



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

공학석사학위논문

Performance Evaluation of Classifiers by Machine Algorithm at Freeway On-Ramp

기계 알고리즘을 적용한 고속도로 합류부
차로변경 분류기의 성능평가

2017년 8월

서울대학교 대학원

건설환경공학부

우 동 준

Performance Evaluation of Classifiers by Machine Algorithm at Freeway On-Ramp

기계 알고리즘을 적용한 고속도로 합류부
차로변경 분류기의 성능평가

지도 교수 이 청 원

이 논문을 공학석사 학위논문으로 제출함

2017년 6월

서울대학교 대학원

건설환경공학부

우 동 준

우동준의 공학석사 학위논문을 인준함

2017년 6월

위 원 장 _____ (인)

부위원장 _____ (인)

위 원 _____ (인)

Abstract

Performance Evaluation of Classifiers by Machine Algorithm at Freeway On-Ramp

Woo, Dong-Joon

Department of Civil and Environmental Engineering

The Graduate School

Seoul National University

At the highway merging section, it is very difficult to predict the lane-change decision due to not only the interaction between the flow on the target lane and the flow on the acceleration lane, but also various influencing factors (*e. g.* traffic conditions, geometry, weaving, and individual reactions of merging vehicles to the mainline vehicles) on driver behavior. Also, NGSIM US-101 datasets are inherently imbalanced such as one large “rejected gaps” class and small/rare “accepted gaps” class since a single driver could participate in several non-merge events under same traffic circumstances, but only one merge event. The strategy of this study is threefold to moderate the imbalanced dataset and to improve the classification performance of proposed classifiers for decision of merging characteristics under

MLC(mandatory lane change) circumstance.

Firstly, the data sampling technique for class imbalance problem will be introduced to show the classification performance by using the corresponding contingency matrices and alternative classification metrics based on skill scores and ROC/PR curves. For this purpose, the generalized Hampel filtering available in MATLAB is applied to decrease measurement errors and two simple approaches including the duplicate elimination by averaging and the sampling time interval are considered for data reduction.

Secondly, the non-parametric classifiers based on the machine algorithms of SVM(Support Vector Machine) and EBM(Ensemble Boosting Method) have been presented to improve the classification performance of the lane-changing characteristics at freeway on-ramp as compared with the commonly used parametric classifier by BLM(Binary Logit Model) on the basis of probabilistic function in combination of linear parameters.

Thirdly, the anticipated gap model suggested by Choudhury(2007) is used to include the gap variation due to the dynamic interaction of lead and lag vehicles with respect to a subject vehicle in addition to the conventional adjacent gap since the critical gap has a significant effect on lane-changing behavior.

To extend this study by using the proposed classifiers, the microscopic traffic analysis has been carried out with the True-Positive vehicles classified by contingency matrices. Not only the driver's decision making process is investigated from the vehicle trajectory plotting, but also the classification of merging patterns are illustrated such as direct, chase merging, and others. For this purpose, the K-means clustering algorithm has been adopted to distinguish the trajectory patterns. The error of lateral position of merging vehicles produced by different classifiers has been compared with that by observed data from the comparison of vehicle trajectories according to direct and chase merging patterns. Also, the performance of classifiers is compared in terms of distribution of distance and time error by the data

reduction by sampling time interval.

Detailed vehicle trajectory data from the Next Generation Simulation (NGSIM) dataset are used for model development and testing (US Highway 101). It may be concluded that non-parametric classifiers based on machine algorithm show a better prediction than conventional parametric model for lane-changing vehicles at the merging location regardless of the imbalanced NGSIM dataset. It is also known that the data resampling techniques and the anticipated gap model have a great effect on moderating the imbalanced dataset as well as improving the data quality.

Keyword : SVM, EBM, Lane-Change Prediction, Data Under-Sampling, Anticipated Gap Model, K-means Clustering, Hampel Filter, Decision Making Process

Student Number : 2015-21300

Table of Contents

Abstract.....	i
List of Figures.....	v
List of Tables.....	vii
List of Abbreviations.....	viii
Nomenclatures	ix
Chapter 1. Introduction.....	1
1.1 Motivation.....	1
1.2 Data Collection.....	3
1.3 Objective.....	5
1.4 Research Outline.....	6
Chapter 2. Classifiers for Prediction Model.....	9
2.1 General.....	9
2.2 Binary Logit Model(BLM).....	9
2.3 Support Vector Machine(SVM).....	11
2.4 Ensemble Boosting Method(EBM).....	13
Chapter 3. Data Resampling for Class Imbalance Problem.....	15
3.1 General.....	15
3.2 Data Processing by Hampel Filter.....	16
3.3 Data Under-sampling Technique.....	18
3.3.1 Data Reduction by Sampling Time Interval.....	18
3.3.2 Duplicate Elimination by Averaging.....	19

3.4 K-means Clustering.....	21
Chapter 4. Metrics for Classification Performance	26
4.1 Contingency Matrix.....	26
4.2 Skill Scores.....	27
4.3 ROC and PR Curves.....	29
Chapter 5. Lane-Change Characteristics.....	32
5.1 Anticipated Gap Model.....	32
5.2 Merging Pattern.....	34
5.3 Decision Making Process	36
Chapter 6. Numerical Results.....	38
6.1 Performance Evaluation of Classifiers by Duplicate Elimination...38	
6.2 Performance Evaluation of Classifiers by Sampling Time Interval.43	
6.3 Decision-Making Process by Vehicle Trajectory.....45	
6.4 Classification of Merging Patterns	58
Chapter 7. Conclusions.....	62
References.....	65

List of Figures

Figure 1.1 The road geometry of US-101 site, Los Angeles California.....	4
Figure 1.2 Schematic diagram for research outline.....	7
Figure 2.1 Basic concept of SVM.....	13
Figure 2.2 Schematic diagram of EBM.....	14
Figure 3.1 Illustration of under-sampling technique.....	15
Figure 3.2 Type of measurement errors.....	17
Figure 3.3 The Silhouette diagram shows how well the data are separated into five clusters from US-101 data.....	23
Figure 3.4 Vehicle trajectories of five clusters by averaging value using US-101 data.....	24
Figure 3.5 Representation of vehicle trajectories of five clusters by piecewise linear fitting using US-101 data.....	25
Figure 4.1 The difference between comparing algorithms in ROC vs. PR space....	30
Figure 5.1 Vehicle relationships in a merging situation (Choudhury, 2007).....	34
Figure 5.2 Data collection site and merging patterns (Chu, 2014).....	36
Figure 5.3 An illustration of how to define the begin/end time points of DLC trajectory in US-101 site (Wang <i>et al.</i> , 2014).....	37
Figure 6.1 ROC and PR curves with adjacent gap model before data under-sampling.....	42
Figure 6.2 ROC and PR curves with anticipated gap model after data under-sampling.....	43
Figure 6.3 Distribution of distance and time errors after data reduction using 1 sec	

time interval.....	44
Figure 6.4 Classification of trajectory patterns by K-means clustering.....	47
Figure 6.5 Lateral position and discretized lateral velocity for Cluster I (direct merging).....	48
Figure 6.6 Lateral position and discretized lateral velocity for Cluster II (chase merging).....	49
Figure 6.7 Comparison of vehicle trajectories between direct and chase merging patterns according to prediction models.....	50
Figure 6.8 Distribution of distance and time errors for Cluster I.....	54
Figure 6.9 Distribution of distance and time errors for Cluster II.....	55
Figure 6.10 Distribution of distance and time errors for Cluster III.....	56
Figure 6.11 Relation between gap and distance to end of acceleration lane by observed model using accepted vehicles in contingency matrix.....	60
Figure 6.12 Relation between gap and distance to end of acceleration lane by BLM using True-Positive vehicles in contingency matrix.....	60
Figure 6.13 Relation between gap and distance to end of acceleration lane by SVM using True-Positive vehicles in contingency matrix.....	61
Figure 6.14 Relation between gap and distance to end of acceleration lane by EBM using True-Positive vehicles in contingency matrix.....	61

List of Tables

Table 3.1 Distribution of merging vehicle types in US-101 site.....	17
Table 3.2 Data Reduction by Sampling Time Interval (e.g. 3sec).....	19
Table 3.3 Duplicate elimination of redundant rows by averaging.....	20
Table 3.4 Number of vehicles according to five clusters in US-101 site.....	24
Table 4.1 Contingency Matrix between observed and predicted data.....	26
Table 4.2 Properties of AUROC.....	31
Table 6.1 Contingency matrix of BLM using adjacent gap model before data under-sampling.....	38
Table 6.2 Contingency matrix of SVM using adjacent gap model before data under-sampling.....	39
Table 6.3 Contingency matrix of EBM using adjacent gap model before data under-sampling.....	39
Table 6.4 Contingency matrix of BLM using anticipated gap model after data under-sampling.....	39
Table 6.5 Contingency matrix of SVM using anticipated gap model after data under-sampling.....	39
Table 6.6 Contingency matrix of EBM using anticipated gap model after data under-sampling.....	40
Table 6.7 Skill scores with adjacent gap model before data under-sampling.....	40
Table 6.8 Skill scores with anticipated gap model after data under-sampling.....	41
Table 6.9 Percent of correctly predicted data by prediction models.....	45
Table 6.10 Errors of prediction models with respect to the observed model.....	52

Table 6.11 Number of True-Positive(TP) vehicles by prediction models.....52

Table 6.12 Percent of correctly predicted data by prediction models for Cluster I.....57

Table 6.13 Percent of correctly predicted data by prediction models for Cluster II.....57

Table 6.14 Percent of correctly predicted data by prediction models for Cluster III....57

List of Abbreviations

AC	Accuracy
AUPR	Area Under the Precision-Recall curve(AUPR)
AUROC	Area Under the Receiver Operating Characteristic curve
BLM	Binary Logit Model
DLC	Discretionary Lane Change
EBM	Ensemble Boosting Method
FAR	False Alarm Ratio
FN	False Negative
FP	False Positive
FPR	False Positive Rate
MLC	Mandatory Lane Changing
MTSM	Microscopic Traffic Simulation Models
NGSIM	Next Generation Simulation
OSH	Optimal Separating Hyperplane
P	Precision
PF	Putative Following vehicle
PL	Putative Lead vehicle
POD	Probability of Detection
PR	Precision-Recall
ROC	Receiver Operator Characteristic
SEMA	Symmetric Exponential Moving Average filter
SVM	Support Vector Machine

TN	True Negative
TNR	True Negative Rate(or Specialty)
TP	True Positive
TPR	True Positive Rate(Sensitivity or Recall)

Nomenclatures

$d(X_i, X_j)$	Euclidean distance of two trajectories
U_i	desirability of choosing particular alternative
α	constant
X_1, X_2, \dots, X_n	variables that influence decision of driver
$\beta_1, \beta_2, \dots, \beta_n$	corresponding coefficients
ξ_i	non-negative slack variable
L_n	length of vehicle
G_{nt}^{lead}	available lead gap
G_{nt}^{lag}	available lag gap
v_{nt}^{lead}	speed of lead vehicle
v_{nt}^{lag}	speed of lag vehicle
a_{nt}^{lead}	acceleration of lead vehicle
a_{nt}^{lag}	acceleration of lag vehicle
$v_{nt}^{subject}$	speed of subject vehicle
$a_{nt}^{subject}$	acceleration of subject vehicle
τ_n	anticipation time

$\tilde{G}_{nt}^{adj}(\tau_n)$ adjacent gap

$\tilde{G}_{nt}^{anti}(\tau_n)$ anticipated gap

Chapter 1. Introduction

1.1 Motivation

The merging section is a key point on the expressway networks. It is regarded as a potential bottleneck and a source of traffic crashes due to the competition of two traffic flows for the same space. Especially in congested situations, the conventional acceptable gaps are often not available and more complicated merging characteristics have been observed. Due to the interaction between the flow on the target lane and the flow on the acceleration lane, drivers merge through courtesy of the lag (or following) driver in the target lane or decide to force in and compel the lag driver to slow down. Therefore, its operations in terms of efficiency and safety are becoming increasingly important concerns. In the last few decades, the microscopic traffic simulation models (MTSMs) have been widely used as effective tools to evaluate the operational policy or new geometric design in terms of efficiency and/or safety of traffic facilities including the merging sections. However, to get reasonable evaluation results, it is very important to take into account of various influencing factors (*e. g.* traffic conditions, geometry, and individual reactions of merging vehicles to the mainline vehicles) on driver behavior for providing a more realistic resampling of traffic operation. Unfortunately, at merging sections, the existing simulation models cannot precisely represent driver behavior under those influencing factors. This study is only focused on the development of prediction models to detect merging vehicles at freeway on-ramp by using the machine algorithms including SVM (support vector machine) and EBM (ensemble boosting method). To improve the prediction power, the data under-sampling technique (Mandalia and Salvucci, 2005) is considered to moderate the imbalanced dataset obtained from NGSIM US-101 observed data. Also, the combined critical gap model, recently proposed by Choudhury(2007), will be used to classify merging

patterns into direct, yield and chase merging as explained in Chapter 4 since the gap acceptance theory is very important to predict the lane-changing decision.

Recent critical gap models (Choudhury, 2007, Marczak, 2013) have been developed not only to distinguish between normal, courtesy, and forced merging, but also to propose the combined critical gap model. In other words, the critical gap has been split into a lead gap and a lag gap, that is, a gap between the PL(putative lead vehicle) and the merging vehicle and a gap between the merging vehicle and the PF(putative following vehicle), respectively. The proposed Choudhury's model is deemed to the best reliable gap model up to now considering details of driver behaviors that will be used in this work.

Also, datasets for many classification problems, especially for the lane-changing decision at merging sections are inherently imbalanced such as one large "rejected gaps" class and small/rare "accepted gaps" class in the contingency matrix defined in Table 4.1. A contingency matrix of size $n \times n$ (Kohavi and Provost, 1998) associated with a classifier shows the predicted and actual classification, where the number of classes denoted by n is fixed as 2 associated with "rejected gap" and "accepted gap". However, the most commonly used classification algorithms do not work well for such problems because they aim to minimize the overall error rate, rather than paying special attention to reduce the "rejected gap" class (Fatourechi *et al.*, 2008). Imbalanced datasets may cause the fraud detection of lane-change.

For instance, the conventional binary logit model(BLM) is recognized as a parametric model as well as a linear model to quantify the influencing factors on the probability whether drivers accept or reject a certain gap. Most of investigators have used the "accuracy" rate to evaluate the classification performance. According to this performance measure, the predictive power of BLM generally shows more or less than 90% of accuracy regardless of imbalanced datasets. However, the "accuracy" rate determined using Eq. (4.1) may not be an adequate performance measure when the number of "rejected cases" or "negative cases" is much greater

than the number of “accepted cases” or “positive cases” (Kubat *et al.*, 1997). Due to this reason, various skill scores or performance measures have been presented to estimate the prediction of detect.

In this study, eight skill scores including “accuracy” index as well as ROC/PR curves are used to evaluate the performance of classifiers or prediction models to classify the “rejected cases” or “accepted cases”. At the same time, we try to make an effort to reduce the number of “rejected gaps”. For example, if two time frames for an identical subject vehicle under the same circumstance have the same observed class as “rejected case”, then they are redundant and one of them can be dropped. Such analysis is performed regardless of the classification method. For this purpose, the data resampling technique (Cano *et al.*, 2006; Mandalia and Salvucci, 2005) was presented to reduce the “negative cases” as deemed to be one of filtering method of imbalanced datasets.

Support vector machine (SVM) and Ensemble boosting method (EBM) belong to non-parametric model or machine algorithms that can perform binary classification or pattern recognition tasks. The motivation behind using machine algorithms in analyzing gap acceptance is its advantages over some other classification techniques like BLM, such as (a) requires less training data, (b) is able to produce nonlinear models, (c) is insensitive to imbalanced datasets, and (d) has a better generalization performance (Pawar *et al.*, 2015).

1.2 Data Collection

The observed data have been collected by a dataset of vehicle trajectory data completed as part of the Federal Highway Administration’s(FHWA) Next Generation Simulation (NGSIM, 2005, 2006) project. According to discussion by Thieman *et al.* (2008) and Punzo *et al.* (2011), the trajectory data from US-101 site have the best accuracy and consistency as compared with other three datasets in NGSIM. The data analyzed in this study represent vehicle trajectories on a segment

of U.S. Highway 101(Hollywood Freeway) in Los Angeles, California collected between 7:50 a.m. and 8:35 a.m. on June 15, 2005. The data was collected using video cameras mounted on a 36-story building, 10 University City Plaza, which is located adjacent to the U.S. Highway 101 and Lankershim Boulevard interchange in the Universal City neighborhood. The site was approximately 604 m in length, with five mainline lanes throughout the section. An auxiliary lane is present through a portion of the corridor between the on-ramp at Ventura Boulevard and off-ramp at Cahuenga Boulevard. Lane numbering is incremented from the left-most lane. Video data were collected using eight video cameras, cameras 1 through 8, with camera 1 recording the southernmost, and camera 8 recording the northernmost section of the study area, as shown in Figure 1.1. Complete vehicle trajectories were transcribed at a resolution of 10 frames per second. A total of 45 minutes vehicle trajectories are being processed from the video data collected. These data have been divided into three 15-minute periods for processing and analysis to identify whether or not congested conditions. Periods of the first 15 min. and the remaining 30 min denote transition condition and congested condition, respectively. Also, the US-101 study area is located between on-ramp and off-ramp as shown in Fig 1.1 that causes very complicated merging patterns due to the weaving phenomenon. Vehicles are classified into three categories: (1) motorcycle, (2) automobile, and (3) truck and buses.

