8-1-2014

A Comparative Study: Utilizing Data Mining Techniques to Classify Traffic Congestion Status

Abbas Mirakhorli
University of Nevada, Las Vegas, mirakhorli.a@gmail.com

A COMPARATIVE STUDY: UTILIZING DATA MINING TECHNIQUES TO CLASSIFY TRAFFIC CONGESTION STATUS

By

Abbas Mirakhorli

Industrial Engineering

Faculty of Industrial Safety and Health, Shaheed Beheshti University of

Medical Sciences and Health services, Tehran, Iran

2008

A thesis submitted in partial fulfillment of the requirements for the

Master of Science in Engineering - Civil and Environmental Engineering

Department of Civil and Environmental Engineering and Construction

Howard R. Hughes College of Engineering

The Graduate College

University of Nevada, Las Vegas

August 2014

UNLV
UNIVERSITY OF NEVADA LAS VEGAS

THE GRADUATE COLLEGE

We recommend the thesis prepared under our supervision by

Abbas Mirakhorli

entitled

A Comparative Study: Utilizing Data Mining Techniques to Classify Traffic Congestion Status

is approved in partial fulfillment of the requirements for the degree of

Master of Science in Engineering -- Civil and Environmental Engineering

Department of Civil and Environmental Engineering and Construction

Alexander Paz, Ph.D., Committee Co-Chair

Brendan Morris, Ph.D., Committee Co-Chair

Mohamed Kaseko, Ph.D., Committee Member

Pramen Shrestha, Ph.D., Committee Member

Venkatesan Muthukumar, Ph.D., Graduate College Representative

Kathryn Hausbeck Korgan, Ph.D., Interim Dean of the Graduate College

August 2014

ABSTRACT

A Comparative Study:

Utilizing Data Mining Techniques to Classify Traffic

Congestion Status

By

Abbas Mirakhorli

Dr. Alexander Paz, Examination Committee Chair

Assistant Professor

Department of Civil and Environmental Engineering and Construction

University of Nevada, Las Vegas

Performance measure is a process of evaluating and quantifying a system. Performance measure provides us with information about how good a system is working and how well the predefined goals are met. In order to analyze the performance of a transportation system, the traffic data such as speed, volume, occupancy and travel time of the system need to be collected. These data will generate valuable historical database that can be used to develop models to improve the quality of service of transportation system. The performance measures in transportation studies can be categorized to following main groups: Congestion, Mobility, Accessibility, Reliability, Safety and Environmental. Traffic congestion is one the important issues in any transportation system. Growing congestion in urban transportation network has enforced significant economic burdens to our current society. It causes waste of time, money, fuel and energy for the commuters

and consequently impacting daily life of people in the society. Based on 2011 Congested Corridors Report presented by Texas A& M Transportation Institute, traffic congestion incurred $121 billion cost for drivers. Based on this report, 5.5 billion additional hours are wasted waiting in traffic in 2011. It means $818 additional fuel and time cost for each commuter. Being aware of the status of congestion in future can help, decision makers, intelligent systems and apps improve their accuracy and help commuters in their travel routing. To achieve these goals accurate traffic status classification techniques is required. Achieving higher accuracy is still one of the influential driving factor for research in this area. The objective of this thesis is to utilize data mining techniques to classify traffic status to congested or non-congested for some point of time in future based on historical traffic parameters (Vehicle Count, Occupancy, Speed). Moreover, to compare the performance of different data mining techniques on this problem. This dissertation examined several classification techniques including J48 Decision Tree, Artificial Neural Network, Support Vector machine, PART and K-Nearest Neighborhood to classify future traffic status to Congested or Non-congested. The one minute traffic data from I-15 Northbound from I-215 up to Desert Inn, Las Vegas, NV were used to run these experiments. Based on the comparison of these algorithms, the J48 algorithm has the best performance.

ACKNOWLEDGEMENTS

I would like to give special thanks to Nevada Department of Transportation for sponsoring this project. I would also like to express my appreciation to Dr. Alexander Paz and Dr. Brendan Morris for their support and attention toward the project. I would like to thank them for their continuous support during my period of study. Without their support and knowledge this thesis could not have been a success. I would also like to express my thanks to other members of my committee: Dr. Mohamed Kaseko, Dr. Pramen P. Shrestha for their time, effort and support. Additionally, I cannot forget my colleagues and friends in TRC for their help and support in completing this research project. Finally, I owe my deepest gratitude to my family who has always been a source of inspiration and support for me.

DEDICATION

This thesis is dedicated to my family.

Thank you all for your love, support and care.

I love you all.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Background

Performance measure is a process of evaluating and quantifying a system. Performance measure provides us with information about how good a system is working and how well the predefined goals are met. The decision makers can also make proactive decisions based on monitoring performance measures. In order to analyze the performance of a transportation system, the traffic data such as speed, volume, occupancy and travel time of the system need to be collected. These data will generate valuable historical database that can be used to develop models to improve the quality of service of transportation system. The performance measures in transportation studies can be categorized to following main groups: Congestion, Mobility, Accessibility, Reliability, Safety and Environmental.

1.2 Problem Statement and Objective

Traffic congestion is a main issue in any transportation system. The decision makers have to take into account congestion in their transportation planning. The commuters have to deal with congestion in their every day trip. Traffic congestion is one the important issues in any transportation system. Growing congestion in urban transportation network has enforced significant economic burdens to our current society. It causes waste of time, money, fuel and energy for the commuters and consequently impacting daily life of people in the society. Based on 2011 Congested Corridors Report presented by Texas A& M Transportation Institute, traffic congestion incurred $121

billion cost for drivers. Based on this report, 5.5 billion additional hours are wasted waiting in traffic in 2011. It means $818 additional fuel and time cost for each commuter.

