The Prediction of Traffic Congestion and Incident on Urban Road Networks Using Naive Bayes Classifier

Guangxing Wang¹, Jiwon Kim¹
¹The University of Queensland, Brisbane St Lucia, QLD 4072, Australia
Email for correspondence: jiwon.kim@uq.edu.au

Abstract

This study proposes a Naive Bayes (NB) classifier model for predicting congestion and incident in urban road networks. NB is a machine learning classification model based on Bayes' theorem with strong (probabilistic) independence assumptions between predictor variables. This study considers congestion or incident as a target variable and applies a NB model to classify its state (i.e., occur vs. not occur). Predictor variables or features considered in this study include network environment variables (time of day, day of week, and weather) and traffic condition variables (speed on bottlenecks). This study develops a data-driven approach for building and validating NB models. The models are trained and tested using actual traffic, incident, and weather data collected from Brisbane, Australia in 2014. The validation results show that the proposed models can successfully predict congestion and incident occurrence with a desired level of accuracy.

1. Introduction

According to a report (2014) from INRIX in 2014, merely in the U.S., traffic congestion has caused 124.2 billion US dollars of economic loss in 2013 and will increase 50% to 186.2 billion dollars by 2030. Over that period, the cumulative congestion cost for US will reach 2.8 trillion dollars. Traffic incidents, especially crashes, cost even more. In 2011, a report from AAA (2011) pointed out that the cost of overall annual crashes is more than three times the congestion cost in US. In 2013, crashes took 1.25 million lives all over the world (WHO, 2015).

In order to mitigate or even prevent congestion and incidents, traffic managers need to predict when and where they will occur. With the availability of large amounts of data, data-driven approaches based on machine learning techniques have been increasingly applied in addressing traffic prediction problems. Among various machine learning approaches, we limit our attention to the approaches based on Bayesian Network (BN) models (Pearl, 1988). Bayesian networks are probabilistic graphical models that encode the joint probability distribution over a set of random variables using a graph representation and have been widely applied in modelling and predicting probabilistic relationships among variables. A brief introduction to BN can be found in our recent study (Kim and Wang, 2016). The existing studies that use BN approaches include:

- **Congestion prediction:** From 2004 to 2006, some of the early works that apply Bayesian networks in traffic forecasting problems were published (Sun et al., 2004, Changshui et al., 2004, Sun et al., 2005, Sun et al., 2006). In the last decade, various studies have developed models for traffic flow forecasting and short-term traffic prediction using Bayesian network method (Castillo et al., 2008, Yu and Cho, 2008, Pascale and Nicoli, 2011, Hoong et al., 2012, Liu et al., 2014, Zhou et al., 2014). In 2012, (Horvitz et al., 2012) proposed a highly accurate congestion prediction model called JamBayes, which is a continuous Bayesian network model. The model uses
temporal traffic variables that indicate the future and past states of congestion of the bottlenecks. Recently, Kim and Wang (2016) have presented an analysis framework that uses a Bayesian network to diagnose and predict traffic congestion.

- **Incident prediction**: (Zhang and Taylor, 2006) proposed a Bayesian network based algorithm for road accident detection. It modelled the causal dependencies between accident and traffic parameters quantitatively and updated accident probability based on real-time traffic data. (Hongguo et al., 2010) developed a causal dependence structure for accident and its causes using K2 algorithm. It provided a new perspective on how to reveal the accident causality mechanisms. (Gregoriades and Mouskos, 2013) illustrated an approach taking site factors, such as bus stop and road work, into account to draw the high accident risk points on map.

In this study, we propose a Naive Bayes classifier model for the prediction of traffic congestion and incidents.