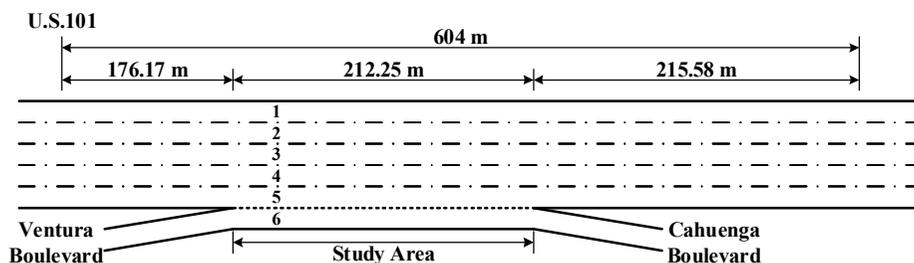


Figure 1.1 The road geometry of US-101 site, Los Angeles California.

1.3 Objective

Many practical classification problems, especially for decision of lane-changing prediction at the merging section are “imbalanced”, i.e., at least one of the classes constitutes only a very small minority of the data. Consider a classifier designed for classifying two classes, “accepted(positive) case” and “rejected(negative) case”. Assume that the “rejected” class is the class with the majority of test samples and the “accepted” class is the class with the minority test samples. In this study, the number of “rejected gaps”(non-merging cases) from the observed NGSIM US-101 dataset is much greater than the number of “accepted gaps”(merging cases). For such a problem, the interest usually leans towards correct classification of the “rare” class (which will refer to as the “accepted(positive) case”. This correct classification is called TP(True-Positive) and is denoted by (1,1) as defined in Table 4.1. However, the most commonly used classification algorithms do not work well for such kind of problems because they aim to minimize the overall error rate, rather than paying attention to the “accepted(positive) class”.

Thus many researchers provided comparative study and analysis of classification techniques in the field of machine learning and data mining to moderate the imbalanced dataset. However, the application of machine algorithms to classifiers for lane-changing behaviors is very limited. Thus the parametric model based on BLM(Binary Logit Model) and non-parametric models including SVM(Support Vector Machine) and EBM(Ensemble Boosting Method) are used for prediction of merging behavior at freeway on-ramp.

The aim of this study is threefold to improve the classification performance of prediction models based on machine algorithm(SVM and EBM) for decision of merging characteristics at the merging section. Firstly, the effect of data resampling technique will be tested to show the classification performance by using the corresponding contingency matrices and alternative classification metrics based on skill scores and ROC/PR curves with respect to different prediction models.

Secondly, the anticipated gap model will be used whether this model may affect the better prediction of merging at freeway on-ramp as compared with the conventional adjacent gap model. Thirdly, the decision making process and merging characteristics will be investigated by using the vehicle trajectories of merging vehicles through the process of K-means clustering.

1.4 Research Outline

To improve the classification performance of imbalanced dataset, three strategies are presented such as data resampling technique, applying machine algorithms based on SVM(support vector machine) and EBM(ensemble boosting method), and the modified gap acceptance model as shown in Figure 1.2.

(1) Data Resampling Technique

It is noted that a single driver could participate in several non-merge events under same traffic circumstances, but only one merge event. This causes extremely imbalanced dataset. To solve this problem, two common approaches are considered on the basis of under-sampling concept; one is simple averaging method for same events, and the other is the under-sampling by sampling interval increased from 1 sec to 5 sec.

(2) Machine Algorithm using SVM and EBM

There are two categories to decide the decision of vehicle merging such as parametric approach as well as non-parametric approach. The BLM(binary logit model) is a commonly used parametric algorithm in lane-changing problem on the basis of probabilistic function in combination of linear parameters. The examples belong to parametric approach are Naïve Bayes, Gaussian discriminant analysis(GDA), Hidden Markov model(HMM) and

Probabilistic graphical model, *etc.* On the other hand, non-parametric algorithms are recently used to improve the performance of a system with experience or sample data at some task in lane-changing problems. In this study, SVM(Support Vector Machine) and EBM(Ensemble Boosting Method) have been adopted for better classification of merge events.

(3) Anticipated Gap Model

Recently, Choudhury(2007) suggested the anticipated gap to include the gap variation in addition to the adjacent gap. However, the driver cannot merge into the target lane in some cases even though the adjacent gap is acceptable. If we trace those vehicles, the gap between subject vehicle and lag vehicle is not enough to merge. Thus, it is necessary to split the adjacent gaps into more precise gaps, called “anticipated gap”, considering the interaction of lead and lag vehicles with respect to a subject vehicle. In this study, the anticipated gap model(Choudhury, 2007) has been used for this reason.

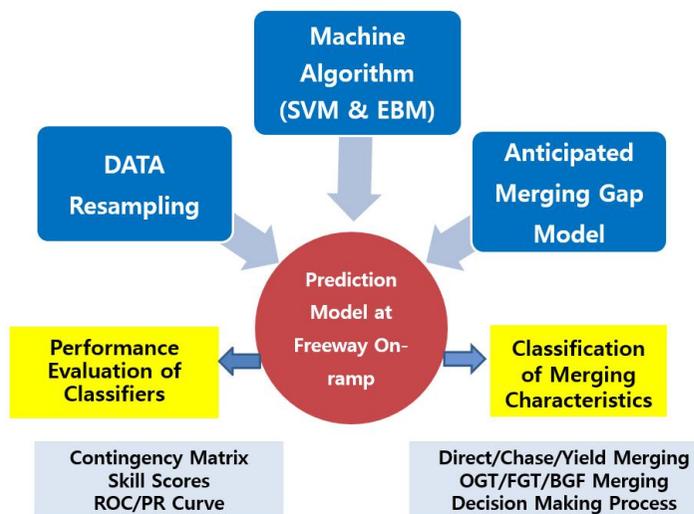


Figure 1.2 Schematic diagram for research outline.

On the basis of three strategies, this study aims to show the classification performance of SVM and EBM as compared with the conventional BLM. For this purpose, various classification metrics based on contingency matrix will be presented by skill scores and ROC and PR curves. In addition these, the decision making process will be investigated by using vehicle trajectories in terms of lateral position of velocity of merging vehicles after Hampel filtering of raw NGSIM dataset and K-means clustering to distinguish the merging patterns.

Chapter 2. Classifiers for Prediction Model

2.1 General

The classifier models for lane-changing expectation can be classified into two categories such as parametric model and non-parametric model. Parametric model assumes some finite set of parameters Θ . Given parameters, future predictions denoted by x are independent of observed data, D : $P(x|\Theta, D) = P(x|\Theta)$. So, the complexity of the model is bounded even if the amount of data is unbounded. This makes them not very flexible. The typical model is BLM(binary logit model). On the other hand, non-parametric model assume that the data distribution cannot be defined in terms of such a finite set of parameters. But they can often be defined by assuming an infinite dimensional Θ . The amount of information that Θ can capture about the data D can grow as the amount of data grows. This makes them more flexible. This model can be used with non-normally distributed data as well as discrete data. Generally, machine algorithms are deemed as non-parametric model such as SVM(support vector machine) and EBM(ensemble boosting method) (Kumar and Sahoo, 2012).

2.2 Binary Logit Model(BLM)

A BLM is recognized as an important modeling tool for studying discrete choices that is a linear model to quantify the influencing factors on the probability whether drivers accept or reject a certain gap. It can be considered as a soft classifier that classifies events according to the estimated class conditional probabilities. Based on these gaps we construct a binary variable which equals 1 when an offered gap is

accepted and 0 when the gap is rejected. The following variables are extracted from datasets such as position of the vehicle on the acceleration lane at the moment a gap is offered, offered gap length, positions of the putative leader and the putative follower, speed difference of merging vehicle and putative follower and speed difference of putative leader and putative follower. Using these variables, we apply an explanatory statistical method, the so-called Principal Component Analysis (PCA), to find the correlation between all variables extracted from the datasets.

The utility of choosing an alternative depends on various factors. A linear-utility expression can be defined in Eq. (2.1) (Ben-Akiva , 1985):

$$U_i = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_n X_n \quad (2.1)$$

Where

- $i = 1, 2, \dots, n$ indicates various alternatives;
- U_i = desirability of choosing particular alternative;
- α = constant;
- X_1, X_2, \dots, X_n = variables that influence decision of driver;
- $\beta_1, \beta_2, \dots, \beta_n$ = corresponding coefficients.

Many models are available to transform the utility function to obtain alternative specific probability. The binary logit model to estimate the probability of choosing an alternative i by the driver is given by Eq. (2.2).

$$P(i) = \frac{1}{1 + e^{-(\alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_n X_n)}} \quad (2.2)$$

The maximum-likelihood method is used for evaluating the coefficients of the utility expression by an iterative search process throughout the dataset.

2.3 Support Vector Machine(SVM)

SVMs do not require any assumption about the distribution of the data being analyzed as these are supervised nonparametric statistical learning algorithms (Cotes, 1995; Vapnik, 1998). SVMs fit an optimal separating hyperplane(OSH) to the underlying data by which the data can be grouped into two classes. In Figure 2, the two-dimensional data are non-separable, and an OSH groups the data into two classes. The OSH is obtained by maximizing the margin between the OSH and the closest training samples, called the “support vectors”(Vapnik, 1998; Burge, 1998). The resulting maximum margin hyperplane has the maximum separation between the decision classes. If the data are linearly non-separable, a linear hyperplane can be fitted by mapping the data into a high-dimensional space(Scholkopf and Smola, 2002). The basic linear SVM learning decision rules are given by $f(x) = w \cdot x + b$ where w is the weight vector and b is a bias. $f(x)$ is the discriminant function associated with the hyperplane. Training data D is a set of n points of the form in Eq. (2.3).

$$D = \{(x_i, y_i) \mid x_i \in R^d, y_i \in \{-1, +1\}\} \quad (2.3)$$

where y_i is either -1 or +1, to which the point x_i belongs in a d -dimensional feature space, R^d . The distance between the optimal separating hyperplane and the origin is given by $|b| / \|w\|$. If the training data are linearly separable, two hyperplanes can be selected that separate the data such that there are no data points between them. The region bounded by two hyperplanes is called the “margin” as shown in Figure 2.1. These two hyperplanes, which are parallel to the OSH can be described by the equations $w \cdot x - b = +1$ and $w \cdot x - b = -1$. For the datasets that cannot be separated cleanly, Cortes(1995) and Vapnik(1998) modified the SVM

algorithm by adding a soft margin. The soft margin method chooses the hyperplane that splits the mislabeled examples as cleanly as possible. The margin between the two hyperplanes is given by $2/\|w\|$, thus minimizing that $\|w\|$ will result in maximizing the margin. The OSH is calculated by maximizing the margin of the two hyperplanes and minimizing the error as given by Eq. (2.4).

$$\min_{w,b,\xi} \left\{ \frac{\|w\|^2}{2} + C \sum_{i=0}^n \xi_i \right\}$$

subject to

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i \quad \text{for } i=1, \dots, n \quad (2.4)$$

The factor C and the slack variable ξ_i in Eq. (2.4) are introduced to take non-separable data into account. The shape of the discriminant function in Eq. (2.4) is controlled by constant C by applying a penalty for the samples that are located on the wrong side of the hyperplane. The non-negative slack variable ξ_i measures the degree of misclassification of the data x_i . Through Lagrange dual optimization, the minimization problem in Eq. (2.4) can be solved. The hyperplane with maximum margin can be represented as in Eq. (2.5) in regard to the support vectors.

$$f(x) = \sum \alpha_i y_i k(x_i, x_j) + b \quad (2.5)$$

where $k(x_i, x_j)$ is kernel function, α_i is Lagrange multipliers, and n is a set of support vectors. A kernel function is used to transform the data into a high-dimensional space. Various kernel functions are as follows: linear kernel $k(x_i, x_j) = (x_i \cdot x_j)$, polynomial kernel $k(x_i, x_j) = (x_i \cdot x_j + 1)^d$, and radial basis function $k(x, y) = \exp(-\|x_i - x_j\|^2 / 2\delta^2)$, where d is the degree of the polynomial kernel and δ^2 is the bandwidth of the radial basis function kernel. These functions

can be used for constructing the optimal separating hyperplane(OSH) for different types of nonlinear input data. In the present study, as data were linearly non-separable, the cubic polynomial kernel function has been used.

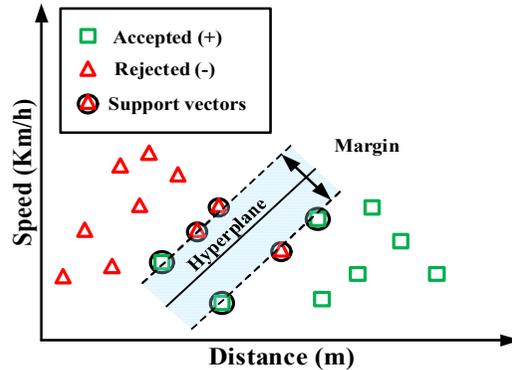


Figure 2.1 Basic concept of SVM.

2.4 Ensemble Boosting Method(EBM)

The ensemble method is to combine the predictions of multiple classifiers. In other words, this method does not learn a single classifier but learn a set of classifiers to achieve more accurate and reliable estimates or decisions than can be obtained from using a single model as shown in Figure 2.2. Ensemble methods can be used for improving the quality and robustness of clustering algorithms(Dimitriadou *et al.*, 2003). Data cases misclassified by earlier classifiers get high weight. Each boosting round learns a new classifier on the weighted dataset by increasing the weight of incorrectly classified dataset. This ensures that they will become more important in the next iterations. These classifiers are weighted to combine them into a single powerful classifier. We stop by using monitoring a cross-validation(Rokach *et al.*, 2014).

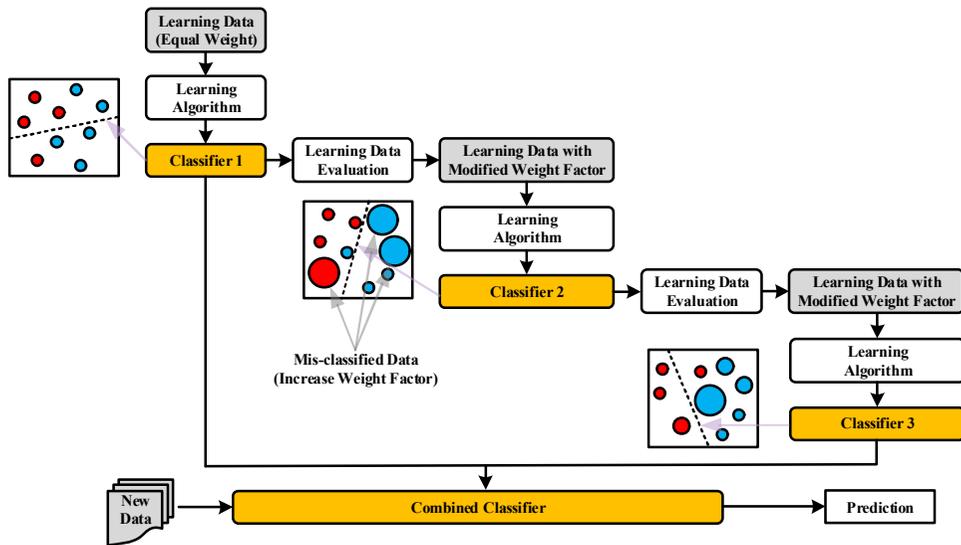


Figure 2.2 Schematic diagram of EBM.

Chapter 3. Data Resampling for Class Imbalance Problem

3.1 General

Data resampling is the process of manipulating the distribution of the training data in an effort to improve the performance of classifiers since there is no guarantee that the training data occur in their optimal distribution in practical problems. The NGSIM dataset is said to present a class imbalance if it contains many more “rejected cases” than “accepted cases”. The problem with class imbalances is that standard learners are often biased towards the majority class. Suppose that there are 99% rejected cases and 1% accepted cases. As a result, the overall accuracy is calculated by 99%. This seems to be biased towards the majority class. Due to this problem, evaluating the performance of a learning system on a class imbalance problem is not done appropriately with the standard “accuracy or error rate” measures. There are several strategies to deal with imbalanced datasets such as over-sampling, under-sampling, cost-sensitive algorithm, add boosting approaches(Sun *et al.*, 2009; Cano *et al.*, 2006; Mandalia and Salvucci, 2005). In this study, the under-sampling techniques based on simple averaging and sampling interval are proposed to reduce the training data as shown in Figure 3.1. In addition to these, the Hampel filtering and K-means clustering are proposed not only to improve the data quality, but also to classify the merging patterns for further studies.