Being aware of the status of congestion in future can help, decision makers, intelligent systems and apps improve their accuracy and help commuters in their travel routing. To achieve these goals accurate traffic status classification techniques is required. Achieving higher accuracy is still one of the influential driving factor for research in this area.

The objective of this thesis is to utilize data mining techniques to classify traffic status to congested or non-congested for some point of time in future based on historical traffic parameters (Vehicle Count, Occupancy, Speed). Moreover, to compare the performance of different data mining techniques on this problem. This dissertation examined several classification techniques including J48 Decision Tree, Artificial Neural Network, Support Vector machine, PART and K-Nearest Neighborhood to classify future traffic status to Congested or Non-congested. The one minute traffic data from I-15 Northbound from I-215 up to Desert Inn, Las Vegas, NV were used to run these experiments. Based on the comparison of these algorithms, the J48 algorithm has the best performance.

1.3 Organization of thesis

This thesis is composed of four chapters: (i) Introduction, (ii) Literature Review, (iii) Methodology, (iv) Conclusions and Future research. The first chapter provides an overview of the study including background, problem statement and objectives. Chapter two presents a literature review about transportation performance measures. Experimental

result associated with each method is presented in chapter 3 and this study ends with conclusions and future research which is presented in chapter 4.

CHAPTER 2

LITERATURE REVIEW ON TRANSPORTATION PERFORMANCE MEASURES AND CONGESTION ANALYSIS

According to the United States Department of Transportation's (U.S. DOT, 2003) Strategic Plan two major goals of U.S. transportation development are to "support a transportation system that sustains America's economic growth" and to "shape an accessible, affordable, reliable transportation system for all people, goods, and regions". In response to the U.S. DOT's strategic plan, the Federal Highway Administration (FHWA) has also enacted its own strategic plan to ensure the satisfaction of the goals. In order to gain the above-mentioned goals, transportation specialists have been trying to improve the efficiency of transportation system for many years. The first step in determining the performance measure of a transportation system is to identify goals and objectives. The selection of goals and objectives should directly reflect the customer needs and the economic costs associated with it. Transportation performance measures can be categorized to following measures:

- Congestion
- Mobility
- Accessibility
- Reliability
- Safety
- Environmental

2.1 Congestion:

There have been a lot of definitions for traffic congestion in the literature (Aftabuzzaman, 2007). The research presented a report to propose a framework for developing a congestion performance measures. In this report some definition for congestion which is defined by previous studies are presented. Based on those researches congestion refers to a situation in which the number of vehicles increases more than the capacity of the roadway resulting in speeds that are slower than the normal or free flow speed. Sarah and Michael (Sahara & Michael 2003) presented a report to specify a performance measure to show the congestion levels on main corridors of Virginia. Moreover, A review of procedures and examples of application of geographic information system (GIS) technology for development of congestion management systems (CMSs) is presented by Quiroga (Quiroga 2000). The paper analyzed different transportation performance measures. Based on this paper the travel time is the most beneficial and understandable performance measure. A lot of performance measures exist in the literature to measure and track congestion. The Texas Transportation Institute (TTI) is a leader in developing measurements for determining congestion. Congestion Measures can be subdivided into Mobility Measures and Reliability Measures. Thus this measure is presented in this study through mobility and reliability.

2.2 Mobility:

Mobility is the ability to easily move and transport product and services between different locations. Average speed is considered as the main factor for mobility measurement (Litman, 2003; Sen et al., 2011). Litman (Litman 2003) measured the performance of a transportation system taking in to account mobility, traffic and

accessibility. The research not only considered the state of mobility management practice throughout Texas, but also overviewed national best practices in mobility management. The research also presents examples of applied mobility management and a series of performance measures which was based on the type and level of program implemented. The Transit Cooperative Research Program (TCRP) Research Results (Simon 1997) examines the impact of implementation of Americans with Disabilities Act (ADA) Para transit requirements on public transportation. The National Council on Disabilities produced a report (Frieden, 2005) that revealed the limitations imposed on people with disabilities due to lack of transportation, which in turn affected their ability to work, socialize, and even attend spiritual events. The research highlighted the difficulties of individuals with disabilities compared to the general public's transportation choices regardless of where they live. Texas Transportation Institute (Texas Transportation Institute, 2005) presented some clue points for estimating mobility in urban areas. The conclusion of their report is that there is no single measure satisfying all the needs. The report concludes that thers no single measure that can represent and quantify mobility status thoroughly. Congestion is a measure of how movement is constrained by too many users for the capacity of the system. Thus congestion is in many respects the inverse of mobility (though mobility can be low even on an uncongested system if there is insufficient network).

These are the five most common measures for mobility:

- Volume-to-Capacity Ratio (V/C Ratio): the volume divided by capacity. This criterion is often used for the Level of Service (LOS) calculations.

- The Level of Service (LOS): it is graded from A to F, which A means free flow and F means very congested. These grades interval means how well an intersection is serving its traffic. LOS is based on a volume to capacity (v/c) ratio and has long been used as the primary measure of congestion for planning purposes. In (V/C) ratio, The Volume is often estimated as the 30th yearly highest volume available.
- Travel Time Index: ratio of average peak travel time to an off-peak (free-flow) standard, in this case 60 mph for freeways. For example, a value of 1.20 means that average peak travel times are 20% longer than off-peak travel times.
- Travel Delay: the amount of extra time which is needed for traveling due to congestion.
- Percent of Congested Travel: the congested vehicle-miles of travel divided by total vehicle-miles of travel. This measure is actually a relative measure of the amount of travel affected by congestion.

Table 1 summarizes the studies on mobility measures.