### 2. Naive Bayes classifier

Naive Bayes (NB) classifier is a special type of Bayesian network designed for classification problems (Friedman et al., 1997; Rish, 2001). This is a simple probabilistic classification model that computes the probability of a target variable (or class variable) given particular instances of feature variables (or predictors, attributes) and then predicts the class of the target variable with the highest posterior probability. This computation is performed effectively by making a strong independence assumption that all the feature variables are conditionally independent given the value of the target variable. The main advantage of NB over other machine learning models is its computational simplicity. NB models can be trained very fast and is highly scalable, which scales linearly with the number of predictors. In addition, NB is robust to noisy data as it is not sensitive to irrelevant features.

A NB model typically contains one target variable (as model output) and multiple feature variables (as model input). Let \( T \) be the state or class of target variable and a vector \( X = (x_1, x_2, \ldots, x_n) \) be the states of \( n \) features. In order to infer the value of \( T \) based on \( X \), the probability of \( T \) given \( X \) need to be calculated first. Below shows the calculation process:

Based on Bayes' theorem, the conditional probability of \( T \) given \( X \) is expressed as:

\[
p(T|X) = \frac{p(X|T)p(T)}{p(X)},
\]

where \( p(X) \) and \( p(T) \) are constants that can be directly derived from data, while \( p(X|T) \) is remaining to be solved. Based on the feature independence assumption of NB, \( p(X|T) \) can be factorized as:

\[
p(X|T) = p(x_1, x_2, ..., x_n|T) = \prod_{i=1}^{n} p(x_i|T).
\]

By combining the previous two equations, we have:

\[
p(T|X) = \frac{p(T)}{p(X)} \prod_{i=1}^{n} p(x_i|T),
\]

where \( p(T) \), \( p(X) \), and \( p(x_i|T) \) are the parameters of NB models. In this study, these parameters are learned directly from training data. Therefore, the conditional distribution of \( T \) given \( X \) can be calculated using Equation (3). The classification outcome, i.e., the value of target variable \( T \) given the values of \( X \), is given as the state of \( T \) with the highest probability.
Prediction of Traffic Congestion and Incidents on Urban Road Networks
Using Naive Bayes Classifier

3. Methodology

NB models considered in this study are a set of Naive Bayes classifiers that have the same features but different targets. In other words, we build a separate NB model for each target variable but use the same model structure and feature variables across all target cases. Table 1 presents the list of variables and their definitions.

3.1. Target Variables

This study uses a binary target variable that has two possible states: alarm (positive) or no alarm (negative). These states are further defined based on the type of target variable (i.e., congestion vs. incident) as follows:

- **Congestion**: For congestion prediction, the target variable indicates “whether traffic congestion will occur on a given link at a time interval of 15 minutes after a specific prediction time horizon”. For example, if we denote a variable by “congestion_1025_30mins”, this represents a binary variable indicating whether congestion will occur on link 1025 during the 15-min time interval after 30 minutes have elapsed. This study considers four different prediction horizons: 15min, 30min, 45min, and 60min, producing four target variables for each link that represent the link congestion states after 15min, 30min, 45min, and 60min, respectively.

The state of congestion is defined as the state where the link speed is lower than 50% of the 99th percentile of the speed distribution on that link. The speed value of the 99th percentile is used as the upper bound of normal speed range in order to prevent the effect of large-value outliers. For example, in link 1025, the highest speed value recorded is 255km/h while the 99th percentile speed is 80km/h. In this case, the speed threshold is set to 40km/h and link 1025 is marked as congested when its speed is lower than 40km/h.

- **Incident**: Unlike traffic congestion, incidents are rare events that occur with much low frequency. As such, predicting them with the same 15-minute time interval is not reliable and is not practically meaningful. Instead, we aim to predict whether an incident will occur within the next 1 hour. Therefore, a target variable for incident prediction is defined as to indicate “whether incident will occur on a given link within a specified time duration”, where the time duration is set to be 1-hour.

There are three categories of incidents considered in this study: crash, hazard, and stationary vehicle. These will be analyzed as separate incident target variables.