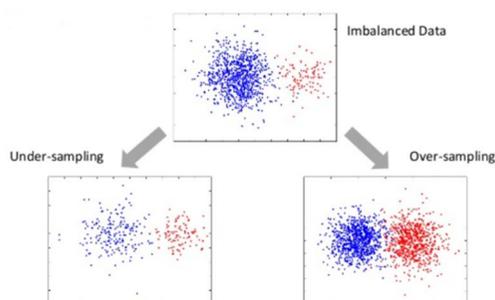
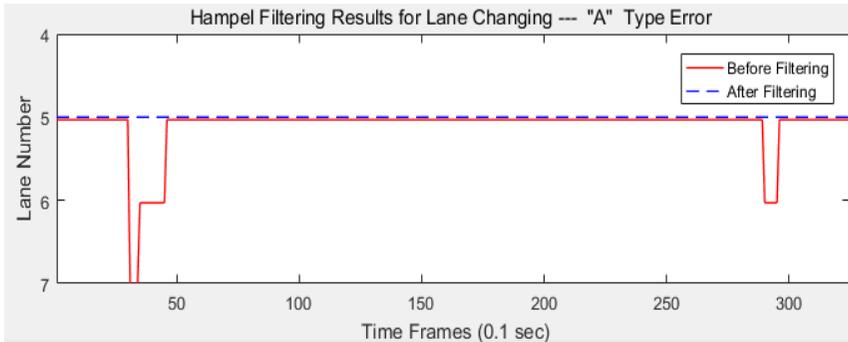


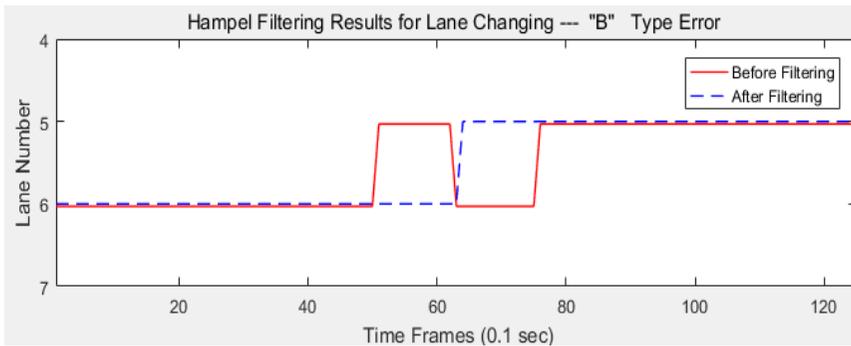
Figure 3.1 Illustration of under-sampling technique.

3.2 Data Processing by Hampel Filter

Before any further data analysis, the following data processing has been carried out: (1) all the merging vehicles from the auxiliary lane 6 to the target lane 5 are sorted that show MLC(mandatory lane change) actions made by on-ramp/off-ramp vehicles; (2) very risky and abnormal vehicles are not considered; (3) the generalized Hampel filter(Pearson, 2016) provided by MATLAB Library is adopted to decrease measurement errors due to the noises introduced by video processing, especially outlier or impulsive noises in the data. Past research studies(Thiemann, 2008; man Punzo, 2011; Hou 2014) have shown that NGSIM speed measurements exhibit noise(random errors). Data smoothing techniques such as moving average, Kalman filtering, and Kalman smoothing have been used to improve data quality. In this study, Hampel filtering method is adopted since this technique is good for smoothing the impulsive and outlier noises(Pearson, 2016). The distribution of vehicle types is shown in Table 3.1 and the sample of Hampel filtering is illustrated in Figure 3.2 to represent type of measurement errors where “A” type and “B” type errors denote impulsive(spike) noise and outlier(oscillation) noise, respectively. After Hampel filtering process with respect to extraordinary or risky lane-changing events of subject vehicles, 365 merging vehicles have been obtained for data analysis. However, 313 merge events are extracted using the moving average filtering with respect to velocity by Hou(2014). On the other hand, 398 valid merging vehicles samples using the symmetric exponential moving average filter(SEMA) were obtained from the US-101 dataset by Wan *et al.*(2016, 2017). It is explained that the difference of merging vehicles is mainly due to the filtering method, time interval of data sorting, and removal of extraordinary or risky errors dependent on investigators.



(a) spike error(impulsive noise)



(b) oscillation error(outlier noise)

Figure 3.2 Type of measurement errors.

Table 3.1 Distribution of merging vehicle types in US-101 site.

Vehicle Type	No. of Vehicles	
	Before Hampel Filtering	After Hampel Filtering
Motorcycle	5	5
Automobile	364	355
Truck and Bus	7	5
Sum	376	365

3.3 Data Under-sampling Technique

There are two types of data imbalance problems; firstly, intrinsic imbalance due to the nature of data-space and secondly, extrinsic imbalance due to time, storage, and other factors. Solutions to imbalanced learning are proposed by sampling methods, cost-sensitive methods and kernel & active learning methods. If data is imbalanced, the sampling methods are used to modify data distribution and thus to create balanced dataset. In this study, two data under-sampling techniques are proposed to handle the class imbalance problem such as “simple time interval” and “duplicate elimination by averaging”.

3.3.1 Data Reduction by Sampling Time Interval

For NGSIM datasets for merging section, detector tracks the vehicles in units of 0.1 second from the point when it first detects the merging vehicle to the point when the merging vehicle completes its lane-change into mainline. Before finishing the lane-change, merging index is displayed as 0 (“rejected”). However, when the merging vehicle makes the lane-change into mainline, merging index is turned into 1 (“accepted”). Due to this characteristic of NGSIM datasets, an imbalance between the majority of “rejected” cases and the minority of “accepted” cases is occurred, which leads to poor prediction power of the lane-change classifier. Since there may be multiple data observations of “rejected” with respect to single “accepted” case. Accordingly, data reduction method based on “sampling time interval” as a key to solve the imbalance problem. The proposed method for data reduction is conducted as following steps. First, find the point for when merging vehicle completes its lane-change. Second, starting from the point of merging (i.e. when it displays “1”), reduce the number of “rejected” cases(0) by applying different sampling time intervals as shown in Table 3.2. Based on Table 3.2, it shows an example of data reduction based on sampling time interval of 3 seconds, and the portion indicated by blue rows means a portion reduced according to the sampling time interval.

Table 3.2 Data Reduction by Sampling Time Interval (e.g. 3sec).

Time Frame	Vehicle ID (Subject)	Vehicle ID (Lead)	Vehicle ID (Lag)	Merging Index (reject=0; accept=1)	Variables (position, velocity, acceleration, etc.)
-	-	-	-	-	-
t_{i-9}	S	E	D	0	V_1
t_{i-8}	S	F	E	0	W_1
t_{i-7}	S	L	F	0	W_2
t_{i-6}	S	L	F	0	W_3
t_{i-5}	S	L	F	0	X_1
t_{i-4}	S	G	L	0	X_2
t_{i-3}	S	G	L	0	X_3
t_{i-2}	S	H	G	0	Y_1
t_{i-1}	S	H	G	0	Y_2
t_i	S	I	H	1	Y_3
-	-	-	-	-	-

3.3.2 Duplicate Elimination by Averaging

The one-second intervals produce the observed data with different sample sizes for both lane changing and non-lane changing events. During every one-second interval, a driver's behavior has been identified as either merge or no-merge. Merge events occurred when a vehicle's front center point shifts to adjacent target lane. A part of the observed data with one-second intervals is presented in Table 3.3. It is noted that a single driver could participate in several non-merge events, but only one merge event. Consequently, there are too many "rejected cases (non-merge events)" for an identical subject vehicle(ID=S) as compared with "accepted cases (merge events)" that may cause imbalanced datasets in lane-changing problem at the US-101 merging section.

When we have a range of data which contains some duplicate entries and now we want to combine the duplicate data with one average corresponding values, the “duplicate elimination by averaging” is commonly used to handle this problem in EXCEL spreadsheet for storage and recording of data (Deepak *et al.*, 2006).

For instance, the current subject vehicle(ID=S) gives three “rejected” while the putative lead vehicle(ID=L) and the putative lag or following vehicle(ID=F) remain unchanged for $t_{i-7} \leq t \leq t_{i-5}$. In a similar manner, two “rejected cases” are detected under same circumstance with lead vehicle(ID=G,H) and lag vehicle(ID=L,G) for $t_{i-4} \leq t \leq t_{i-3}$ and $t_{i-2} \leq t \leq t_{i-1}$, respectively. Thus for every N “rejected cases”, the under-sampling technique is used for data reduction by considering multiple non-merge events as a single non-merge event. Instead, the corresponding variables for N time frames such as global position of vehicle, speed, and acceleration, time headway, *etc.* are assumed by a single average value.

Table 3.3 Duplicate elimination of redundant rows by averaging.

Time Frame	Vehicle ID (Subject)	Vehicle ID (Lead)	Vehicle ID (Lag)	Merging Index (reject=0; accept=1)	Variables (position, velocity, acceleration, <i>etc.</i>)	Data Under-sampling	
						Reduced Time Frame	Averaged Value
-	-	-	-	-	-	-	-
t_{i-9}	S	E	D	0	T_1	t_{j-5}	T_1
t_{i-8}	S	F	E	0	U_1	t_{j-4}	U_1
t_{i-7}	S	L	F	0	V_1	t_{j-3} $= \frac{t_{i-5} + t_{i-6} + t_{i-7}}{3}$	V_{ave} $= \frac{V_1 + V_2 + V_3}{3}$
t_{i-6}	S	L	F	0	V_2		
t_{i-5}	S	L	F	0	V_3		
t_{i-4}	S	G	L	0	W_1	$t_{j-2} = (t_{i-3} + t_{i-4})/2$	W_{ave} $= \frac{W_1 + W_2}{2}$
t_{i-3}	S	G	L	0	W_2		
t_{i-2}	S	H	G	0	X_1	$t_{j-1} = (t_{i-1} + t_{i-2})/2$	$X_{ave} = \frac{X_1 + X_2}{2}$
t_{i-1}	S	H	G	0	X_2		
t_i	S	I	H	1	Y_1	t_j	Y_1
-	-	-	-	-	-	-	-

In other words, combining duplicate data can be represented by a single average value. This idea, called “duplicate elimination by averaging”, has been commonly used in EXCEL Spreadsheet for storage system and recording(Deepak, 2006). For the first case, three time frames are reduced to a single time frame, and the three corresponding variables can be represented by a single average value, i.e. $V_{ave} = (V_1 + V_2 + V_3)/3$. Similarly, $W_{ave} = (W_1 + W_2)/2$ and $X_{ave} = (X_1 + X_2)/2$ for a single time frame.

3.4 K-means Clustering

In general, clustering is the classification of objects into different groups, or more precisely, the portioning of a dataset into clusters, so that the data in each cluster share some common trait, merging pattern in this study according to some defined distance measure. There are two types of clustering such as hierarchical cluster and partitioned clustering. Partitioned algorithm determines all cluster at once that includes K-means clustering, Fuzzy C-means clustering and QT clustering (Hartigan, 1979). In case of K-means clustering, the distance measure will determine how the similarity of two elements is calculated and it will influence the shape of the clusters. Common distance measures include the Euclidean distance, the Euclidean squared distance and the Manhattan distance. The Euclidean distance is given by;

$$d(x, y) = \sqrt{\sum_{i=1}^N (x_i - y_i)^2} \quad (3.1)$$

1. The Euclidean Squared distance is given by;

$$d_{sq}(x, y) = \sum_{i=1}^N (x_i - y_i)^2 \quad (3.2)$$

2. The Manhattan distance is given by;

$$d(x, y) = \sqrt{\sum_{i=1}^N |x_i - y_i|^2} \quad (3.3)$$

The K-means clustering is an algorithm to group N objects based on attributes into K partitions (or number of clusters), where $K < N$. It assumes that the object attributes form a vector space using distance measure. The Euclidean measure corresponds to the shortest geometric distance between two points. Thus an algorithm for clustering N data points into K disjoint subsets S_j containing data points so as to minimize the objective function J based on sum-of-squares criterion where x_n is a vector representing the n^{th} data point and μ_j is the geometric centroid of data points in S_j . Each cluster is associated with a centroid. The centroid can be calculated the average of data points of S_j .

$$J = \sum_{j=1}^K \sum_{n \in S_j} |x_n - \mu_j|^2 \quad (3.4)$$

In case of the K-means clustering algorithm (Hartigan, 1979) for lane-changing problem, each cluster of vehicle trajectory is represented by the center of the cluster and the algorithm converges to stable centroids of clusters. Here ‘‘K’’ stands for number of clusters. In principle, optimal partition can be achieved by minimizing the sum of squared Euclidean distance in each cluster as described above. Suppose the i -th trajectory segment is denoted as a vector $X_i = [x_i(1), x_i(2), \dots, x_i(N)]$ where $x_i(k)$ denotes the lateral position of the i -th vehicle at the k -th time point. Thus the Euclidean distance of two trajectory segments is calculated by Eq. (3.5). This distance measure will be small, if the two trajectories are similar each other.

$$d(X_i, X_j) = \sum_{k=1}^N [x_i(k) - x_j(k)]^2 \quad (3.5)$$

The following figures illustrate the K-means algorithm on a 2-dimensional dataset. The positive part of the Silhouette diagram in Figure 3.3 shows that there is a clear separation of the points between the clusters. On the other hand, the negative parts denote a conflict.

For instance, the vehicle trajectories can be classified into 5 clusters when $K=5$. The corresponding silhouette diagram is illustrated in Figure 3.3. The variations of lateral position of the vehicles have been plotted with respect to the truncated time interval of vehicle trajectories ranging from -10 sec to +10 sec in reference to the merging point as shown in Figs. 3.4-3.5. Since the original NGSIM data are produced by 0.1 sec time interval, the truncated time interval of vehicle trajectories varies from 0 sec to 20 sec. The positions of the crossed lane boundaries are shifted to 0 for all trajectories in order to set the merging point to be located at the middle point of vehicle trajectories. The lateral positions obtained by averaging are illustrated in Figure 3.4 according to five cluster types. To find the break point easily, Figure 3.5 has been represented by the piecewise linear fitting using Figure 3.4. It is noted that the number of clusters are determined by setting the K-value.

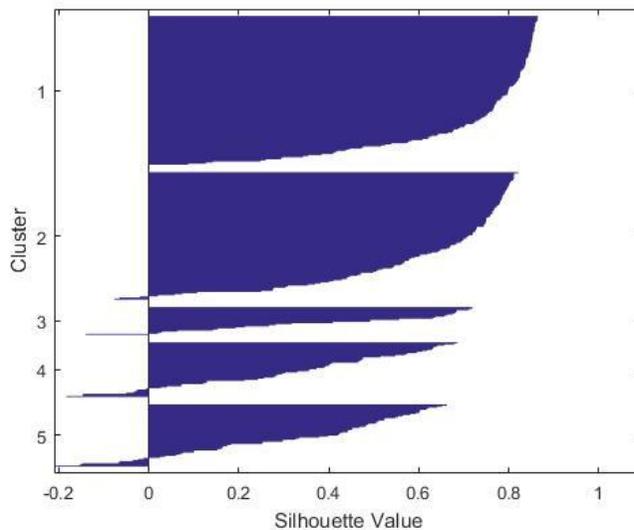


Figure 3.3 The Silhouette diagram shows how well the data are separated into five clusters from US-101 data.

Table 3.4 Number of vehicles according to five clusters in US-101 site.

Cluster Type	I (Direct Merging)	II (Chase Merging)	III	IV	V	Total
No. of Vehicles	139	119	22	40	45	365

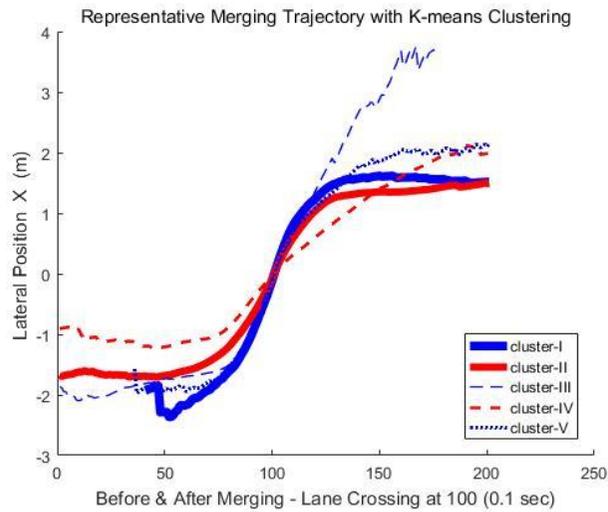


Figure 3.4 Vehicle trajectories of five clusters by averaging value using US-101 data.