Table 1. Mobility

Year	Authors	Notes
1997	TCRP	Examines the impact of implementation of ADA Para transit requirements on public transportation
2002	Black et al	preserve a mobility management program
2003	Litman	Considering mobility, traffic and accessibility performance measures
Year	**Authors**	**Notes**
2005	Texas Transportation Institute and Texas A&M University	Presenting some clue points for estimating mobility in urban areas
2005	Frieden	Considered the mobility management plan for people with disabilities.
2010	Williamsa and Saggerman	A guide for review and evaluation of local mobility plan
2011	Lomax et al	Focusing on urban mobility information affecting traffic delays
2011	Lalita et al	Considered the state of mobility management practice throughout Texas

2.3 Accessibility:

Accessibility is a measure or indicator of the performance of transportation systems in serving individuals living in a community. Farrington and Farrington (Farrington and Farrington , 2005) defined accessibility as ‘‘the ability of people to reach and engage in opportunities and activities’’ while Pirie (Prie, 1981) defined accessibility as being similar to reachability and convenience. The paper meant how easily the infrastructures can be reached by people. Gulliford et al (Gulliford et al., 2002) considered the accessibility from two different perspectives. ‘‘having access’’ that refers to availability of services and ‘‘gaining access’’ that refers to individual’s ability to utilize the available services. The literature presented various other approaches to conceptualize and define access. Aday and Andersen (Aday and Andersen, 1974) presented a framework that identifies different aspect of accessibility like financial, informational and behavioral. The authors distinguish between socio-economic and spatial perspectives of accessibility and relate different aspects of accessibility to system level and individual level factors. The number of goods transferred and number of people accessing the system are considered to be indicators of transportation accessibility by Bertini et al (Bertini et al., 2000).

Eisele, et al (Eisele, et al., 2005) described the importance of access management and how the use of raised medians has an effect on access management. They presented that net delay can be reduced significantly by using a raised median.

Five major theoretical approaches for accessibility measurement found in the literature are as follows(Koenig, 1978; Morris et al., 1978):

1) travel-cost approach : The first class of accessibility indicators embodies those measuring the ease with which any land-use activity can be reached from a location using a particular transportation system.
2) gravity or opportunities approach : Indicators based on spatial opportunities available to travelers are among the first attempts to address the behavioral aspects of travel.
3) constraints-based approach : based on the fact that individual accessibility has both spatial and temporal dimensions. Opportunities or potential to opportunities for an individual are not only constrained by the distance between them, but also by the time constraints of the individual.
4) utility-based surplus approach : This class of accessibility indicators is another attempt to include individual behavior characteristics in accessibility models. Utility-based indicators have their roots in travel demand modeling
5) composite approach : Representation of the multiple-purpose property of trips is lacking in the utility-based measures. Space-time and the utility-based models are combined with each other to develop composite approach

Geurs and Ritsema (Geurs and Ritsema, 2001) presented a literature study and three case studies trying to review accessibility measures for their ability to evaluate the accessibility impact of national land use and transport scenarios and related social and economic impacts. Murray and Wu (Murray and Wu, 2003) have presented two spatial optimization models for addressing accessibility in the provision of transit service. These models simultaneously take into account access and geographic coverage. Table 2 summarizes the studies on accessibility measures.

Table 2. Accessibility

Year	Authors	Notes
1974	Aday and Andersen	A framework that identifies different aspect of accessibility
1981	Pirie	defining accessibility as how easily the infrastructures can be reached by people
2001	Geurs and Ritsema	Presenting a literature study and three case studies trying to review accessibility measures
Year	**Authors**	**Notes**
2002	Bertini et al	Considering the number of goods transferred and number of people accessing the system to be indicators of transportation accessibility
2002	Shaw	Percentage of urban population within X mile of transit is used to evaluate the transit service accessibility
2002	Gulliford et al	considering the accessibility as having access and gaining access and presenting literature review about other approaches to conceptualize and define access
2003	Murray and Wu	Presenting two spatial optimization models for addressing accessibility in the provision of transit service

2.4 Reliability:

Reliability is defined as day-to-day change in travel times experienced by travelers. For a transportation system, the reliability is usually associated with unprecedented delay. The two methods to measure travel time reliability are the 90th or 95th percentile travel time's method and planning time method. The 90th or 95th percentile travel time's method, predicts delay on specific routes during the heaviest traffic days (US Department of Transportation (2005). The one or two bad days each month mark the 95th or 90th percentile, respectively. The buffer index represents the amount of extra time which is needed to be added to average travel time to ensure on-time arrival. For example, a buffer index of 40 percent means that for a trip that usually takes 20 minutes a traveler should budget an additional 8 minutes to ensure on-time arrival most of the time. The 8 extra minutes is called the buffer time. Therefore, the traveler should allow 28 minutes for the trip in order to ensure on-time arrival 95 percent of the time. The planning time index estimates the total amount of time needed to ensure on-time arrival. The buffer index represents the additional travel time that is necessary for on-time travel, but the planning time index estimates the total travel time that is necessary. For example, a planning time index of 1.60 means that for a trip that takes 15 minutes in light traffic a traveler should budget a total of 24 minutes to ensure on-time arrival 95 percent of the time.

The measures that look the most promising or may provide some good material for other analyses are as follows (Lomax et al., 2003):

- Travel time window: The standard deviation of travel time or travel rate can be combined with the average for any of several measures to create a variation or reliability measure.

  Travel Time Window = Average Travel Time ± Standard Deviation

- Percent variation: The average and standard deviation values can also be combined in a ratio to produce a value that the 1998 California Transportation Plan calls percent variation: CV=(Standard Deviation)/(Average Travel time) ×100

- Misery Index: This measure focuses on the length of delay of only the worst trips. The average travel rate is subtracted from the upper 10%, 15% or 20% of travel rates to get the amount of time beyond the average for some amount of the slowest trips.

  $$Misery\ index = \left[\frac{\begin{matrix}average\ of\ the\ travel \\ rate\ for\ longset \\ 20\%\ of\ the\ trip\end{matrix} - Average\ travel\ rate\ for\ all\ trip}{Average\ travel\ rate}\right]$$

- Buffer time = this measures the amount of extra time needed to be on time for 95% of the trips.

