5	19152	57.10514	0.1213988
6	9818	68.6234	0.04459194

Feature Analysis (contd...)

- The mean flow during the forth slot is highest which is coincident with the fact that most number of incidents took place during that time period
- The flow decreases during the early hours and late hours of day the speed is more, as less flow means vehicles can move freely

Feature Analysis (contd...)

Analysis of features with respect to the incident count

FLOW		Class				
	Class Attribute	1	2	3	4	5
%age Contribution		0.82	0.17	0.01	0	0
	Mean	15836.41	19039	19146.5	18039.78	18892.47
	std. deviation	7165.739	5771.523	5449.169	5725.543	3727.44
	weight mean	7343	1517	115	21	3

OCC		1	2	3	4	5
	mean	0.0844	0.114	0.1207	0.1268	0.1438
	std. deviation	0.0578	0.0651	0.0643	0.065	0.089
	weight mean	7343	1517	115	21	3

SPEED		1	2	3	4	5
	mean	62.2365	58.2673	56.7113	52.5823	52.8296
	std. deviation	9.2311	10.154	10.788	13.73	16.3071
	weight mean	7343	1517	115	21	3

Analysis of Prediction Model

- For all the classifiers the training set used was data for the year 2014(8999 records) and data for the year 2015(10143 records) was used as testing set
- The target class(no of incidents) has target variables as 1,2,3,4 and 5

Confusion matrix

Table: Confusion Matrix of Naive Bayes Classifier for all classes

a	**b**	**c**	**d**	**e**	**<− classified as**
7114	777	0	56	24	a = 1
1576	341	0	19	11	b = 2
123	35	0	6	0	c = 3
30	9	0	2	1	d = 4
10	6	0	3	0	e = 5

- Correctly Classified Instances 73.5187% (7457)

Table: Confusion Matrix of Simple CART Classifier for all classes

a	**b**	**c**	**d**	**e**	**<– classified as**
6864	1038	56	13	0	a = 1
1585	351	10	1	0	b = 2
134	30	0	0	0	c = 3
32	9	1	0	0	d = 4
14	5	0	0	0	e = 5

- Correctly Classified Instances 71.1328% (7215)

Analysis of Prediction Model (contd...)

Table: Confusion Matrix of Random Tree Classifier for all classes

a	**b**	**c**	**d**	**e**	**<– classified as**
6489	1311	139	22	10	a = 1
1472	429	35	10	1	b = 2
127	31	5	1	0	c = 3
27	13	2	0	0	d = 4
15	4	0	0	0	e = 5

- Correctly Classified Instances 68.254% (6923)

Analysis of Prediction Model (contd...)

Modification of number of classes

- Zero true poitive for class three onwards except for random tree which is able to classify two instances of class 3 properly
- The accuracy of random tree is less compared to naive bayes
- Therefore target class was modified to two classes, one and more than one

Confusion Matrix for two classes

Table: Confusion Matrix of Naive Bayes Classifier for two classes

a	**b**	**<– classified as**
7087	884	a = 1
1712	460	b = >1

- Correctly Classified Instances 74.406% (7547)

Table: Confusion Matrix of Simple CART Classifier for two classes

a	**b**	**<– classified as**
6861	1110	a = 1
1716	456	b = >1

- Correctly Classified Instances 72.1384% (7317)

Analysis of Prediction Model (contd...)

Confusion Matrix for two classes

Table: Confusion Matrix of Random Tree Classifier for two classes

a	**b**	**<− classified as**
6515	1456	a = 1
1677	495	b=>1

- Correctly Classified Instances 69.1117% (7010)
- Slight increase in the accuracy compared to the model which consists of all the target classes

Conclusion

- Two classes classification gives slightly better results compared to five class classification
- Naive Bayes classifier performed the best with 72% accuracy while for Random Tree classifier the accuracy is 69%
- The Random Tree classifier works better for class >1 records
- Random tree preferable as with little loss in overall accuracy(3%) >1 class is classifed with greater accuracy

Future Work

- Investigating correlations between features and the target variable and draw inferences
- Enhancement of the training set
- Replicate the generated model to data sets acquired from other sources
- Deriving inference based on the context
- Spatio temporal analysis of vehicular trajectory data

References

- **Chang, Li-Yen, and Wen-Chieh Chen.** 'Data mining of tree-based models to analyze freeway accident frequency.' *Journal of Safety Research* 36.4 (2005): 365-375.
- **Ying Jin, Jing Dai, Chang-Tien Lu** 'Spatial-Temporal Data Mining in Traffic Incident Detection.' *Proc. SIAM DM 2006 Workshop on Spatial Data Mining* Vol. 5. 2006.
- **Min, Wanli, and Laura Wynter.** 'Real-time road traffic prediction with spatio-temporal correlations.' *Transportation Research Part C: Emerging Technologies* 19.4 (2011): 606-616.
- **Liu, Wei, et al.** 'Discovering spatio-temporal causal interactions in traffic data streams.' *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM, 2011.

Thank You