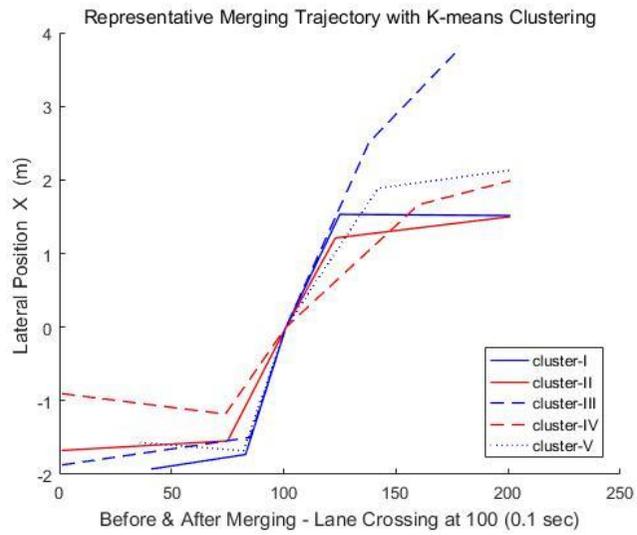


Figure 3.5 Representation of vehicle trajectories of five clusters by piecewise linear fitting using US-101 data.

Chapter 4. Metrics for Classification Performance

4.1 Contingency Matrix

To measure the performance of a gap acceptance prediction algorithm, it is necessary to compute categorical statistics and scalar skill scores according to a “confusion matrix” or “contingency matrix” as shown in Table 4.1. The contingency matrix(Kohavi and Provost, 1998) is a two-dimensional square table that contains information about observed and predicted classifications done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix. The following table shows the contingency matrix for a two class classifier. The entries in the confusion matrix have the following meaning in the context of our study: a is the number of “no” (rejected gap) predictions that matches with the observed “no” (rejected gap), or the number of correct negatives; b is the number of “yes” (accepted gap) predictions when the observation is “no” (rejected gap), or the number of false alarms; c is the number of “no” (rejected gap) predictions when observations are “yes” (accepted gap), or the number of misses; d is the number of “yes” (accepted gap) predictions that matches with the actual “yes” (accepted gap) observations, or the number of hits.

Table 4.1 Contingency Matrix between observed and predicted data.

Observed Class	Predicted Class	
	NO (0)	YES (1)
NO (0)	True Negative(TN) a	False Positive(FP) b
YES (1)	False Negative(FN) c	True Positive(TP) d

4.2 Skill Scores

Since the accuracy measure treats every class as equally important, it may not be suitable for analyzing imbalanced datasets, where the rare class is considered more interesting than the majority class. For binary classification, the rare class is often denoted as the positive class, while the majority class is denoted as the negative class. A contingency matrix that summarizes the number of instances predicted correctly or incorrectly by classification models. The counts in a contingency matrix can also be expressed in terms of various skill scores.

(1) Accuracy(AC)

This is the proportion of the total number of predictions that were correct. The accuracy score may not be an adequate performance measure when the number of negative cases is much greater than the number of positive cases and the classified data (2x2 matrix) are imbalanced (Kubat *et al.*, 1997).

$$AC = \frac{a+d}{a+b+c+d} \quad (4.1)$$

(2) Precision(P)

Precision is the ratio of the number of accepted gaps correctly predicted by SVM to the total number of predicted accepted gaps. The precision value gives the percent of selected gap events that are correct. The formula used for calculating the precision is shown below.

$$P = \frac{d}{b+d} \quad (4.2)$$

(3) Probability of Detection(POD) or Recall

POD is also known as recall, TPR(True Positive Rate), and sensitivity. This is the ratio of the number of accepted gaps correctly predicted by classifiers to the total number of observed accepted gaps. The POD gives the fraction of observed gap events that are correctly forecast. The value of POD ranges from 0 to 1, and POD=1 indicates that the classifier correctly detect all accepted gaps.

$$POD \text{ or Recall} = \frac{d}{c+d} \quad (4.3)$$

(4) TNR(True Negative Rate) or Specialty

This is defined as the fraction of rejected gaps predicted correctly by the model.

$$TNR \text{ or Specialty} = \frac{a}{a+b} \quad (4.4)$$

(5) Bias

The bias is the ratio of the number of predicted accepted gaps to the total number of observed accepted gaps. This value indicates whether the classifiers (or prediction models) underestimate(when bias is less than 1) or overestimate(when bias is greater than 1) the number of accepted gaps.

$$Bias = \frac{b+d}{c+d} \quad (4.5)$$

(6) F-Measure

This is a weighted harmonic mean of the prediction and recall. The F-measure close to 1 indicates the best score; the F-measure close to 0 indicates the worst score.

$$F - Measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4.6)$$

(7) False Alarm Ratio(FAR)

This is the ratio of the number of incorrect gaps predicted by the classifiers to the total number of accepted gaps predicted by the SVM. The FAR value indicates the fraction that the SVM detects accepted lane-changing frequencies that were not detected in the observed accepted lane-changing data. The FAR has a best score of 0 with a range of 0 to 1.

$$FAR = \frac{b}{b+d} \quad (4.7)$$

(8) Heidke's Skill Score

This is popularly used in forecasting since all elements from the contingency matrix are considered. Perfect prediction receives an HSS of 1, a prediction equivalent to the reference prediction receives a score zero, and the predictions worse than the reference prediction receive negative scores.

$$HSS = \frac{2(ad-bc)}{(a+b)(b+d)+(c+d)(a+c)} \quad (4.8)$$

4.3 ROC and PR Curves

In a binary decision problem, a classifier labels cases as either “positive” or “negative”. The decision made by the classifier can be represented in a structure known as a contingency table or confusion matrix. Given the contingency table, we are able to plot the Receiver Operator Characteristic(ROC) curve as well as the Precision-Recall(PR) curve to evaluate an algorithm’s performance. ROC curve plots the true positive rate(TPR, sensitivity, recall) against the false positive rate(FPR). It shows how the number of correctly classified positive cases varied with the number of incorrectly classified negative cases and can also present an overly optimistic view of an algorithm’s performance(Provost *et al.*, 1998). Despite its popularity, the ROC curve has some drawbacks when dealing with highly

skewed or imbalanced datasets, especially when the “negative” cases greatly exceed “positive” cases. On the other hand, the PR(Precision-Recall) curve plots the precision(positive predictive case) against the recall(true positive rate) that has been cited as an alternative to ROC curve for tasks with a large skew in the class distribution(Bockhorst & Craven, 2005; Davis *et al.*, 2006; Keilwagen *et al.*, 2014). An important difference between ROC space and PR space is the visual representation of curves. Looking at PR curves can expose differences between algorithms that are not apparent in ROC space. Sample ROC and PR curves are shown in Figure 4.1. The goal in ROC space is to be in the upper-left-hand corner, and when one looks at the ROC curves in Figure 4.1(a) they appear to be fairly close to optimal. In PR space, the goal is to be in the upper-right-hand corner.

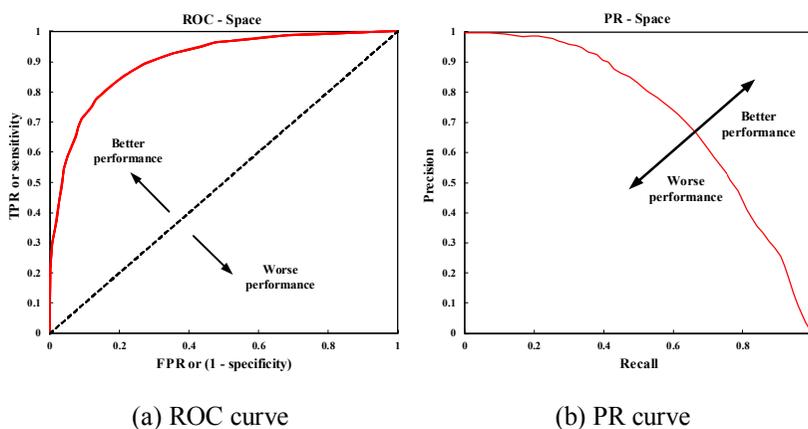


Figure 4.1 The difference between comparing algorithms in ROC vs. PR space.

In order to evaluate classifier performance, we measured both the area under the receiver operating characteristic curve(AUROC) and area under the precision-recall curve(AUPR) of models. Here, we propose a piecewise-defined function allowing to compute the AUROC and AUPR by a sum of integrals where p and r denote

FPR(false positive rate) and sensitivity for AUROC, on the other hand, precision and recall for AUPR, respectively. The evaluation of AUROC and AUPR has been presented in Table 4.2.

$$AUROC \text{ or } AUPR = \int_0^1 p(r)dr \quad (4.9)$$

Table 4.2 Properties of AUROC.

AUROC	1.0	0.9	0.8	0.7	0.6	0.5	<0.5
Prediction	Perfect	Excellent	Good	Mediocre	Poor	Random	Something wrong!

Chapter 5. Lane-Change Characteristics

5.1 Anticipated Gap Model

The various merging behavior models have been proposed on the basis of gap acceptance theory. Many investigators have attempted to define the critical gap, typically minimum acceptable gap. Herman and Weiss (1961), and Miller (1972) were pioneers in the development of gap acceptance models based on critical gap. They assumed that critical gap follows a normal distribution and they used a probabilistic model to estimate. However, not all of these models are applicable for expressway merging sections where drivers have to change lane within limited length of road and where no complete stop situation occurs before a lane-change. At expressway merging sections, several studies have been carried out to model gap acceptance. Kita (1993) made use of a binary logit model and found that the gap length, remaining distance to the end of acceleration lane, and relative speed were significant explanatory variable. However, all of these models are applicable only under uncongested conditions. To overcome these limitations, several models have recently been developed to represent gap acceptance for vehicles merging under congested conditions (Ahmed, 1999; Hidas, 2005). Under congested conditions, where there are few acceptable gaps, they proposed “forced” and “cooperative” lane-change models. These models are capable of representing instances of merging through the creation of gap either by yielding of the following vehicle in the target lane or by forcing the following vehicle to slow down. However, the influence of acceleration lane length on gap acceptance has not been considered. Marczak (2013) claimed that the proceeding researches did not make efforts in observing the rejected gap. For instance, Choudhury (2007) and Kondyli and Elefteriadou (2011) observed gap acceptance, however, they did not take into account of gap rejection. Based on this fact, Marczak (2013) collected video data at two different sites

(Bodegraven in Netherland and Grenoble in France) using the helicopter technique to study the merging maneuvers. They applied the binary logit model(BLM) to develop the gap acceptance for each study site, respectively. The probability of gap acceptance was modeled as a function of the remaining distance, the space gap, the relative speed between (i) PL(putative leading) vehicle and PF(putative following) vehicle, (ii) PL vehicle and merging(or subject) vehicle. They found that the aggressive drivers are influenced by gap distance, remaining distance of acceleration lane and congestion level on the mainline. In other words, the merging section geometry and traffic condition can be important influencing factors.

As we reviewed papers, there are a lot of scenarios for lane-changing characteristics according to driver's behaviors including normal, courtesy and forced merging, and adjacent traffic circumstances between merge lane and target lane. As we aware of it, the gap-acceptance model is based on field observations of adjacent gaps defined in Fig. 5.1. However, the adjacent gap denoted by $\tilde{G}_{nt}^{adj}(\tau_n)$ in Eq. (5.1) may be varied while a subject vehicle is merging into the target lane. If the adjacent gap denoted by $\tilde{G}_{nt}^{adj}(\tau_n)$ is not acceptable to make a normal merge, the merging vehicle evaluates the speed, acceleration and relative position of the passing vehicles in the traffic direction and approximates an expected or anticipated gap that is going to open up after time τ_n . Because of the difference in perception among individuals, the anticipation time τ_n may vary among individuals. Choudhury(2007) proposed the estimated distribution of anticipation time ranging from 0 sec to 4 sec by using the probability density function.

In other words, the courtesy of discourtesy of the lag driver is reflected in the anticipated gap in Eq. (5.2). If the lag driver has decided to provide courtesy to a merging vehicle and has started to decelerate, the anticipated gap increases. If the anticipated gap is unacceptable, the drivers decide whether to force their ways to the target lane compelling the lag driver to slow down or not. This dynamic influence causes the variation of lead and lag gaps. Choudhury(2007) suggested the

anticipated gap denoted by $\tilde{G}_{nt}^{anti}(\tau_n)$ to include the gap variation in addition to the adjacent gap as shown in Eq. (5.2).

$$\tilde{G}_{nt}^{adj}(\tau_n) = G_{nt}^{lead} + G_{nt}^{lag} + L_n \quad (5.1)$$

$$\tilde{G}_{nt}^{anti}(\tau_n) = G_{nt}^{lead} + G_{nt}^{lag} + L_n + \tau_n(v_{nt}^{lead} - v_{nt}^{lag}) + \frac{1}{2}\tau_n^2(a_{nt}^{lead} - a_{nt}^{lag}) \quad (5.2)$$

Where for individual n at time t , L_n is the length of the vehicle, G_{nt}^{lead} and G_{nt}^{lag} are available lead and lag gaps, v_{nt}^{lead} and v_{nt}^{lag} are the speeds of the lead and lag vehicles, a_{nt}^{lead} and a_{nt}^{lag} are the acceleration of the lead and lag vehicles, respectively, as shown in Fig. 5.1.

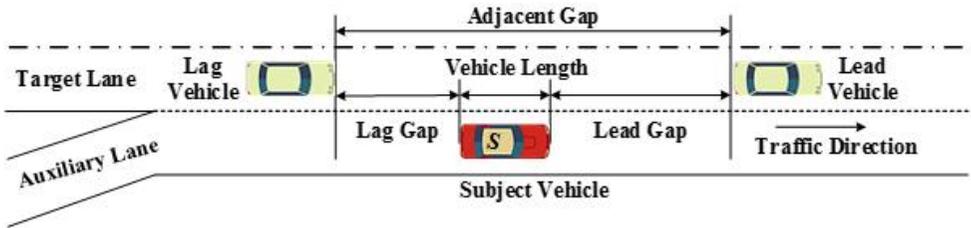


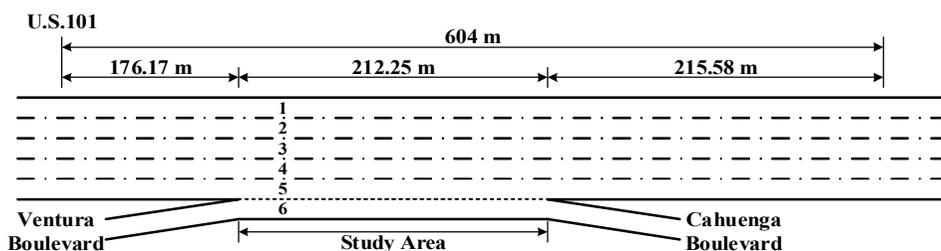
Figure 5.1 Vehicle relationships in a merging situation (Choudhury, 2007).

5.2 Merging Pattern

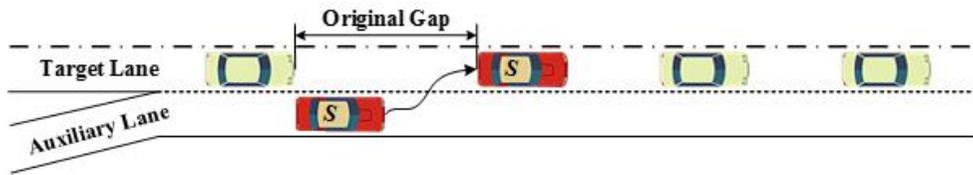
It is very important to take into account of various influencing factors (*e. g.* traffic conditions, geometry, and individual reactions of merging vehicles to the mainline vehicles) on merging pattern for providing a more realistic resampling of traffic operation. Fig. 5.2 illustrates the merging patterns and gap selection scenarios. Merging vehicle taking their original gap as their targeting gap are called original-gap-targeting(OGT) vehicles as shown in Fig. 5.2(b). On the other hand, merging vehicles taking the gap in front of the original gap in Fig.5.2(c) are called forward-gap-targeting(FGT) vehicles (Wan *et al*, 2016).

According to Wan *et al.* (2016, 2017), 398 valid merging vehicles samples were obtained from the US-101 dataset. 242 merging vehicles merged into their “original gap”, or the gap between a PL and a PF that is faced by a merging vehicle when it arrives at the auxiliary lane. 156 merging vehicles did not choose their original gap. Among these, 151 merging vehicles accepted a gap in front of the original gap, and only 5 merging vehicles eventually accepted a gap behind of their original gap as accepted gap. However, the number of backward-gap-targeting(BGT) vehicles were limited to only five. In this study, the number of merging vehicles are detected by 365 after Hampel filtering. Also, 313 merge events were obtained by Hou(2014) using the moving average method. It is expected that the difference of merging vehicles is mainly due to the filtering method, time interval of data sorting, and removal of extraordinary or risky errors dependent on investigators.