  Buffer Time = 95% percent travel time for a trip-Average Travel Time

- Buffer Time Index: Using the Buffer Time concept and the travel rate simultaneously (in minutes per mile), rather than average travel time, can address the concerns about identifying an average trip. This measure is used as the reliability performance measure in the Mobility Monitoring Program reports.

$$\text{Buffer time index} = \left(\frac{95th\ precentile\ travel\ rate - average\ travel\ time\ rate\ (in\ minutes\ per\ mile)}{Average\ travel\ rate\ (in\ minutes\ per\ mile)}\right) \times 100$$

- Variability Index = the index is a ratio of peak to off-peak variation in travel conditions. The index is calculated as a ratio of the difference in the upper and lower 95% confidence intervals between the peak period and the off-peak period (Equation 3).

$$Variability\ index = \frac{Difference\ in\ peak-period\ confidence\ intervals\ (Upper\ 95\%\ value - Lower\ 95\%\ value)}{Difference\ in\ off\ peak-period\ confidence\ intervals\ (Upper\ 95\%\ value - Lower\ 95\%\ value)} \times 100$$

- Planning Time Index = the upper end of the Buffer Time Index can also be concerned as an useful measure in some situations. The 95th percentile Travel Time Index or the travel rate (expressed in minutes per mile) is a good measure to estimate of travel time budget and is calculated as part of the Buffer Time Index process. Planning time index is relatively easy to communicate and is a good estimate of trip planning measure for trips that require on-time arrivals.

  Planning Time Index = 95th Percentile Travel Time Index (of all peak period travel)

- Florida Reliability Method: The Florida reliability method uses a percentage of the average travel time in the peak to estimate the limit of the acceptable additional travel time range. The sum of the additional travel time and the average time defines the expected time.

  Florida Reliability Statistics (% of unreliable trip): 100% - (percent of trip with travel time greater than expected) = 100% - (percent of trips with travel rate

greater than the average for the time period plus 5%, 10%, 15% and 20% of the average).

- On-Time Arrival: A concept similar to the Florida method uses an acceptable "lateness threshold" of some percentage to indicate the percentage of trip travel times that can be termed reliable. This measure is used in a variety of travel modes and services and might be particularly useful in cross-modal comparisons.

  On time-Arrival = 100% - (Percent of travel rate greater than 110% of the average travel rate) = 100% - (percent of daily peak period travel rate average that are greater than 110% of average peak period travel rate)

Table 3. Reliability

Year	Authors	Notes
2005	US Department of Transportation	Travel Time reliability
2005	Economic Development Research Group	Examines the importance of travel time reliability
2012	Douglas et al	Developing a travel time reliability model

2.5 Safety:

Safety is the state of being "safe". Safety is an inherent performance measure for transportation. A transportation system without high safety is unreliable and inefficient. The most common indicators of safety are fatalities per 100 million vehicle-mile of travel

and number of accidents per 100 million vehicle-miles of travel *(50)*. Different modes of transportation have different causes to influence safety, so safety measures are different according to the mode for different modes in the transportation system. For example, for highways, the measure is usually the number of fatalities within a certain length of Vehicle miles travel; whereas for airborne transportation, the measure is usually identified by fatal aviation accidents per 100,000 departures (Dumbaugh & Meyer, 2003). In general, accident rates, fatality rates, and injury rates are directly related to the loss due to accidents. Besides these, transportation is also associated with many other safety measures: for example, average time between notification and response/arrival clearance, total duration of incidents, etc. The number of accidents, fatalities, and injuries are some appropriated performance measures to evaluate the safety of a transportation system. The National Highway Traffic Safety Administration (NHTSA) and the Governors Highway Safety Association (GHSA) have presented a minimum set of performance measures to be used for safety plans and programs (Hedlund, 2008). In this research Performance measures were considered for the following ten areas. The safety plan contains 14 measures: ten core outcome measures, one core behavior measure, and three activity measures. Botha (Botha, 2005) conducted a research about measuring road traffic safety performance. The purpose of the paper is to provide some information about the measures associated with road traffic safety. The current measures are mainly based on un-planned random incidents: crashes and causalities. The paper developed road safety index (RSI) which can be used in future as the main indicator of the level of safety on the road and street network. Susan et al (Herbel et al., 2011) conducted a research about Safety Performance Measures for the Transportation Planning Process. Their research

presented a tool to help transportation decision makers identify safety performance measures as a part of the transportation planning process. Table 4 summarizes the studies on safety measures.

Table 4. Safety

Year	Authors	Notes
2003	Dumbaugh and Meyer	Presenting the indicators of safety
2005	Botha	measuring road traffic safety performance
2008	Hedlund	Presenting set of performance measures to be used for safety plans and programs
2011	Susan et al	Safety Performance Measures for the Transportation Planning Process

2.6 Environmental:

The impact of transportation system on human and natural environment is one of the important in transportation planning. Because of increasing costs of environmental operations, selecting an effective tool for measuring environmental performance has received more attention these years. Estimating the emissions from all the mobile sources is one of the most important performance measures for the system. The DOT uses “Tons (in millions) of mobile source emissions from one-road vehicles” as one of the major performance measures (Gudmundsson, 2000). Noise is another unwanted effect of transportation. Aviation and railways are main contributors of noise pollution. Based on

Global Environmental Management Initiative (Global Environmental Management Initiative, 1998), Environmental indicators are classified to lagging and leading indicators. Most environmental metrics programs will contain both types of measures.

Lagging Indicators

Lagging indicators are the mostly used metrics. These indicators measure the results of environmental practices or operations. The performance measures consist of the following data: number of accidents or lost work days, tons of generated waste, number of fines and violations, or pounds of produced package.