3.2. Feature Variables

Details of feature variables used in this study are provided below:

- **Time of day**: it represents the hour value of the recording time. For example, state “h_0” represents the record time of 0:00am to 0:59am.
- **Day of week**: it represents the day of week of the recording time. For example, “Monday” represents that the recording time is Monday.
- **Weather**: it represents rain intensity. For example, if there is no rain, then the state is “Clear”. Its discretization details are presented in Table 1.
- **Speed**: it represents the traffic speed of a specific link. It is discretized based on the upper bound of speed which is mentioned in 3.1. Speed data are normalized by dividing the speed values for each link by its associated upper bound. Once speed data are normalized to range between 0 and 1, the three values of 0.25, 0.50, and 0.75 are set as break points to discretize data into four states {Very low, Low, High, Very high} accordingly. The discretization details are shown in Table 1.
Table 1 Variables and State Definitions for the Proposed NB Model

<table>
<thead>
<tr>
<th>Variable Type</th>
<th>Name</th>
<th>States and States Definitions</th>
<th>Variable Defined</th>
</tr>
</thead>
</table>
| Target variables | Congestion   | -Alarm: occurs after a specified time horizon  
-Non alarm: not occurs after a specified time interval                                           | For each link;                    |
|               | Crash         | -Alarm: occurs within the next one hour  
-No alarm: not occurs within the next one hour                                                     | For each link                     |
|               | Hazard        |                                                                                                 |                                  |
|               | Stationary vehicle |                                                                                              |                                  |
| Feature variables | Time of day  | -h_n (n = 0, 1, 2, ..., 23)  
Sunday - Monday - Tuesday - Wednesday - Thursday - Friday - Saturday                                | Network-wide                      |
|               | Day of week   |                                                                                                 | Network-wide                      |
|               | Weather       | -Clear: 0mm/h  
-Light rain: < 2.5mm/h  
-Moderate rain: 2.5 – 7.6mm/h  
-Heavy rain: ≥ 7.6mm/h  
-Very low: < 0.25*  
-Low: 0.25 – 0.5  
-High: 0.5 – 0.75  
-Very high: ≥ 0.75  

*based on normalized value

3.3. Model Training

Once model variables are defined, the next step is to train the model to estimate NB model parameters described in Eq (3) in Section 2 (i.e., p(T), p(X), and p(x_i|T)). These parameters can be learned from the training data set by performing the Maximum-likelihood estimation.

4. EXPERIMENTAL SETUP

4.1. Study Site

The study site selected for this study is the greater Brisbane region shown in Figure 1. As indicated in Table 1, congestion, incidents, weather and speed variables are defined for a specific location (e.g., either link or weather station). Each of these location-specific variables are therefore defined for its associated location.

Five weather variables are defined for the five weather stations distributed in the study site as shown in Figure 1. For the link-specific variables, we selected 35 links for each variable group based on the selection criteria in Table 2. Four sets of 35 links were selected: one set for the speed and congestion variable groups, which represents top 35 bottleneck links, and the other three sets for the three incident variable groups, respectively, each of which represents incident hot-spots for the respective incident type.

Figure 2 shows the locations of the links included in these sets. The four sets are not necessarily the same. They, however, share many links in common, which include Gympie Arterial Rd, Gateway Motorway, Pacific Motorway, Centenary Motorway, and Ipswich Motorway. This indicates that there exists a correlation between traffic bottlenecks and
incident hot-spots, which meets our expectation that including bottlenecks and incident hot-spots in the model could help the prediction of a target variable (congestion or incident) given the predictor variables (speed).

Table 2 Link number and selection criteria of variable groups

<table>
<thead>
<tr>
<th>Variable group</th>
<th>Number of links</th>
<th>Selection criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speed and congestion</td>
<td>35</td>
<td>Top 35 bottlenecks: Annual congestion time* &gt; 350 hrs/yr, and Average flow &gt; 1200 veh/h</td>
</tr>
<tr>
<td>Crash</td>
<td>35</td>
<td>Top 35 links for crash</td>
</tr>
<tr>
<td>Hazard</td>
<td>35</td>
<td>Top 35 links for hazard</td>
</tr>
<tr>
<td>Stationary vehicle</td>
<td>35</td>
<td>Top 35 links for stationary vehicle</td>
</tr>
</tbody>
</table>