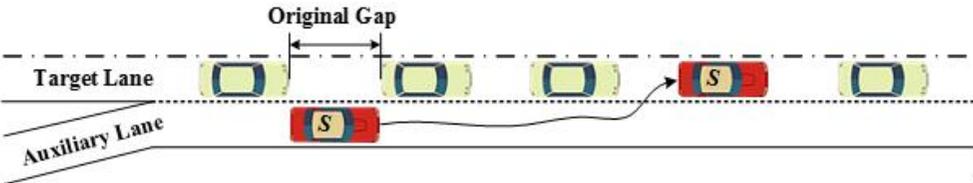
Similarly, Chu(2014) classified the merging patterns into three groups such as direct merging, chase merging and yield merging as shown in Fig. 5.2(b) and 5.2(c). He used the anticipated gap model proposed by Choudhury(2007) as the gap acceptance model. This classification can be corresponded to OGT(original gap targeting), FGT(forward gap targeting) and BGT(backward gap targeting) presented by Wan *et al.* (2016, 2017). If a merging vehicle overtakes the mainline vehicle and chooses FGT, it is called “chase merging”. On the other hand, if a merging vehicle follows OGT and BGT, they are called “direct merging” and “yield merging”, respectively. It is noted that “yield merging” patterns are excluded in this study since the number of vehicles for “yield merging” based on BGT are very limited.



(a) US-101 Site



(b) Direct merging pattern



(c) Chase merging pattern

Figure 5.2 Data collection site and merging patterns (Chu, 2014).

5.3 Decision-Making Process

There are many existing models to describe the decision-making process for lane changing (Hidas, 2002; Sun and Elefteriadou, 2010), but neglected the detailed driving actions. Conventionally, a lane-change was viewed as an instantaneous event with zero time or an event with constant duration time. This simplification is mainly due to the lack of abundant and accurate vehicle trajectories. With the development of GPS and video-based monitoring technology, vehicle trajectories contain rich information on individual driver behaviors and allow inference on interactions between drivers. However, the NGSIM data show that we may draw biased conclusions if raw data were directly used without proper examination. Generally, two important problems must be solved as; firstly filtering out abnormal lane-change actions from the sampled raw data, and secondly definition of begin/end points of a lane-change action. To answer these two problems, the trajectory clustering method using K-means clustering has been used to filter out

some complex abnormal merging patterns, and the begin/end points are derived from the vehicle trajectory for lateral movement and lateral velocity with respect to time.

The begin/end time points of a lane-change are usually defined as “the time instances when the lateral movement of the subject vehicle begins and ends, respectively” (Toledo, 2007). However, the initiation and completion time points of a lateral movement are difficult to determine in practice, because the slopes of the beginning part and ending part of trajectories are very gentle. Wang *et al.*(2014) proposed the how to decide the begin/end points denoted by T_{begin} and T_{end} using DLC(Discretionary Lane Change) trajectories from NGSIM data as shown in Fig. 5.3. From the upper subfigure, it is difficult to find the begin/end time points or breaking points since every normal DLC trajectories are fitted by the fifth-order polynomials. Thus, lower subfigure is plotted to denote the corresponding discretized lateral velocity. They calculated the discretized lateral velocity from the empirical DLC trajectories every 0.5 sec by

$$\bar{v}(t) = \frac{x_i(t) - x_i(t-0.5)}{0.5} \quad (5.3)$$

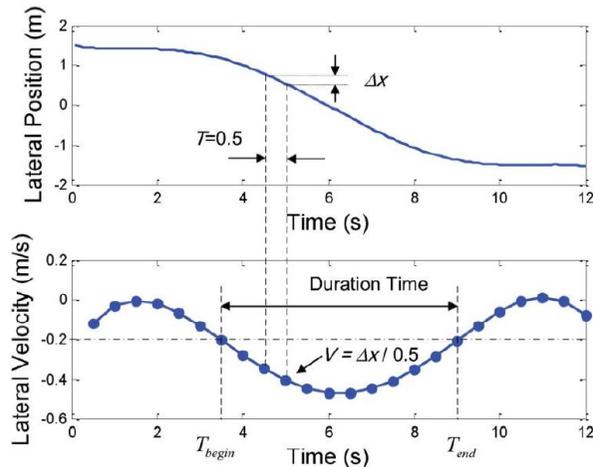


Figure 5.3 An illustration of how to define the begin/end time points of

Chapter 6. Numerical Results

6.1 Performance Evaluation of Classifiers by Duplicate Elimination

The multi-paradigm numerical computing tool, MATLAB, has been used for developing the classifiers considering a data under-sampling as mentioned before. The one-second intervals produce the observed NGSIM US-101 data with different sample sizes for both lane changing and non-lane changing events. The contingency matrices by BLM, SVM and EBM are compared to each other where the cubic polynomial kernel function is used for SVM. Before data under-sampling, there are too many “rejected cases(0,0)” as compared with “accepted cases(1,1)” that may cause imbalanced datasets in lane-changing problem at the US-101 merging section. Nevertheless, the classification by SVM and EBM not only shows better prediction of merging, but also increases the number of actual lane-changing events denoted by (1,1) as shown from Table 6.1 to Table 6.6.

Table 6.1 Contingency matrix of BLM using adjacent gap model before data under-sampling.

Observed Decision	Predicted Decision			Observed(%)
	Rejected	Accepted	Row Total	
Rejected	2145 (84.4%)	33 (1.3%)	2178	85.7%
Accepted	344 (13.5%)	21 (0.8%)	365	14.3%
Column Total	2489	54	2543	100.0%
Predicted(%)	97.9%	2.1%	100.0%	-

Table 6.2 Contingency matrix of SVM using adjacent gap model before data under-sampling.

Observed Decision	Predicted Decision			Observed(%)
	Rejected	Accepted	Row Total	
Rejected	2078 (81.7%)	100 (3.9%)	2178	85.6%
Accepted	280 (11.0%)	85 (3.4%)	365	14.4%
Column Total	2358	185	2543	100.0%
Predicted(%)	92.7%	7.3%	100.0%	-

Table 6.3 Contingency matrix of EBM using adjacent gap model before data under-sampling.

Observed Decision	Predicted Decision			Observed(%)
	Rejected	Accepted	Row Total	
Rejected	2133 (83.9%)	45 (1.8%)	2178	85.7%
Accepted	291 (11.4%)	74 (2.9%)	365	14.3%
Column Total	2424	119	2543	100.0%
Predicted(%)	95.3%	4.7%	100.0%	-

Table 6.4 Contingency matrix of BLM using anticipated gap model after data under-sampling.

Observed Decision	Predicted Decision			Observed(%)
	Rejected	Accepted	Row Total	
Rejected	755 (61.4%)	109 (8.9%)	864	70.3%
Accepted	189 (15.4%)	176 (14.3%)	365	29.7%
Column Total	944	285	1229	100.0%
Predicted(%)	76.8%	23.2%	100.0%	-

Table 6.5 Contingency matrix of SVM using anticipated gap model after data under-sampling.

Observed Decision	Predicted Decision			Observed(%)
	Rejected	Accepted	Row Total	
Rejected	740 (60.2%)	124 (10.1%)	864	70.3%
Accepted	128 (10.4%)	237 (19.3%)	365	29.7%
Column Total	868	361	1229	100.0%
Predicted(%)	70.6%	29.4%	100.0%	-

Table 6.6 Contingency matrix of EBM using anticipated gap model after data under-sampling.

Observed Decision	Predicted Decision			Observed(%)
	Rejected	Accepted	Row Total	
Rejected	737 (60.0%)	127 (10.3%)	864	70.3%
Accepted	100 (8.1%)	265 (21.6%)	365	29.7%
Column Total	837	392	1229	100.0%
Predicted(%)	68.1%	31.9%	100.0%	-

However, it is clearly evident that three prediction models considering the anticipated gap model perform reasonably well after data under-sampling. The corresponding contingency matrices have been moderately balanced. It is noted that the “rejected cases(0,0)” are decreased, on the other hand, “the accepted cases(1,1)” are increased. Table 6.7 and Table 6.8 show the skill scores for three different prediction models.

Table 6.7 Skill scores with adjacent gap model before data under-sampling.

Skill Scores	BLM	SVM	EBM	Remark
Accuracy	0.852	0.851	0.868	The higher, the better. (Improper index for imbalanced dataset)
Precision	0.389	0.459	0.622	The higher, the better.
Bias	0.148	0.507	0.326	Bias>1 : over-estimate Bias<1 : under-estimate Bias=1 : perfect
Sensitivity (or Recall, POD)	0.058	0.233	0.203	1 (best score) 0 (worst score)
FAR (False Alarm Ratio)	0.611	0.541	0.378	0 (best score) 1 (worst score)
F-Measure	0.100	0.309	0.306	1 (best score) 0 (worst score)
HSS (Heidke Skill Score)	0.066	0.235	0.253	if HSS=1 : perfect prediction

It is also known that the skill scores with the anticipated gap model after data under-sampling are much more accurate than those with the conventional adjacent gap model before data under-sampling. The prediction models by SVM and EBM are insensitive to the imbalanced datasets consisting of “rejected cases” and “accepted cases”, and have a very good potential to be an alternative classifier as compared with BLM. In case of under-sampling data, the bias score for SVM is closer to 1, which indicates that the SVM predicts gap acceptance and rejection reasonably well. The recall values show that the EBM has a higher score than BLM and SVM, and the EBM also has a lower FAR(false alarm ratio) score as compared with BLM and SVM where FAR is 0 for the best score. On the whole, the EBM shows better prediction power than other models.

Table 6.8 Skill scores with anticipated gap model after data under-sampling.

Skill Scores	BLM	SVM	EBM	Remark
Accuracy	0.758	0.795	0.815	The higher, the better. (Improper index for imbalanced dataset)
Precision	0.618	0.657	0.676	The higher, the better.
Bias	0.781	0.989	1.074	Bias>1 : over-estimate Bias<1 : under-estimate Bias=1 : perfect
Sensitivity (or Recall, POD)	0.482	0.649	0.726	1 (best score) 0 (worst score)
FAR (False Alarm Ratio)	0.382	0.343	0.324	0 (best score) 1 (worst score)
F-Measure	0.542	0.653	0.700	1 (best score) 0 (worst score)
HSS (Heidke Skill Score)	0.380	0.507	0.567	if HSS=1 : perfect prediction

The performance of the prediction models appear to be comparable in ROC/PR spaces. Figure 6.1 illustrates ROC/PR curves according to different prediction models with adjacent gap model before data under-sampling. From Fig. 6.1(a), the

area under receiver operating characteristic curve(AUROC) by EBM model is 0.931, on the other hand, the AUROC by SVM and BLM models are 0.906 and 0.778, respectively. The AUPR values are 0.659, 0.673 and 0.307 according to EBM, SVM, and BLM. The effect of data under-sampling on ROC/PR curves are illustrated in Fig. 6.1(b). To validate the prediction power of classifiers considering the anticipated gap model as well as data under-sampling technique, similar ROC/PR curves have been plotted in Fig. 6.2. The area under receiver operating characteristic curve(AUROC) by EBM model is 0.972, on the other hand, the AUROC by SVM and BLM models are 0.987 and 0.845, respectively. It is concluded that SVM and EBM models belong to “excellent” prediction and BLM shows “good” prediction from the properties of ROC in Table 4.2. In cases of SVM and EBM, the corresponding AUPR values are 0.970 and 0.801, respectively. However, the AUPR by BLM is 0.658 that means worse performance than SVM and EBM.

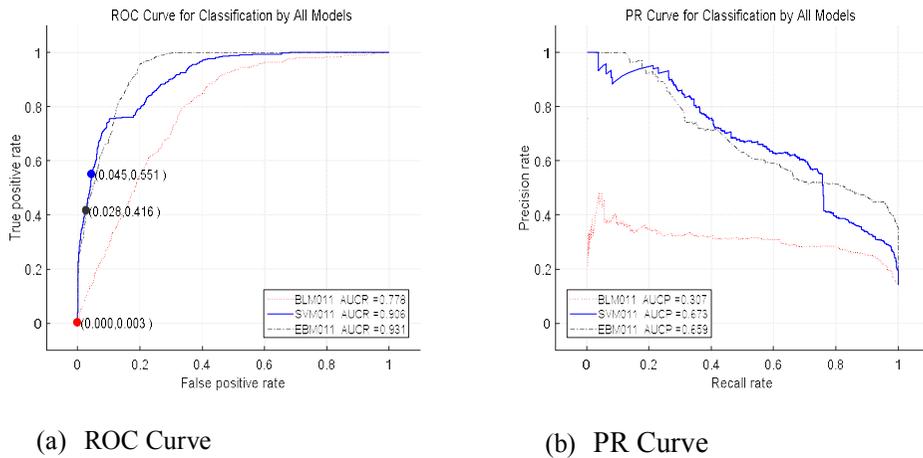
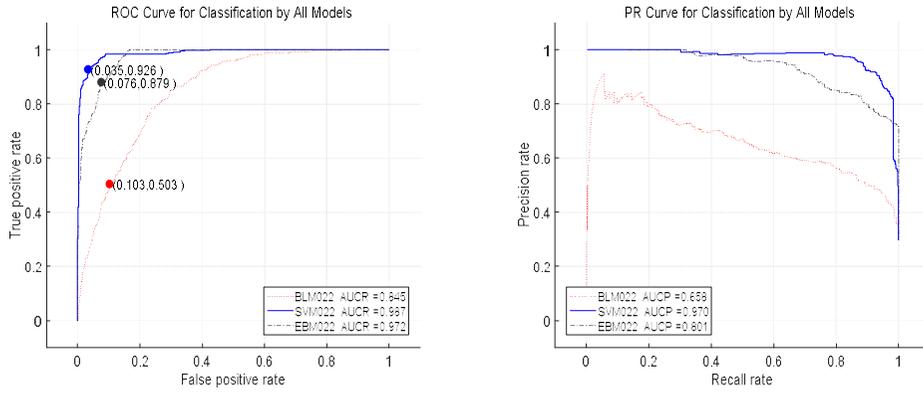


Figure 6.1 ROC and PR curves with adjacent gap model before data under-sampling.



(a) ROC Curve

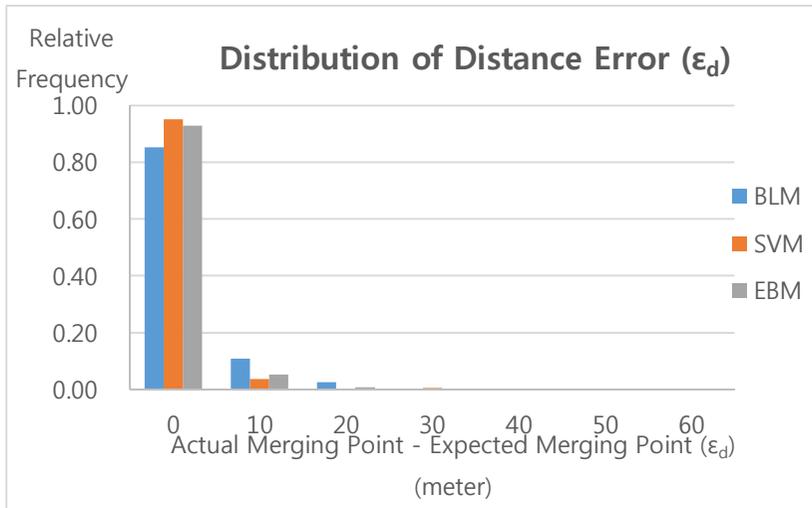
(b) PR Curve

Figure 6.2 ROC and PR curves with anticipated gap model after data under-sampling.

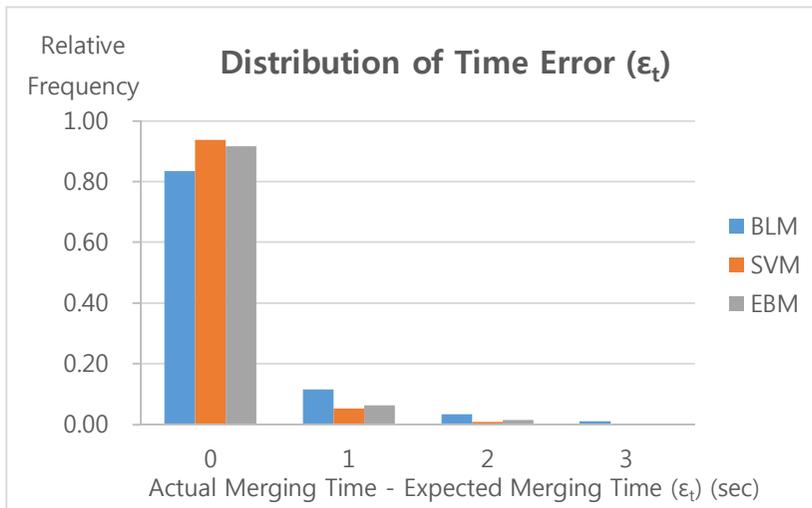
6.2 Performance Evaluation of Classifiers by Sampling Interval

Due to the fact that NGSIM datasets have a problem of imbalance between numbers of “rejected” cases and “accepted” cases, which causes the poor prediction power of the lane-change prediction model; therefore, “sampling time interval” method is applied. “Sampling time interval” method is conducted by following steps: first is to find the point for when merging vehicle completes its lane-change; and second is to reduce the number of “rejected” cases(0) by applying different sampling time intervals by starting from the point of merging (i.e. when it displays “1”).