Leading Indicators

The Leading indicators evaluate the amount of improvement in environment made by implemented policies. As an instance, number of health and safety compliance is used instead of numbers of fines and violations. Usually by implementing corrective programs to identify and omit the environmental problems, the amount of fines and violation will be decreased. Developing metrics for sustainable transportation is another issue in environmental performance measurement. Zeng et al (Zeng et al., 2013) presented a process for developing such metrics in the form of a composite index. His research provides guidance for selecting an appropriate index and developing the new index. Cory Searcy (Searcy, 2012) conducted a research in design, implementation and evaluation of Sustainability Performance Measurement. Moreover, the paper presents a literature review of published paper between 2000 and 2010. National Cooperative Highway Research Program (NCHRP) Project 25-25, Task 23 (US. Environmental Protection Agency, 2011) presented an instruction for the design and implementation of environmental performance measurements for state departments of transportation (DOT).

The research also presented practical procedures to integrate environmental measurements into agency practices and decision-making process. Table 5 summarizes the studies on environmental measures.

Table 5. Environmental

Year	Authors	Notes
1998	Global Initiative	Classifying environmental indicators as lagging and leading indicators
2000	Ministry of Environment and Energy	Presenting indicators and Performance Measures for transportation, environment, and sustainability
2003	U.S. DOT	Presenting number of people who are exposed to significant noise
2011	U.S. Environmental Protection Agency	Instruction for the design and implementation of environmental performance measurements
2012	Cory Searcy	Evaluation of Sustainability Performance Measurement
2013	Jason Zeng et al	Providing guidance for selecting an appropriate index and developing the new index

2.7 Congestion Analysis

Congestion analysis is a topic which is drawing research's attention during last decade. The researchers were trying to predict status of the highways whether there is congestion

or not. These researches were able to classify real-time status of congestion. Yu et al (Yu et., al 2010) presented a logistic regression model to measure congestion intensity. Their model can be used to specify the intensity of traffic congestion for different roadways. Hongsakham et al (Hongsakham et al., 2008) developed a technique based on neural network to estimate road traffic congestion levels. Neural network was then trained and tested. Their congestion estimation model had a recall of 79.43% and precision ranging from 73.53% to 85.19%. The studies in Pongpaibool et al (Pongpaibool et al., 2007) utilized fuzzy logic and neuro-fuzzy techniques to estimated the congestion level using data from traffic camera. The proposed techniques had accuracy of 88% and 75% respectively. Porikli and Li, (Porikli and Li, 2004) used hidden Markov approach to estimate congestion status. The accuracy of their developed model is 95%. Tsai et al (Tsai et al., 2011) developed a traffic congestion classification framework that classifies congestion to four level accuracy. Automatic roadway detection, bidirectional roadway analysis and Virtual detector setting method are presented as the three procedure of their framework to classify congestion status. The accuracy of their approach was 93.2%. Lu and Cao (Lu and Cao, 2003) also used fuzzy techniques to detect and evaluate congestion status. Elhenawy and Rakha (Elhenawy and Rakha, 2014) presented a Machine Learning Classifiers based on adaptive boosting method to predict the status of congestion. The algorithm showed high performance for real time congestion prediction. The true positive and false positive prediction rates are 0.99 and 0.01 respectively. Zhan-quan et al (Zhan-quan et al., 2012) used support vector machine algorithm to predict the real-time congestion status. The precision of their algorithm was 94%. They used speed, volume and occupancy as their features. Wang et al (Wang et al., 2006) combined clustering and

classification technique to classify the real-time congestion status. They used decision three to classify the real-time prediction. Their developed classification algorithm was 99.3% accurate.

CHAPTER 3

METHODOLOGY

3.1 Data Classification

Data classification concept is a two-step procedure in which, the first step (figure 1) of this procedure tries to develop a model that represents a predetermined set of data classes or concepts and in the second step (Figure 2), the developed model is used for classification (Dunham, 2003). In classification problems, each record belongs to a prespecified class. Figure 1 and figure 2 shows the two step of classification process. Figure 1 shows learning process in which classification algorithm analyze the training data set. This example classifies credit cart status to high or excellent.