*total times with the state of congestion defined in 3.1

Figure 1 Study Site (blue points represent weather stations)
Figure 2 Selected 35 Links for Each Variable Group (red lines)

4.2. Data Collection and Overview

For this case study, one-year of data from 2014 were used. Data were collected from different sources: traffic speed, incidents and weather data. All the data were processed such that observations are aggregated over a 15-minute interval and merged by matching the observation time intervals. The merged data starts from 2014-1-1 to 2014-12-31, which contain a total of 35040 records.

- **Traffic speed data** is from Queensland Department of Transport and Main Road (DTMR) through STREAMS Business Intelligent (BI). Its original time interval was 1-3 minutes. Therefore, it was processed to fit 15-minute intervals through linear interpolation.
- **Incident data** is from DTMR through STREAMS Incident Management System (SIMS). There were 34365 incidents in total. Since the original incident data were not mapped onto the underlying network, a mapping between incident sites and the underlying links was performed. For each incident, we inferred the associated incident link based on the nearest links whose distance to the incident site is less
than 20 meters. The definition of distance is the possible shortest length between a random point on link to incident point.

- **Weather data** is from Bureau of Meteorology (BoM). Its original time interval was 30 minutes. Therefore, it was processed to fit 15-minute intervals through linear interpolation.

After fusing these data, all the records are aligned such that data in each row are from the same 15-minute time interval. For each target variable, a 35040-by-43 matrix is constructed, where 43 variables consist of 1 target variable and 42 feature variables with the 42 features containing 1 time-of-day variable, 1 day-of-week variable, 5 weather variables, and 35 speed variables (from 35 bottlenecks). This matrix is then used as training data. An example of this training matrix is shown in Figure 3, which is for target variable “Congestion_link1026_after_45mins”.

**Figure 3 Example data matrix**

<table>
<thead>
<tr>
<th>Con_1026_after_45mins</th>
<th>Day_of_Week</th>
<th>Time_of_Day</th>
<th>Weather_station1</th>
<th>...</th>
<th>Weather_station5</th>
<th>Speed_1</th>
<th>...</th>
<th>Speed_35</th>
</tr>
</thead>
<tbody>
<tr>
<td>NoAlarm</td>
<td>Wednesday</td>
<td>h_0</td>
<td>NA</td>
<td></td>
<td>NA</td>
<td>VeryHigh</td>
<td>...</td>
<td>High</td>
</tr>
<tr>
<td>NoAlarm</td>
<td>Wednesday</td>
<td>h_0</td>
<td>NA</td>
<td></td>
<td>Clear</td>
<td>VeryHigh</td>
<td>...</td>
<td>Low</td>
</tr>
<tr>
<td>NoAlarm</td>
<td>Wednesday</td>
<td>h_0</td>
<td>Clear</td>
<td></td>
<td>Clear</td>
<td>VeryHigh</td>
<td>...</td>
<td>Low</td>
</tr>
<tr>
<td>NoAlarm</td>
<td>Wednesday</td>
<td>h_1</td>
<td>Clear</td>
<td></td>
<td>Clear</td>
<td>VeryHigh</td>
<td>...</td>
<td>Low</td>
</tr>
<tr>
<td>NoAlarm</td>
<td>Wednesday</td>
<td>h_1</td>
<td>Clear</td>
<td></td>
<td>Clear</td>
<td>VeryHigh</td>
<td>...</td>
<td>Low</td>
</tr>
<tr>
<td>NoAlarm</td>
<td>Wednesday</td>
<td>h_1</td>
<td>Clear</td>
<td></td>
<td>Clear</td>
<td>VeryHigh</td>
<td>...</td>
<td>Low</td>
</tr>
<tr>
<td>NoAlarm</td>
<td>Wednesday</td>
<td>h_1</td>
<td>Clear</td>
<td></td>
<td>Clear</td>
<td>VeryHigh</td>
<td>...</td>
<td>Low</td>
</tr>
<tr>
<td>NoAlarm</td>
<td>Wednesday</td>
<td>h_2</td>
<td>Clear</td>
<td></td>
<td>Clear</td>
<td>VeryHigh</td>
<td>...</td>
<td>Low</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
5. MODEL VALIDATION