With resampled dataset based on “sampling time interval”, distribution of distance error and time error are plotted as shown in Figure 6.3. Distance Error (in meter) indicates the difference between actual merging point and expected merging point by prediction models. On the other hand, Time Error (in second) means the difference between actual merging time and expected merging time by prediction models.



(a) Distribution of distance error (ϵ_d).



(b) Distribution of Time Error (ϵ_t).

Figure 6.3 Distribution of distance and time errors after data reduction using 1 sec time interval.

Table 6.9 Percent of correctly predicted data by prediction models.

Prediction Model	Percent(%) of correctly predicted data	
	Distribution of Distance	Distribution of Time
BLM	85	84
SVM	95	94
EBM	93	92

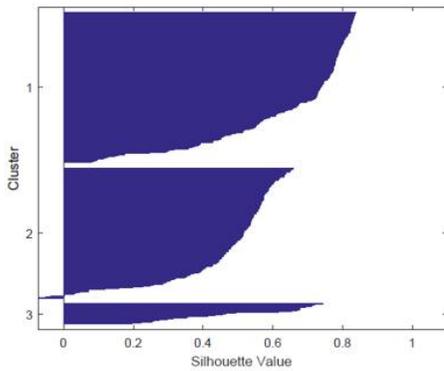
The performance of the prediction models is compared in terms of distribution of distance and time error. Figure 6.3(a) illustrates the distribution of distance as well as Figure 6.3(b) shows the distribution of time. It is observed from Figure 6.3 that more number of predicted data derived from SVM and EBM is located more concentrated at error (ϵ_d or ϵ_t) =0. Table 6.9 summarizes the result from distribution of distance and time error (i.e. percent of data correctly predicted by prediction models (BLM, SVM, EBM)). From Table 6.9, the percent of correctly predicted data based on distribution of distance by SVM model is 95%, on the other hand, the percent of correctly predicted data by EBM and BLM models are 93% and 85%, respectively. The percent of correctly predicted data in terms of distribution of time values are 94%, 92% and 84% according to SVM, EBM, and BLM. This means that SVM and EBM show better prediction performance as compared to BLM.

6.3 Decision-Making Process by Vehicle Trajectory

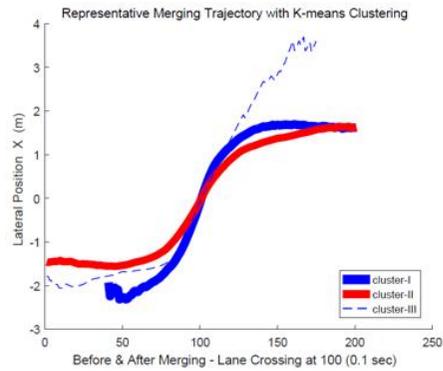
It is very important to draw the plot of vehicle trajectories to identify the decision making process of merging vehicles. For this purpose, the silhouette diagram by K-means clustering has been illustrated in Figure 6.4(a) showing 3 clusters according to different vehicle trajectory patterns. The corresponding vehicle trajectories are shown in Figure 6.4(b). For simplicity, the segments of lane changing trajectories are only considered with uniform length (20 sec) in this study since a normal lane

changing is finished in 20 sec according to Salvucci and Liu(2002) and Toledo and Zohar(2007). The variation of lateral position of vehicles has been plotted with respect to the truncated time interval of vehicle trajectories ranging from -10 sec to +10 sec in reference to the merging point of lane boundary. The similar vehicle trajectories are shifted to a single merging point to estimate the overall merging patterns. As shown in Figure 6.4(b), the Cluster III represents that the merging vehicles move aggressively from Lane 6 to Lane 4 that is excluded in this study. The number of vehicles belonging to Cluster III is only 22. The difference of Cluster I(186 vehicles) and Cluster II(157 vehicles) is the required access times for passing the merging point of lane boundary. Cluster II needs more access times to be merged into the target lane as compared with Cluster I. Thus Cluster-I and Cluster-II classified by vehicle trajectory patterns represent “direct merging” and “chase merging”, respectively.

To confirm the merging patterns using cluster types, Figure 5.2 illustrates the merging patterns and gap selection scenarios. Merging vehicles taking their original gap as their targeting gap are called “direct merging” (Wan et al, 2016; Chu, 2014) as shown in Figure 5.2(b). On the other hand, if a merging vehicle follows the original-gap-target, this is called “direct merging” as shown in Figure 5.2(c). According to these classification of merging patterns, it is noted that “Cluster-I” and “Cluster-II” may belong to “direct Merging” and “chase Merging”, respectively, as mentioned before.



(a) Silhouette diagram showing 3 clusters.



(b) Representative trajectory patterns.

Figure 6.4 Classification of trajectory patterns by K-means clustering.

The discretized lateral velocity from the MLC vehicle trajectories every 0.5 sec based on lateral position has been illustrated in Figure 6.5 according to direct merging pattern as well as chase merging pattern. In other words, the MLC trajectories in terms of lateral position are fitted by a fifth-order polynomial showing “S”-like curves and thus the discretized lateral velocity can be obtained numerically by differentiating as defined in Eq. (5.3).

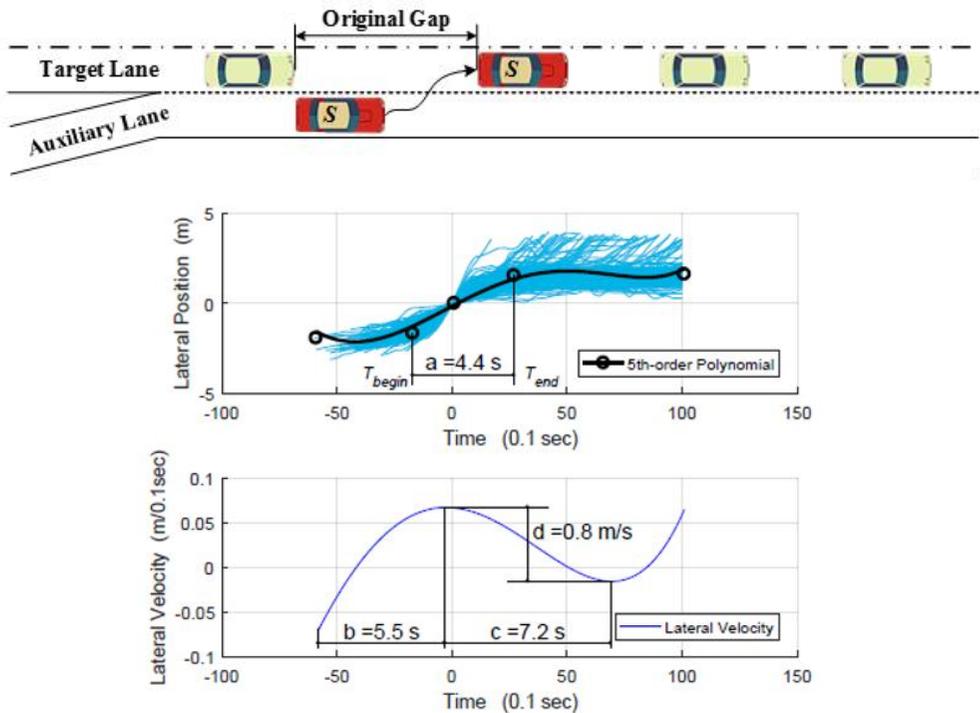


Figure 6.5 Lateral position and discretized lateral velocity for Cluster I (direct merging).

As shown in Figure 6.5, the vehicle trajectory and corresponding discretized lateral velocity for direct merging are illustrated with respect to time t . Due to the characteristics of direct merging(Cluster-I), merging of vehicles completes within approximately 5.5 sec. The duration time (a) is estimated by 4.4 sec that is counted by crossing time of lane boundary. Also, the times for merging acceleration (b) and merging relaxation (c) after passing the lane boundary are detected by 5.5 sec and 7.2 sec, respectively. Applying merging acceleration can be proceeded with the increase of vehicle speed to reach the desired merging position. The merging relaxation is detected when the merging vehicle feels comfortable to apply the continuous lane-change maneuver after moving into the target lane. The amplitude of lateral velocity (d) is read by 0.8 m/sec.

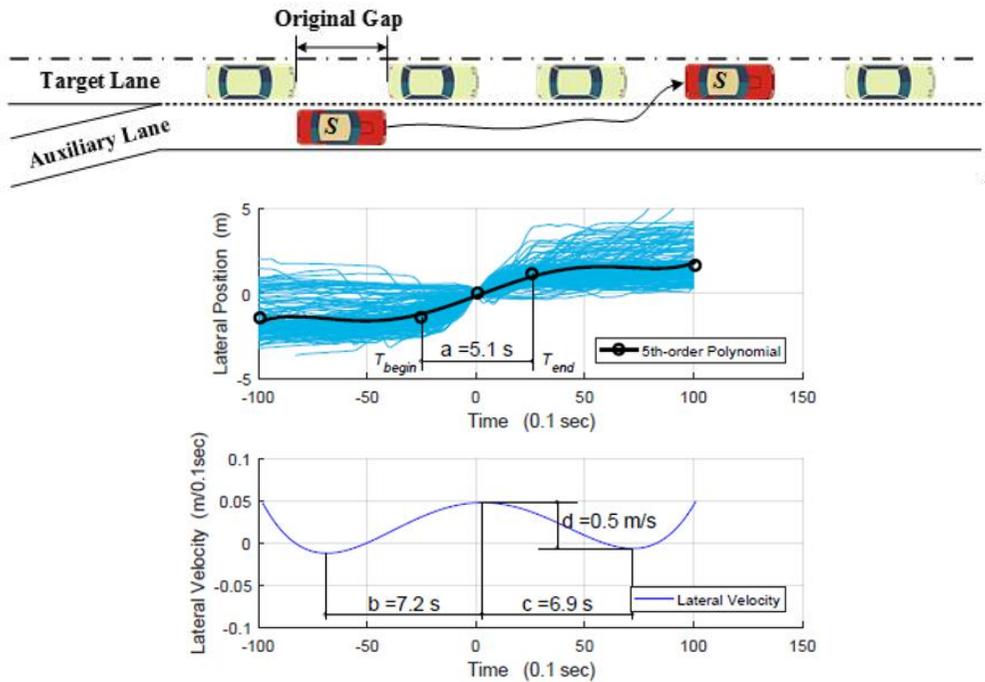


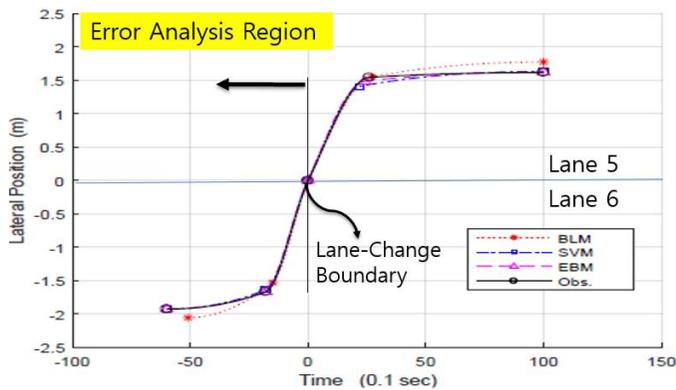
Figure 6.6 Lateral position and discretized lateral velocity for Cluster II (chase merging).

In case of chase merging(Cluster-II) shown in Figure 6.6, however, duration time (a) is estimated by 5.1 sec and the times for merging acceleration (b) and merging relation to be stable position are detected by 7.2 sec and 6.9 sec, respectively. The amplitude of lateral velocity (d) is read by 0.5 m/sec. In general, the duration time of lane-change can be influenced by the merging patterns or risk-taking characteristics of drivers with respect to length of acceleration lane, traffic condition of target lane and existence of off-ramp, *etc.* From the analysis of vehicle trajectories, it is noted that the average lane-change duration and merging acceleration times for direct merging pattern decrease as compared with chase merging pattern. This is mainly due to the characteristics of chase merging may last a much longer time for lane-change. Similar finding is detected in case of lateral velocity amplitude. The amplitude of lateral velocity of direct merging read by 0.8

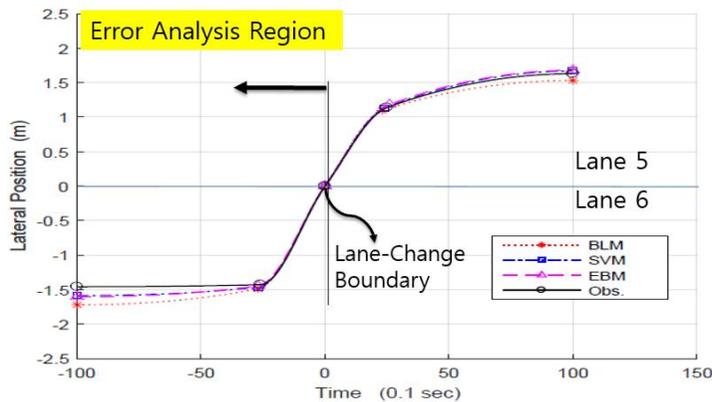
m/sec is greater than that of chase merging read by 0.5 m/sec that is caused by the accelerated action to move into a mainline within the shorter time. To show the errors in reference to the observed model, two performance indices are used such as the mean absolute error(MAE) and the root of mean squared error(RMSE), respectively.

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i^{obs} - y_i^{pred}| \quad (6.1)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i^{obs} - y_i^{pred})^2} \quad (6.2)$$



(a) Vehicle trajectory for direct merging pattern.



(b) Vehicle trajectory for chase merging pattern.

Figure 6.7 Comparison of vehicle trajectories between direct and chase merging

patterns according to prediction models.

The estimated errors are based on the half of whole vehicle trajectory ($-10 \text{ sec} < t < 0 \text{ sec}$) as compared with the observed merging vehicles classified in Table 6.10 considering Cluster-I(direct merging) and Cluster-II(chase merging). In cases of prediction models after data under-sampling, the vehicles classified as True-Positive(TP) cases in Figure 6.7 are selected for comparison purpose. It is noted that BLM holds the highest error, especially in direct merging pattern as compared with SVM and EBM. However, there is no significant discrepancy of errors among three prediction models in case of chase merging pattern.

Table 6.10 Errors of prediction models with respect to the observed model.
 (-10 sec<t<0 sec: error analysis region before crossing lane boundary)

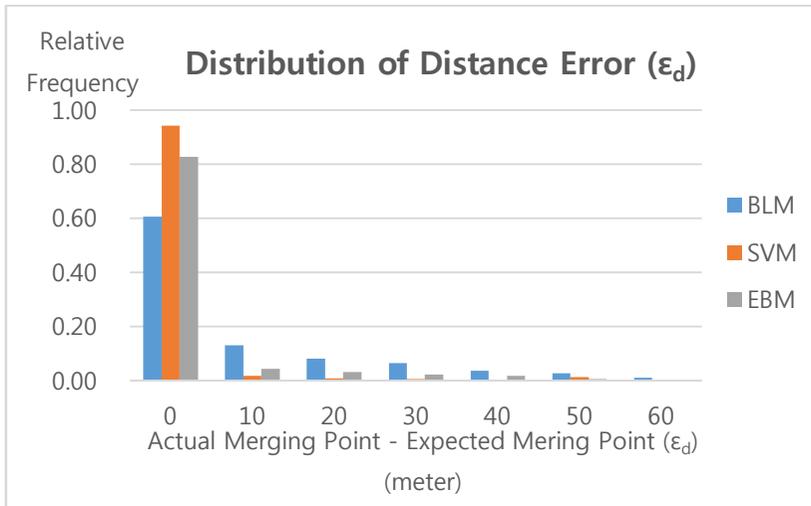
Cluster Type	Prediction Model	Error Index	
		RMSE	MAE
Cluster-I (Direct Merging)	BLM	2.450	10.503
	SVM	0.127	0.016
	EBM	0.033	0.001
Cluster-II (Chase Merging)	BLM	0.795	0.764
	SVM	0.751	0.629
	EBM	0.757	0.646

Table 6.11 Number of True-Positive(TP) vehicles by prediction models.