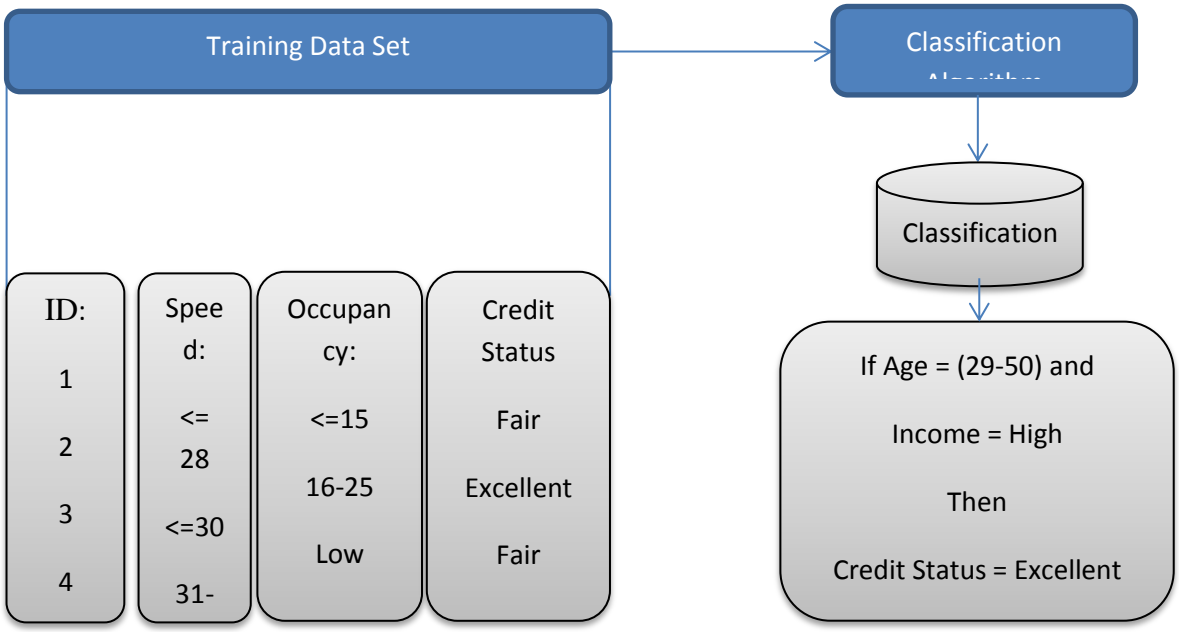

Figure 1. Training Phase

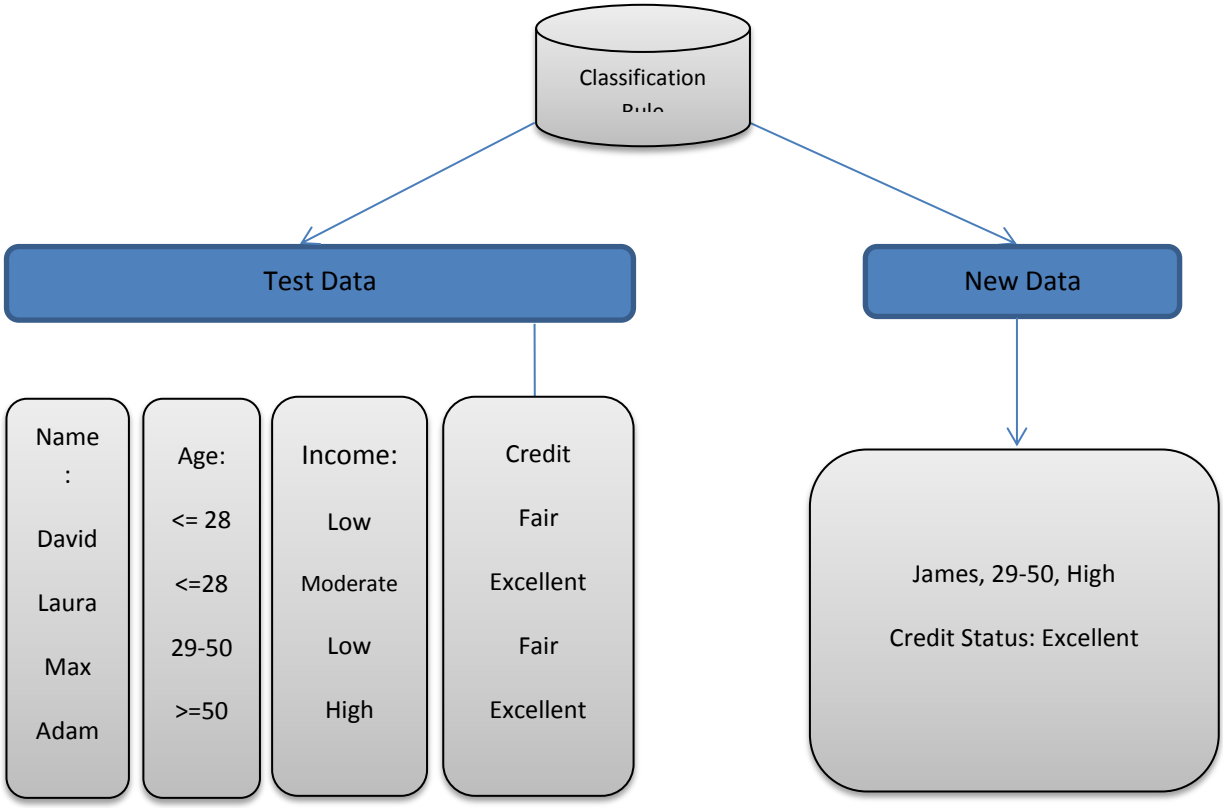

Figure 2. Test Phase

3.2 Classification Techniques

Classification techniques try to specify a certain outcome based on a given input features. The techniques try to find out relationships between the attributes that would make it possible to predict the outcome. The algorithm analyses the input and produces a classification algorithm. There are generally two types of learning process in classification domain. In supervised classification, a label for each pattern is provided and the algorithm tries to learn the rule form labeled training data, While In unsupervised classification there is no explicit label, and the system forms clusters of the input patterns. A good classification model is a model that fits the training data thoroughly and can precisely classify the new unseen data.

In order to solve a classification problem, a training set with a known class labels should be provided. The training data set is utilized to develop a classification model that is applied to the test data set. The developed classification model is evaluated based on the counts of test records correctly and incorrectly classified by the model. These counts are tabulated in a table known as a confusion matrix. Table 1 presents the confusion matrix for a binary classification problem. The values fij in this table represents the number of records from class i classified to be of class j. As an instance, f01 is the number of records from class 0 incorrectly classified as class 1. Based on the values in the confusion matrix, (f11 + f00) and (f10 + f01) is the total number of correct and incorrect predictions made by the model respectively.

Table 6. Confusion Matrix

confusion matrix		**predicted**	
		Class = 0(negative)	**Class = 1(positive)**
Actual	**Class = 0 (negative)**	*F00*	*F01*
	Class = 1(positive)	*F10*	*F11*

By interpreting a confusion matrix we can determine how well a classification model performs. Interpretation is done by summarizing confusion matrix information with indices. This can be done using a performance metric such as Recall (True Positive) and Precision, which are defined as follows:

The recall or true positive rate (TP) is the proportion of the positive cases that were correctly specified. Recall is calculated as follows:

Recall: $\frac{F11}{F11+F10}$

Precision: proportion of the predicted positive cases that were correct. Precision is calculated as follows:

Precision: $\frac{F11}{F01+F11}$

Most of the classification algorithms try to develop a model that attains the highest recall and precision.