Once models are built, it is important to assess their prediction performance based on actual data. In this study, prediction values are first generated through k-fold validation. Then validation parameters are calculated through comparing actual values and prediction values. The sections below will show detailed validation procedures and final validation results.

5.1 K-fold Validation

K-fold validation is used to generate prediction values for validation purposes. Below are the detailed procedures:

- **Separation:** The full model data set is separated into k equal-sized subsamples randomly. In this study, k = 10 is used. This divides the original data matrix of 35040 rows into 10 random subsamples that have 3504 rows each.

- **Loop:** For each subsample, the model is first trained with all other subsamples (as training set). Then the prediction values for that subsample are generated through inputting features of it row by row. For example, in the last (tenth) iteration, first to ninth subsamples are used to train the model. Then for each row in tenth subsample, the prediction value of that row is generated through inputting the feature values of that row into the model. The prediction value is the state with highest probability in target node.

- **Combine:** Combine the k parts of prediction values and form a new column of data matrix. After the loop above is finished, all rows in data matrix have their own prediction value. The prediction column has the same length as other variable in matrix so can be appended into it.

After this process, the prediction values of the data are produced. It will be pasted to next procedure of validation.
5.2. Validation Parameters

The proportion of records that have congestion or incidents alarm is less than 5% so target variables are all skewed classes. It means evaluating model performance using classification accuracy is not appropriate. For example, a simple algorithm always predicts no alarm will have more than 95% accuracy in this study.

In order to assess prediction performance of models properly, in this study applied precision, recall, and F1 score as validation parameters. Below is their detailed information:

- **Precision** is the ratio of true positive prediction to all positive prediction. In this study, it equals to correct alarm prediction over all alarm prediction. In other word, it represents the proportion of alarm prediction that is actually right.
- **Recall** is the ratio of true positive to all actual positive records. In this study, it equals to correct alarm prediction over all actual alarm records. In other word, it represents the proportion of alarm record that can be predicted by model.
- **F1 score** takes precision and recall into consideration. It is the harmonic mean of precision and recall:

\[
F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}
\]

For each model in this study, classification accuracy and above 3 validation parameters are calculated.

5.3. Validation Results

After running through all the procedures in 5.1 and 5.2 for all 245 NB models, calculated validation parameters can be used for assessing the overall prediction performance of this study. Below sections will show some of the statistics as the validation results of this study.

**Congestion prediction performance:**

For congestion prediction, the mean value of precision, recall, and F1 score are respectively 0.56, 0.85, and 0.66. Box-plot as Figure 5 can show more statistics of the performance.

Time attribute do affect prediction performance. Generally, the further the future we try to predict, the lower the performance is. All the statistics, such as mean and median, decrease as the time attribute become longer. It can be proved by Figure 6 below. Such phenomenon fits common expectations that further future is more unpredictable.

**Incidents prediction performance:**

For incidents prediction, the mean value of precision, recall, and F1 score are respectively 0.02, 0.29, and 0.04. Box-plot in Figure 7 shows more statistics of incidents prediction performance.

Different categories incidents have different prediction performance. If only the average F1 score is considered, the prediction of stationary vehicle is the best, crashes go second, and hazard is the worst. Below Figure 8 shows more comparison details. Hazard prediction being the worst also fits common expectation because many hazard records, such as fire, are irrelevant to traffic system.
**Figure 5** Box-plot of congestion prediction validation

![Box-plot of congestion prediction validation](image)

**Figure 6** Congestion prediction validation parameters with different time attributes

![Congestion prediction validation parameters with different time attributes](image)
Prediction of Traffic Congestion and Incidents on Urban Road Networks
Using Naive Bayes Classifier

Figure 7 Box-plot of incidents prediction validation parameters

Figure 8 Prediction validation parameters of different kinds of incidents
6. DISCUSSION

6.1. Methodology of NB Model Development

These findings are about NB model development, which can be a reference source of similar researches.