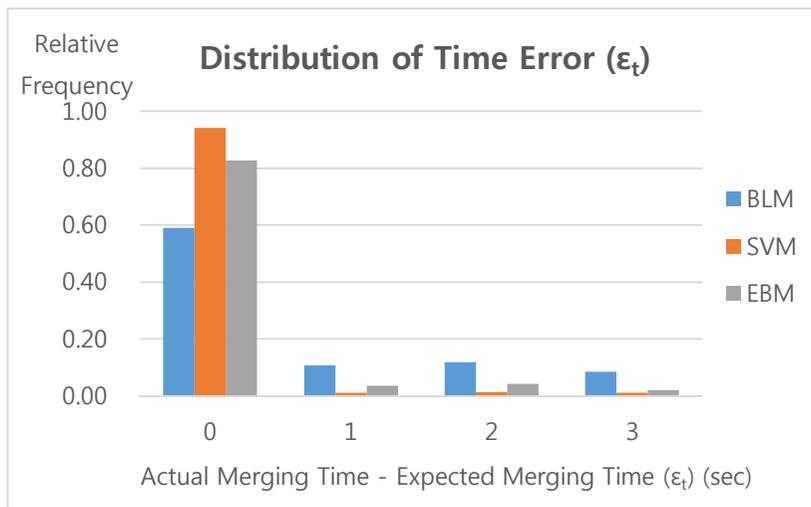
Cluster Type	Observed Merging Vehicles	Number of True-Positive Vehicles by Prediction Models		
		BLM	SVM	EBM
Cluster – I (Direct Merging)	184	56	112	132
Cluster – II (Chase Merging)	159	106	115	122
Cluster III	22	14	10	11
Total	365	176	237	265

Table 6.10 gives the errors by MAE and RMSE have been calculated focused on the merging region (-10 sec<t<0 sec) since prediction models are related to merging decision according to whether the center of front bumper of merging vehicle crosses the lane boundary. Thus the estimated errors are based on the half of whole vehicle trajectory as compared with the observed merging vehicles classified in Table 6.11 considering Cluster-I(direct merging) and Cluster-II(chase merging). In cases of prediction models after data under-sampling, the vehicles classified as True-Positive(TP) cases, as shown in Tables 6.4-6.6 or Table 6.11, are selected for comparison purpose. The detailed number of classified vehicles are presented in

Table 6.10. It is noted that BLM holds the highest error, especially in direct merging pattern(MAE=10.543, RMSE=2.450) as compared with SVM(MAE=0.016, RMSE=0.127) and EBM(MAE=0.001, EBM=0.033). However, there is no significant discrepancy of errors among three prediction models in case of chase merging pattern.

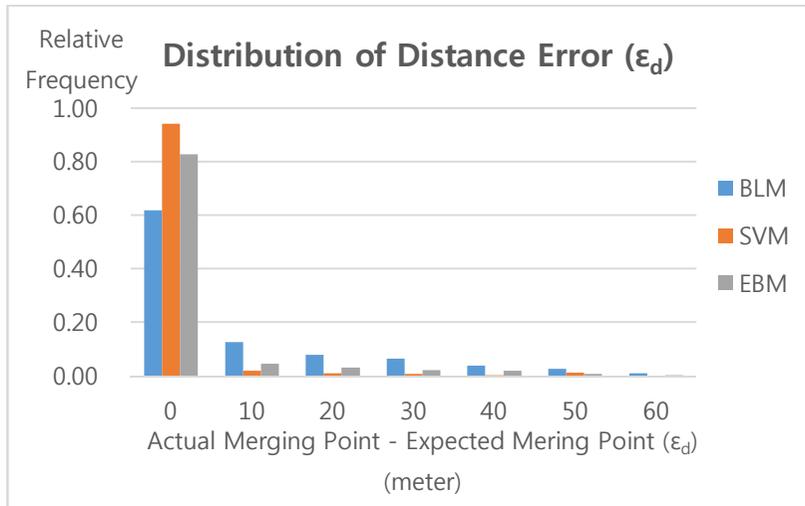


(a) Distribution of distance error for Cluster I (ϵ_d).

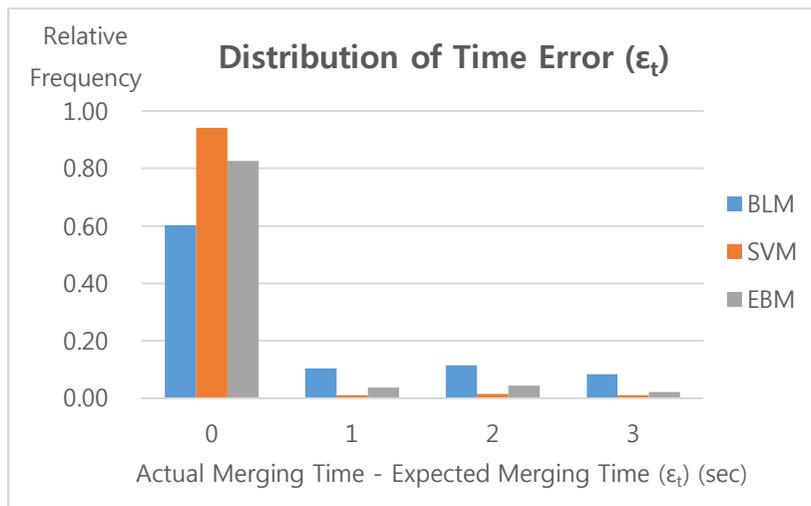


(b) Distribution of time error for Cluster I (ϵ_t).

Figure 6.8 Distribution of distance and time errors for Cluster I

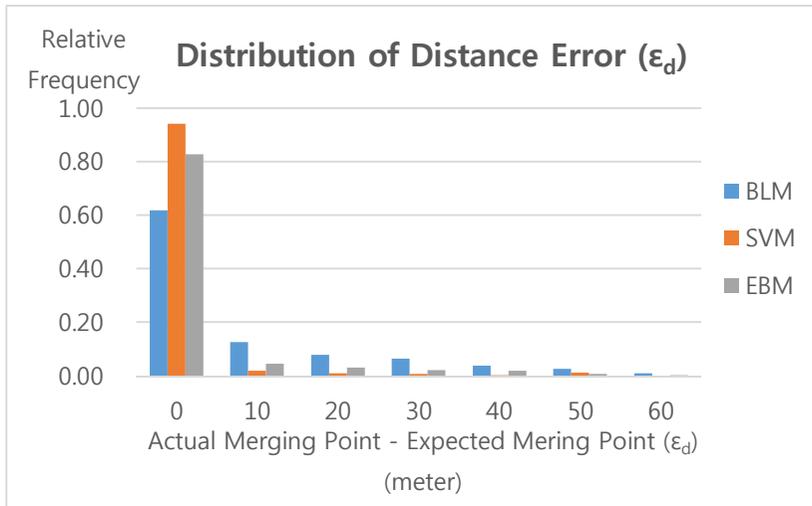


(a) Distribution of distance error for Cluster II (ϵ_d).

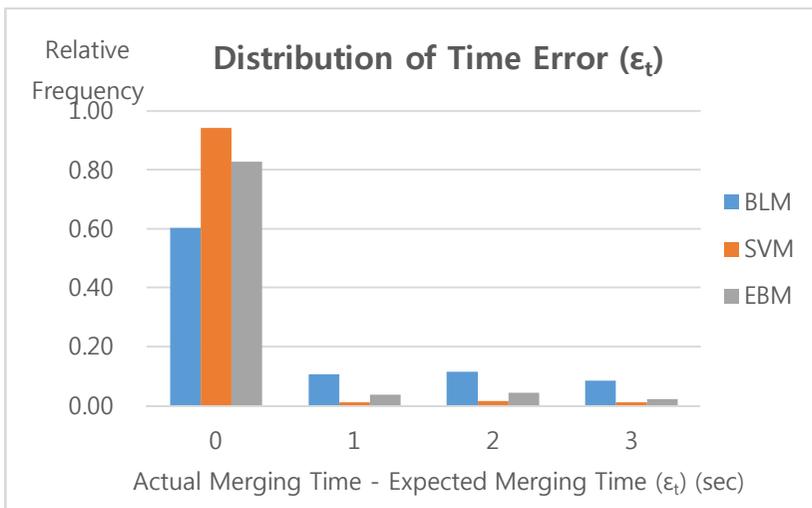


(b) Distribution of time error for Cluster II (ϵ_t).

Figure 6.9 Distribution of distance and time errors for Cluster II



(a) Distribution of distance error for Cluster III (ϵ_d).



(b) Distribution of time error for Cluster III (ϵ_t).

Figure 6.10 Distribution of distance and time errors for Cluster III

Table 6.12 Percent of correctly predicted data by prediction models for Cluster I.

Prediction Model	Percent(%) of correctly predicted data	
	Distribution of Distance	Distribution of Time
BLM	61	59
SVM	94	94
EBM	83	83

Table 6.13 Percent of correctly predicted data by prediction models for Cluster II.

Prediction Model	Percent(%) of correctly predicted data	
	Distribution of Distance	Distribution of Time
BLM	62	60
SVM	94	94
EBM	83	83

Table 6.14 Percent of correctly predicted data by prediction models for Cluster III.

Prediction Model	Percent(%) of correctly predicted data	
	Distribution of Distance	Distribution of Time
BLM	62	60
SVM	94	94
EBM	83	83

The percent of correctly predicted data can be computed separately depending on types of clusters. Figure 6.8-6.10 illustrates the distribution of distance as well as shows the distribution of time. Based on Figure 6.8-6.10, it is observed that more number of predicted data derived from SVM and EBM is located more concentrated at error (ϵ_d or ϵ_t) = 0. Table 6.12-6.14 summarize the result from distribution of distance and time error (i.e. percent of data correctly predicted by prediction models

(BLM, SVM, EBM)). From Table 6.12, the percent of correctly predicted data based on distribution of distance for Cluster I by SVM model is 94%, on the other hand, the percent of correctly predicted data by EBM and BLM models are 83% and 61%, respectively. The percent of correctly predicted data in terms of distribution of time values are 94%, 83% and 59% according to SVM, EBM, and BLM. This means that SVM and EBM show better prediction performance as compared to BLM. From Table 6.13, the percent of correctly predicted data based on distribution of distance for Cluster II by SVM model is 94%, on the other hand, the percent of correctly predicted data by EBM and BLM models are 83% and 62%, respectively. The percent of correctly predicted data in terms of distribution of time values are 94%, 83% and 60% according to SVM, EBM, and BLM. This means that SVM and EBM show better prediction performance as compared to BLM. From Table 6.14, the percent of correctly predicted data based on distribution of distance for Cluster III by SVM model is 94%, on the other hand, the percent of correctly predicted data by EBM and BLM models are 83% and 62%, respectively. The percent of correctly predicted data in terms of distribution of time values are 94%, 83% and 60% according to SVM, EBM, and BLM. This also shows that SVM and EBM have better prediction performance as compared to BLM.

6.4 Classification of Merging Patterns

Choudhury *et al.* (2007) are one of the few authors explicitly classify the driver's behaviors into timid and aggressive where the anticipated gap is split into a lead gap and a lag gap. As shown in Figures 6.11-6.14, the boundary curves proposed by Choudhury *et al.* (2007) have been marked by a solid line and a dotted line to define timid and aggressive drivers. As for direct and chase merging patterns, the observed model using 365 accepted vehicles (direct: 184, chase: 159) shows the relation between anticipated gap and distance to end of acceleration lane. It is found that two different types of merging pattern are clearly distributed into two groups by circular

and triangular tick marks. In case of direct merging, merging vehicles complete the lane change near the start line of acceleration lane as it is expected. It is also known that the classified direct merging may belong to aggressive drivers. However, the chasing merging pattern is widely spread out from center line to end line of acceleration lane. In other words, it is not clearly grouped in contrast to direct merging due to the weaving action since the off-ramp is installed near the end of acceleration lane. Nevertheless, the chase merging pattern is close to timid driving behavior. However, relation between anticipated gap and distance to end of acceleration lane by prediction models using True-Positive vehicles classified in contingency matrix is shown in Figures 6.11-6.14. In case of BLM, number of True-Positive vehicles is 176(direct: 56, chase: 106). On the other hand, 237 accepted True-Positive vehicles (direct: 112, chase: 115) by SVM, and 265 accepted True-Positive vehicles (direct: 132, chase: 122) by EBM, respectively as shown in Table 6.11. It is noted that the distribution of merging patterns by SVM and EBM using only accepted True-Positive vehicles shows an excellent agreement with the distribution of merging patterns by the observed model. However, there are some discrepancies in the distribution of merging patterns by BLM as compared with the observed model, mainly due to the insufficiency of classified True-Positive vehicles.

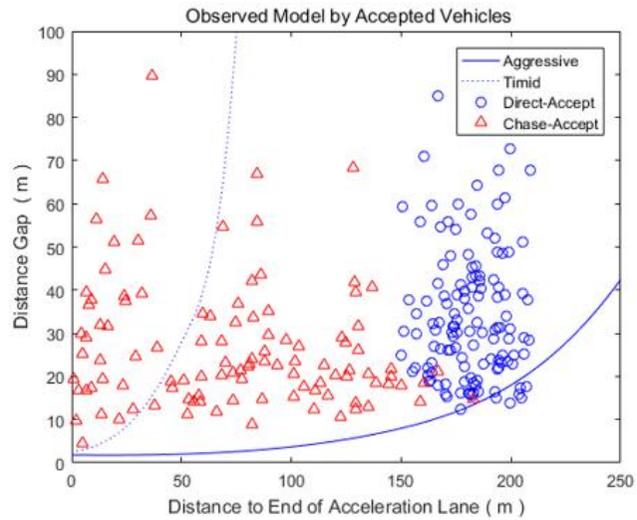


Figure 6.11 Relation between gap and distance to end of acceleration lane by observed model using accepted vehicles in contingency matrix.

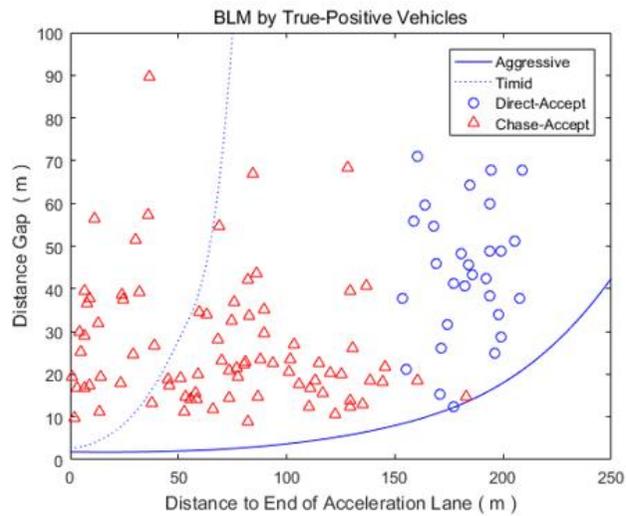


Figure 6.12 Relation between gap and distance to end of acceleration lane by BLM using True-Positive vehicles in contingency matrix.

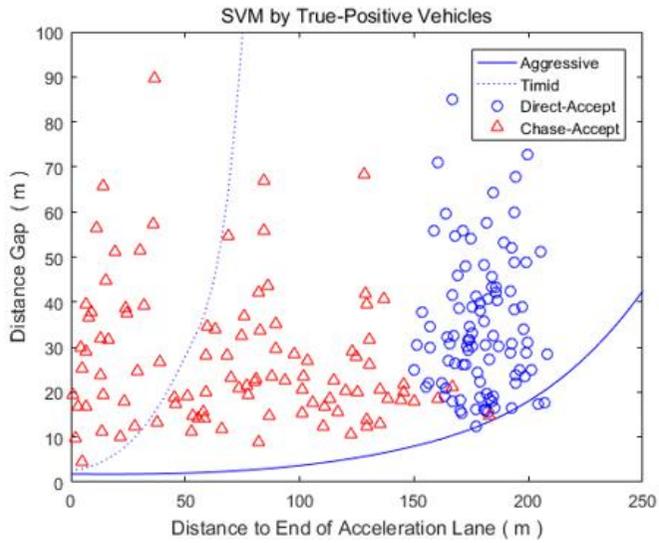


Figure 6.13 Relation between gap and distance to end of acceleration lane by SVM using True-Positive vehicles in contingency matrix.

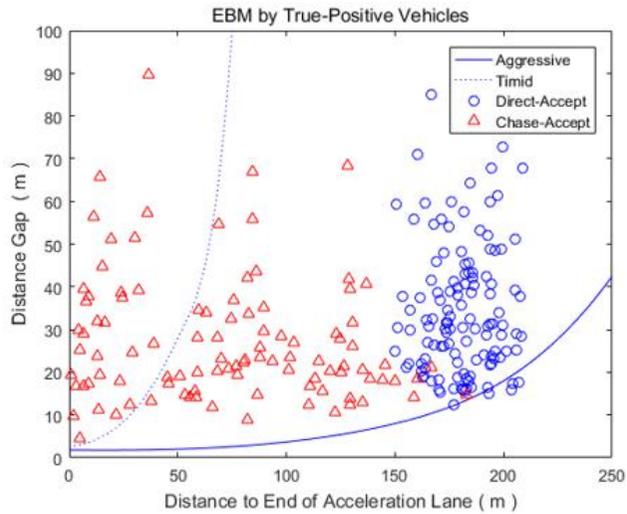


Figure 6.14 Relation between gap and distance to end of acceleration lane by EBM using True-Positive vehicles in contingency matrix.