3.3 Data Preparation

In order to improve the accuracy and efficiency of the classification procedures, the following preprocessing steps may be applied to the data:

- Data cleaning: The missing values should be removed from the data.
- Relevance analysis: any redundant or irrelevant feature should be removed from the learning process.
- Data transformation: Some attribute can be manipulated to extract some other information from them.

This section presents the application of the methodology to a real-life freeway corridor in Las Vegas, Nevada. The data are collected in I-15 Northbound from I-215 up to Desert Inn. In this study we use the one-minute traffic data downloaded from our new website. This data includes speed, number of vehicles and occupancy. The schema of the data set in presented in Figure 3.

	A	B	C	D	E
1	id	time	count	occupancy	speed
2	70_2_21	2/1/2014 0:00	3.2	1.2	81.2
3	70_2_21	2/1/2014 0:01	3.8	2	82.4
4	70_2_21	2/1/2014 0:02	4.4	2	66.8
5	70_2_21	2/1/2014 0:04	4	1.4	81
6	70_2_21	2/1/2014 0:05	6	2	66.8
7	70_2_21	2/1/2014 0:06	5.2	1.6	65
8	70_2_21	2/1/2014 0:07	4.2	1.8	76.8
9	70_2_21	2/1/2014 0:08	5.6	2.6	66.6
10	70_2_21	2/1/2014 0:09	6.6	2.8	67.4
11	70_2_21	2/1/2014 0:10	6.6	2.4	69.8
12	70_2_21	2/1/2014 0:11	4.6	2	69.8
13	70_2_21	2/1/2014 0:12	7	3.2	69.2
14	70_2_21	2/1/2014 0:13	7.6	3.4	69.4
15	70_2_21	2/1/2014 0:14	6.6	2.4	69.8
16	70_2_21	2/1/2014 0:15	6.2	2.8	70
17	70_2_21	2/1/2014 0:18	4.6	2.4	70.8
18	70_2_21	2/1/2014 0:19	6.6	3.2	68.2
19	70_2_21	2/1/2014 0:20	8	4	68.6
20	70_2_21	2/1/2014 0:21	7	2.4	69.8
21	70_2_21	2/1/2014 0:22	6.6	2.6	69.6
22	70_2_21	2/1/2014 0:23	4.4	2.2	69.4

Figure 3. One-minute data

The count, occupancy and speed associated with each time interval is presented. Classifying the next state of the traffic is the goal of this research. J48 Classification technique is used to reach this goal. In order to use this classification method we need to generate a training data set that let us know about the status of traffic (whether it was Congested or Non-congested). In order to increase the accuracy of the classification technique both real-time and historical data are put in our training data base. In our training data set, we need to label our record. In our study we label our record as congested or non-congested. Congested refers to the condition that there is a congestion and non-congested refer to the condition that there is no congestion. There are three main

approaches for labeling training data set as presented in the literature. These approaches are as follows:

Watching video data:

It is the commonly used and reliable approach (Yu ., et al 2010) . This approach can be used for real-time classification and future prediction.

Threshold:

In this approach (Tsai ., et al 2011, Elhenawy, Rakha 2014) a threshold for traffic parameter like speed is chosen. And when the speed falls below the threshold we label it as congestion. This approach can be used for future traffic prediction. We use the real-time and historical data to predict the future traffic status for next 1, 2,..., 5 minute. . We develop a general rule for predicting the congestion status for next minutes.

Clustering (2):

This approach (Wang ., et al. 2006) use clustering for labeling data set with this assumption that the cases with the same traffic status will go to the same classes.

The threshold approach is used in this study. There are different threshold for congestion measurement. Table 2 shows some of these measures (NCHRP report: 398. 1997). The *TSR* performance measure is used in this study.

Table 7. Congestion Performance Measures.

Congestion Performance Measure	Description
Roadway congestion index	This index focuses on the physical capacity of the roadway in term of vehicles. This index measure the congestion by concentrating on daily vehicle miles traveled on roads. $RCI = \frac{(Freeway\ VMT\ per\ lane - mile) * freeway\ VMT + (Principal\ Arterial\ VMT\ per\ lane - mile) +}{13000 * freeway\ VMT + 5000 * Principal\ Arterial\ VMT}$
Travel Speed Rate	Travel speed rate is the rate of reduction in speed from free flow speed due to congestion $TSR = \frac{Free\ flow\ speed\ (spped\ limit) - average\ speed}{Free\ flow\ speed}$ TSR > 0.5 congested condition
Travel time index	This index compares peak period travel and free flow travel while considering for both recurring and incident conditions. This index specify how long it to travel peak hour TTI = (Delay time + travel time)/travel time
Travel delay	Travel delay is the extra amount of time spent traveling due to congested conditions $delay = \frac{daily\ vehicle\ miles\ of\ travel}{speed} - \frac{daily\ vehicle\ miles\ of\ travel}{speed\ limit\ (free\ flow\ speed)}$
Annual Hours of Delay (AHD)	Travel time above a congestion threshold (defined by State DOTs and MPOs) in units of vehicle -hours of delay reduced by the latest annual program of CMAQ projects. $delay = \frac{daily\ vehicle\ miles\ of\ travel}{speed} - \frac{daily\ vehicle\ miles\ of\ travel}{speed\ limit\ (free\ flow\ speed)}$
Buffer index	The buffer index computes the extra percentage of travel time a traveler should consider when making a trip in order to be on time 95 percent of the time $BI = \left(\frac{95th\ precentile\ travel\ rate - average\ travel\ time\ rate\ (in\ minutes\ per\ mile)}{Average\ travel\ rate\ (in\ minutes\ per\ mile)}\right) \times 100$