- **Model level**: This study sets link-level as model level. In other word, all traffic related variables such as congestion, incident and speed, are indicating the states of responding links.

- **Model number**: This study builds a separate NB model for each target variable. Although building a large model (such as an Bayesian Network) which includes all target and feature variables seems to be a better option mathematically and theoretically, multiple model number does not harm the computation speed and prediction performance of NB.

- **Target variable definition**: As targets, latent variables indicating congestion and incidents state in the future need have a different temporal definition because incidents are much rarer than congestion. Congestion target variables are indicating the congestion state in a specific future time point. While incidents target variables are predicting whether any incident will happen in one hour. Such incident variable definition makes prediction possible.

- **Feature variable selection**: The selection is mainly about selecting links for speed data. This study selects bottlenecks that have high volume and congestion time because they can map the general traffic status of the whole city. Such selection is satisfying to NB because it has a good performance both in computation speed and prediction performance.

6.2. Validation of NB Model

The validation procedures also provide us some findings that can be used in future studies.

- **About accuracy rate**: more than 95% of congestion and incident records have state “no alarm”. So they are skewed classes cannot be assessed by accuracy rate in classification validation.

- **K-fold validation**: usually K-fold validation is used to directly calculate the average validation parameters among k parts. But this study uses it to generate prediction values, which are used for validation parameters calculation with actual values.

- **Congestion prediction validation**: the mean value of precision, recall, and F1 score are respectively 0.56, 0.85, and 0.66. All validation parameters decrease as the prediction time increases.

- **Incident prediction validation**: compared to congestion, incidents are much harder to predict. The mean value of precision, recall, and F1 score are respectively 0.02, 0.29, and 0.04. Stationary vehicle prediction has the best performance among all kinds of incidents. Crashes go second and hazard is the worst.

6.3. Limitations and further studies

Some limitations of the current study, plans for addressing those limitations, and other future research directions are discussed below.

- **Classification threshold**: currently, a state is selected in classification process when it has the highest probability in responding target variable. Such a selection criterion is equivalent to setting classification threshold at 50% because all target
variables are binary. By selecting appropriate threshold, a better prediction performance can be achieved from the exact same model. For example, if the threshold decreases to 40%, precision may decrease by 0.1 and recall increase by 0.2, which may be satisfying to traffic manager.

Such a threshold tuning process is not included in this study. Future study may select a better threshold for all models or each of them by considering the need of traffic management and precision-recall trade off weight.

- **Continuous variable**: this study uses discrete data for all variables, it causes information lost. Future study may use continuous data as much as possible (some variables are originally discrete, such as day of week).
- **Other model forms**: NB is a very “old” model form. Future study will try some more advanced models to further improve the prediction performance. The possible options are augmented Naive Bayes, continuous variable normal BN, and Neuron Network using deep learning techniques. They will be validated and compared together with NB.

7. Conclusion

This study proposes a Naive Bayes (NB) classifier model for predicting congestion and incidents in road networks. The model was calibrated and tested using traffic, incident and weather data from the Brisbane area. The validation results show that proposed models can successfully predict congestion and incident occurrence with a desired level of accuracy. The methodology for building and validating a NB classifier model described in this study is general and flexible, allowing the NB-based traffic prediction framework to be easily applicable to other regions and cities. The proposed model could be easily integrated in real-time traffic and incident management systems, where the model is built for each target link offline (with parameters learned from historical data) and the prediction of congestion or incident can be performed online by updating the values of feature variables based on real-time traffic, weather, time-of-day information. Further studies will include the tuning of classification threshold, the application of continuous variables and other related model forms.

Reference