Chapter 7 Conclusions

The evaluation study applying SVM(support vector machine) and EBM(ensemble boosting method) algorithms to lane-change detection has been carried out at the US-101 site. For this study, the data processing by Hampel filter, data under-sampling technique and K-means clustering have been considered to improve the data quality as well as to create the balanced dataset. Firstly, impulsive noise and outlier noise of data were removed by Hampel filtering. Secondly, duplicate elimination by averaging was adopted to combine duplicate data with a single averaging value as well as sampling time interval technique was used to reduce the number of “rejected” cases by a certain time interval. Thirdly, the merging patterns of drivers were classified into direct, chase and yield merging by K-means clustering. At the same time, the “anticipated gap acceptance” theory proposed by Choudhury(2007) was used by considering the interaction of lead and lag vehicles with respect to a subject vehicle. From the numerical analyses, the following results can be summarized:

- (1) The proposed prediction models by SVM and EBM improve not only the overall classification power, but also significantly reduce the prediction errors for lane change detection at the freeway on-ramp.
- (2) For performance evaluation, the contingency matrix, ROC/PR curves and skill scores such as bias, precision, recall(POD), F-measure, FAR, and HSS. It is concluded that the machine algorithms outperform the conventional BLM predictor for imbalanced datasets and are insensitive to the relative numbers of training data in accepted and rejected cases. Thus these results evidence the applicability of machine algorithm in traffic data analysis.

- (3) It is noted that the data under-sampling techniques have a great effect on creating the balanced dataset as well as improving the data quality.
- (4) The merging patterns under MLC condition of US-101 site can be classified into direct, chase and yield merging in spite of weaving phenomenon. However, the number of vehicles corresponding to yield merging is only five. For this reason, the yield merging pattern has been excluded in this study.
- (5) The duration time and amplitude of lateral velocity for direct merging and chase merging are investigated by 4.4 sec and 5.1 sec as well as 0.8 m/sec and 0.5 m/sec, respectively.
- (6) The BLM predictor holds the highest error, especially in direct merging pattern(MAE=10.543, RMSE=2.450) as compared with SVM(MAE=0.016, RMSE=0.127) and EBM(MAE=0.001, RMSE=0.033). However, there is no significant discrepancy of errors among three prediction models in case of chase merging pattern.
- (7) As for direct and chase merging patterns, the observed model using 356 accepted vehicles (direct: 184, chase: 159) in contingency matrix shows the relation between anticipated gap and distance to end of acceleration lane. However, prediction models using True-Positive vehicles classified in contingency matrix. In case of BLM, number of True-Positive vehicles is 176(direct: 56, chase: 106). On the other hand, 237 accepted True-Positive vehicles (direct: 112, chase: 115) by SVM, and 265 accepted True-Positive vehicles (direct: 132, chase: 122) by EBM, respectively.
- (8) It is noted that the distribution of merging patterns by SVM and EBM using only accepted True-Positive vehicles shows an excellent agreement with the distribution of merging patterns by the observed model. But, there are

some discrepancies in the distribution of merging patterns by BLM as compared with the observed model that are mainly due to the insufficiency of classified True-Positive vehicles.

REFERENCES

- Ahmed, K.I.(1999). Modeling drivers' acceleration and lane changing behavior. *Ph.D Dissertation*, Civil Engineering Department, MIT, U.S.A..
- Ben-Akiva, M. E., and Lerman, S. R. (1985). *Discrete Choice Analysis: Theory and Application to Travel Demand*, Vol. 9, MIT Press, Cambridge, Mass.
- Bockhorst, J. and Craven, M. (2005). Markov networks for detecting overlapping elements in sequence data. *Neural Information Proceeding Systems 17(NIPS-17)*. pp.193-200, MIT Press.
- Burges, C. J. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, Vol. 2, No. 2, pp. 121–167.
- Cano, J.R., Herrera, F., and Lozano, M. (2006). On the Combination of Evolutionary Algorithms and Stratified Strategies for Training Set Selection in Data Mining. *Applied Soft Computing*, Vol. 6, pp. 323-332.
- Choudhury, C.F., Ben-Akiva, M., Toledo, T., Lee, G. and Rao, A. (2007). Modeling cooperative lane changing and forced merging behavior. *The 86th Annual Meeting of the Transportation Research Board, Washington*.
- Chu, T.D. (2014). A Study on Merging Behavior at Urban Expressway Merging Sections. *Ph.D. Dissertation*, Civil Engineering Department of Nagoya University, Japan.
- Cortes, C., and Vapnik, V. N. (1995). Support Vector Networks. *Machine Learning*, Vol. 20, pp. 273–297.
- Davis J. and Goadrich M. (2006). The relationship between precision-recall and ROC curves. *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, PA.:ACM pp.233-240.
- DOT (US DOT) and FHWA(Federal Highway Administration) (2006). Next Generation SIMulation(NGSIM) traffic analysis tools. <http://>

ops.fhwa.dot.gov/trafficanalysistools/ngsim.htm.

- Deepak, R.B. and Suresh J. (2006). Improving duplicate elimination in storage systems. *ACM Transactions on Storage*, Vol.V, No. N, pp.1-23.
- Dimitriadou, E., Weingessel, A., Hornik, K. (2003). *A cluster ensembles framework, Design and application of hybrid intelligent systems*, IOS Press, Amsterdam, The Netherlands.
- Domingos, P. (1999). Metacost: A general method for making classifiers cost-sensitive. In *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 155–164)., San Diego. ACM Press.
- Fatourechi, M., Ward, R. K., Mason, S. G., Huggins, J., Schlögl, A., and Birch, G. E. (2008). Comparison of Evaluation Metrics in Classification Applications with Imbalanced Datasets. *Seventh International Conference on Machine Learning and Applications*, pp.777-782.
- Hartigan, J.A. and Wong, M.A. (1979). A K-means clustering algorithm. *Applied Statistics*, Vol.28, No.1, pp.100-108.
- Herman, R. and Weiss, G.H. (1961). Comments on highway-crossing problem, *Operation Research*, Vol.9, pp.828-840.
- Hidas, P. (2002). Modeling lane changing and merging in microscopic traffic simulation. *Transportation Research, Part C, Emerging Technologies*, Vol.10(No.5-6), pp.352-371.
- Hidas, P. (2005). Modeling vehicle interactions in microscopic simulation of merging and weaving. *Transportation Research Part C: Emerging Technologies*, Vol.10, pp.351-371.
- Hou, Y., Edara, P. and Sun, C. (2014). Modeling mandatory lane changing using Bayes classifier and decision trees. *IEEE Transactions of Intelligent Transportation Systems*, Vol.15, No.2, pp.647-655.
- Keilwagen, J., Grosse, I., and Grau, J. (2014). Area under precision-recall curves for

- weighted and unweighted data. *PLOS ONE*, 9(3): <http://dx.doi.org/10.1371/journal.pone.0092209>.
- Kita, H. (1993). Effects of merging lane-length on the merging behavior at expressway on-ramps. *Proceedings of the 12th International Symposium on the Theory of Traffic Flow and Transportation*, pp.37-51.
- Kohavi, R., Provost, F. (1998). Glossary of terms. *Machine Learning*, Kluwer Academic Publishers, Vol.30, pp. 271-274.
- Kondyli, A. and Elefteriadou, L. (2011). Modeling driver behavior at freeway-ramp merges. *Transportation Research Record*, Vol.2249, pp.29-37.
- Kubat, M., Holte, R. and Matwin, S. (1997). Addressing the curse of imbalanced datasets: One-sided sampling. *Proceedings of the 14th International Conference on Machine Learning*, Morgan, Kaufmann(pp.179-186).
- Kumar, Y. and Sahoo, G. (2012). Analysis of parametric and non-parametric classifiers for classification technique using WEKA. *Information Technology and Computer Science*, Vol.7, pp.43-49.
- Mandalia, H.M., & Salvucci, D.D. (2005). Using support vector machines for lane change detection. *Proceedings of the Human Factors and Ergonomics Society 49th Annual Meeting* (pp. 1965-1969). Santa Monica, CA: Human Factors and Ergonomics Society.
- Marczak, F., Daamen, W. and Buisson, C. (2013). Merging behavior: Empirical comparison between two sites and new theory development. *Transportation Research Part C*, Vol.36, pp.530-546.
- Miller, A.J. (1972). Nine estimations of gap acceptance parameters. *Proceeding of the 5th International Symposium on the Theory of Traffic Flow and Transportation*, pp.215-235.
- NGSIM U.S. 101 Data Analysis (2005). Summary Report, Federal Highway Administration, Cambridge Systematics, Inc.
- Pawar, D.S., Patil, G.R., Chandrasekharan, A., and Upadhyaya, S. (2015).

- Classification of gaps at uncontrolled intersections and midblock crossings using support vector machines. *Transportation Research Record*, Vol.2515, pp.26-33.
- Pazzani, M., Merz, C., Murphy, P., Ali, K., Hume, T., & Brunk, C. (1994). Reducing misclassification costs. *Proceedings of the 11th International Conference on Machine Learning*, San Francisco. Morgan Kaufmann.
- Pearson, R.K., Neuvo, Y., Astola, J. and Gabbouj, M. (2014). Generalized Hampel Filters. *EURASIP Journal on Advances in Signal Processing*. 2016:87, DOI 10.1186/s13634-016-0383-6.
- Provost, F., Fawcett, T. and Kohavi, R. (1998). The case against accuracy estimation for comparing induction algorithms. *Proceedings of the 15th International Conference on Machine Learning* (pp.445-453). Morgan Kaufmann, San Francisco, CA.
- Punzo, V., Borzacchiello, M. T., and Ciuffo, B. (2011). “On the assessment of vehicle trajectory data accuracy and application to the Next Generation SIMulation (NGSIM) program data.” *Transportation Research Part C*, Vol.19, No.6, pp.1243–1262.
- Rokach, L., Schclar, A. and Itach, E. (2014). Ensemble methods for multi-label classification. *Expert Systems with Applications*, Vol. 41, pp. 7507–7523.
- Salvucci, D.D. and Liu, A. (2002). The time course of a lane change: Driver control and eye-movement behavior. *Transportation Research, Part F, Traffic Psychology and Behaviour*, Vol.5, No.2, pp.123-132.
- Scholkopf, B., and Smola, A.J. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press, Cambridge, Mass., 2002.
- Sun, A., Lim, E.-P., Liu, Y. (2009). On strategies for imbalanced text classification using SVM: A comparative study. *Decision Support Systems*, Vol.48 (1), pp. 191-201.

- Sun, D. and Elefteriadou, L. (2010). Research and implementation of lane-changing model based on driver behavior. *Transportation Research Record*, Vol.2161, pp.1-10.
- Thiemann, C., Treiber, M., and Kesting, A. (2008). “Estimating acceleration and lane-changing dynamics based on NGSIM trajectory data.” *Transportation Research Record*, Vol.2088, pp.90–101.
- Toledo, T. and Zohar, D. (2007). Modeling duration of lane changes. *Transportation Research Record*, Vol.1999, pp.71-78.
- Vapnik, V. N. (1998). *Statistical Learning Theory*, Vol. 2. John Wiley & Sons, New York.
- Wan, X., Jin, P.J., Yang, F. and Ran, B. (2016). Merging preparation behavior of drivers: How they choose and approach their merge positions at a congested weaving area. *Journal of Transportation Engineering, ASCE* DOI: 10.1061/(ASCE)TE.1943-5436.0000864, pp.05016005-1~10.
- Wan, X., Jin, P.J., Gu, H., Chen, X. and Ran, B. (2017). Modeling freeway merging in a weaving section as a sequential decision-making process. *Journal of Transportation Engineering, Part A: Systems, ASCE*, Vol.143, No.5, pp.05017002-1 ~13.
- Wang, Q., Li, Z. and Li, L. (2014). Investigation of discretionary lane-change characteristics using NGSIM data sets., *Journal of Intelligent Transportation Systems*, 18:246-253.

국문 초록

고속도로 합류부에서는 본선부차선과 합류부차선의 교통류가 혼입됨에 따른 상호작용뿐만 아니라 다양한 교통영향인자(즉, 교통량의 상태, 도로기하구조, weaving, 운전자 행태에 따라 달라지는 개인적 반응 등)로 인하여 차로변경여부를 예측하는 것은 매우 어렵다. 또한, NGSIM US-101 데이터는 동일한 교통상황에서 한 운전자가 여러 번에 걸쳐 비합류 결과(non-merge event)를 발생시키는 반면, 1개의 합류결과(merge event)만을 생성하는 데이터 구조를 갖기 때문에 본질적으로 불균형 데이터가 만들어 진다. 즉, 다수의 “rejected cases”에 비해 소수의 “accepted cases”로 구성된다. 이와 같이 불균형 데이터 구조를 갖고 차로변경여부의 판정은 다수의 경우에 편중(biased)하게 되기 때문에 종종 “accepted cases”를 “rejected cases”로 오분류될 수 있다. 강제차로변경(MLC) 환경 하에서 합류차량의 차로변경의 판정을 위해 제안된 분류기들의 성능을 향상시키고, 불균형 데이터를 완화시키기 위해서 본 연구의 전략은 3가지로 요약될 수 있다.

첫째는, 자료 불균형문제 해결을 위해 데이터 샘플링기법을 도입하여 분할표(contingency matrix)와 이를 활용한 다양한 평가지표(skill scores) 및 ROC/PR 곡선을 통해 분류성능을 보이고자 한다. 이 목적을 위해 먼저 MATLAB 프로그램에 내장되어 있는 Hampel 필터링 기법을 사용하여 비정상적 이상치의 제거와 측정오차를 저감시켰다. 또한 “rejected cases”의 개수를 줄이기 위해 데이터저장을 위해 EXCEL Spread Sheet에 많이 사용되는 평균화에 의한 복제데이터 제거(duplicate elimination by averaging)와 샘플링 시간간격 조절에 의한 데이터 축약(data reduction by sampling time interval) 방법이 모색되었다.

둘째는, 최근 통계학, 의학 및 전산학 등에 많이 사용되는 기계학습(machine algorithm)에 기초를 둔 비모수법 형태의 SVM(서포트벡터 기계학습)과 EBM(양상불 부스팅법)으로 기존에 많이 통용되어 왔던 모수법인 BLM(이분형 로지스틱 모델)과 비교하여 합류부에서의 차로변경 여부에 대한 예측을 상호비교 하였다. 참고로 BLM은 여러 매개변수의 선형조합으로 정의되는 확률함수로 차로변경 여부를 판정한다.

셋째는, MIT공대의 Choudhury가 2007년에 제안한 예상간격모델(anticipated gap model)로 기존에 사용되었던 합류차량과 주변차량이 등속운동을 한다는 가

정하에 계산되는 인접간격(adjacent gap)에 합류차량을 중심으로 선행 및 후행차량이 차로변경을 하는 동안 가속운동을 한다는 가정하에 주변차량의 동적효과에 의한 추가적인 간격변동을 고려하는 방식이다. 차로변경을 결정하는 임계간격(critical gap)은 차로변경에 큰 영향을 주기 때문에 새롭게 제안된 모델을 반영하여 분류성능 효과를 확인하고자 하였다.

한편 제안하는 기계학습 기반 분류기들의 확장성을 보이기 위해 분할표에서 True-Positive(분류기에 의해서도 합류판정, 측정값도 합류판정인 차량)로 분류된 차량만을 갖고 실제 미시교통해석(microscopic traffic analysis)을 수행하였다. 차량궤적의 그래프작성을 통해 차로변경 결정(decision making process)을 구분하고 합류행태를 직접합류(direct merging), 추적합류(chase merging) 및 기타합류로 분류하였다. 이 목적을 위해 차량궤적을 구분하기 위해 K-means 클러스터링 알고리즘이 적용되었다. 각각의 분류기에서 생성된 합류차량의 횡방향 변위와 실제 측정치와의 오차분석을 수행하였다. 특히, 직접합류의 경우 BLM은 많은 오차를 SVM과 EBM에 비해 보여주었다. 또한, 샘플링 시간간격을 통한 데이터 축약기법에 의해 데이터의 횡방향 변위와 시간에 대한 분포를 도시하여 분류기의 성능평가 및 오차분석이 수행되었다.

자세한 차량궤적의 데이터는 NGSIM(Next Generation Simulation) US-101 구간 데이터가 사용되었다. 여러 분석과 평가를 통해 다음과 같은 결과가 도출되었다. 즉, 기계학습 기반의 비모수 분류기는 NGSIM 데이터의 불균형 정도에 상관없이 기존의 모수법 기반의 분류기에 비해 개선된 예측정확도를 보여준다. 그리고 데이터 샘플링기법과 예측간격모델(anticipated gap model)은 데이터 불균형을 완화시키고 데이터의 질을 높여 주는 것으로 판단된다.

핵심용어 : SVM(서포트벡터 기계학습), EBM(양상불 부스팅법), 차로변경 예측, 데이터 샘플링기법, 예측간격모델(Anticipated Gap Model), K-means 클러스터링, Hampel 필터링, 차로변경 결정(Decision Making Process)

학번: 2015-21300