In the TSR index the free flow speed is equal to posted speed. In our study area the posted speed is 65 mile per hour. If the TSR index is greater than .5 we will label it as

congested. We increase the threshold to 0.6 to be sure of congestion condition (pessimistic view). As mentioned before, the historical data were also included in the training data set to increase the capability and accuracy of the model. Figure 7 shows the real-time data and up to three minutes historical data. Figure 4 shows the labeled data set which is composed of real-time and historical data.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	AG	AH
1	id	time	COUNT	OCCUPANCY	SPEED	COUNT-1	OCCUPANCY-1	SPEED-1	COUNT-2	OCCUPANCY-2	SPEED-2	COUNT-3	OCCUPANCY-3	SPEED-3	TSR	Cong
2	1	3/1/2014 0:00	8	3	65	11	4	66	11	4	67	13	4	66	0.015385	NON
3	2	3/1/2014 0:01	14	9	64	8	3	65	12	4	66	11	4	67	0.015385	NON
4	3	3/1/2014 0:02	13	5	65	14	9	64	8	3	65	12	4	66	0.62451	NON
5	4	3/1/2014 0:03	12	4	66	13	5	65	14	9	64	8	3	65	0.68371	CON
6	5	3/1/2014 0:04	9	2	69	12	4	66	14	5	65	14	9	64	0.613902	CON
7	6	3/1/2014 0:05	11	3	69	9	3	69	12	4	66	14	5	65	0.601724	CON
8	7	3/1/2014 0:06	7	2	68	11	3	69	9	3	69	12	4	66	0.613452	CON
9	8	3/1/2014 0:07	14	4	68	7	3	68	11	3	69	9	3	69	0.602513	CON
10	9	3/1/2014 0:08	13	4	68	14	5	68	7	3	68	11	3	69	0.632341	CON
11	10	3/1/2014 0:09	12	5	67	13	4	68	14	5	68	7	3	68	0.608123	CON
12	11	3/1/2014 0:10	13	5	67	12	5	67	13	4	68	14	5	68	0.613905	CON
13	12	3/1/2014 0:10	13	4	67	13	4	67	12	5	67	13	4	68	0.682531	CON
14	13	3/1/2014 0:12	14	5	65	13	4	67	13	4	67	12	5	67	0.635469	CON
15	14	3/1/2014 0:13	11	4	66	14	5	65	13	4	67	13	4	67	0.345213	NON
16	15	3/1/2014 0:14	11	3	70	11	4	66	15	5	65	13	4	67	0.082051	NON
17	16	3/1/2014 0:15	12	4	71	11	3	70	12	4	66	15	5	65	0.097436	NON
18	17	3/1/2014 0:16	10	3	67	12	4	71	11	3	70	12	4	66	0.34523	NON
19	18	3/1/2014 0:17	9	3	67	10	3	67	12	4	71	11	3	70	0.481237	NON
20	19	3/1/2014 0:18	12	3	65	9	3	67	11	3	67	12	4	71	0.601245	CON
21	20	3/1/2014 0:19	8	3	66	12	4	65	9	3	67	11	3	67	0.625834	CON
22	21	3/1/2014 0:20	12	4	66	8	3	66	12	4	65	9	3	67	0.602691	CON

Figure 4. Labeled training data set

For classifying the future traffic status, the historical data in each point of time that there was congestion has been analyzed. Thus our model is trained to find the rule that exist between traffic parameters that will lead to congestion. In each point of time the model considers up to five minutes historical traffic data to classify the future traffic status.

The classifier vector includes Vehicle count, speed, occupancy along the road segments at the times [t-m+1, t-m+2, …, t-1, t0] where m is the parameter that indicates how far back we need to look in order to classify the future traffic status. The training classifier vector is presented as follows:

$$X_{t_o} = (speed_{\mathrm{t-m+1}}, count_{\mathrm{t-m+1}}, occupancy_{\mathrm{t-m+1}}, speed_{\mathrm{t-m+2}}, count_{\mathrm{t-m+2}}, occupancy_{\mathrm{t-m+2}}, \dots) \quad (6)$$

And the response variable is $Y_{t+\Delta t}$ which classify the state of the traffic in time t+Δt. The training dataset is the collection of all the (X_{t_o} , $Y_{t+\Delta t}$). This training dataset is used to learn the rule that exist between historical traffic parameters that lead to congestion situation. This rule can be used to classify the future traffic status when a new unseen predictor vector arrives.

The abovementioned data set was used to train and evaluate the classification model. Our data set consisted of fifteen attributes. The first three attributes consist of count, occupancy and speed are real-time data that is collected. The second three attributes consist of count-1, occupancy-1 and speed-1 which are the data for first minute in the past and the third three attributes are count-2, occupancy-2 and speed-2 that are the data for the second minute in the past and so on and so forth. The last feature is congestion status which get the values CON or NON representing congestion or non-congestion status respectively. This feature is labeled based on TSR metric as presented above. This research utilized WEKA data mining tool. WEKA is a machine learning tool developed by the University of Waikato. This tool is a collection of machine learning algorithms for data mining tasks.

3.4 Overview of Classification Techniques

There are several different techniques for data classification (Jiawei et al,. 2003). J48 Decision tree, Artificial Neural Networks, Support Vector Machines, PART and K-Nearest Neighborhood algorithms are used here to classify future traffic state. The comparative study shows that the J48 Decision Tree has the best performance in comparison with other methods.

3.5 J48 Decision Tree

Decision Tree is one of the classification techniques that is widely used by researchers. The main reason for popularity of tree-based methods is the fact that, in contrast to other methods, decision trees represent rules. Rules can be easily expressed in a different language that everybody can understand. It can also be expressed in a database access language, like SQL. This algorithm tries to divide the large data into smaller sets until the most homogeneous sets (classes) are generated. In the division process, each attribute is compared to a defined value(s) and separated accordingly. Decision tree can be binary where each attribute value has two options only as presented in figure 3, and the classifier has two classes. Or, it can be N dimension tree which the attribute value is examined against N options, and N classes are resulted as presented in figure 5.

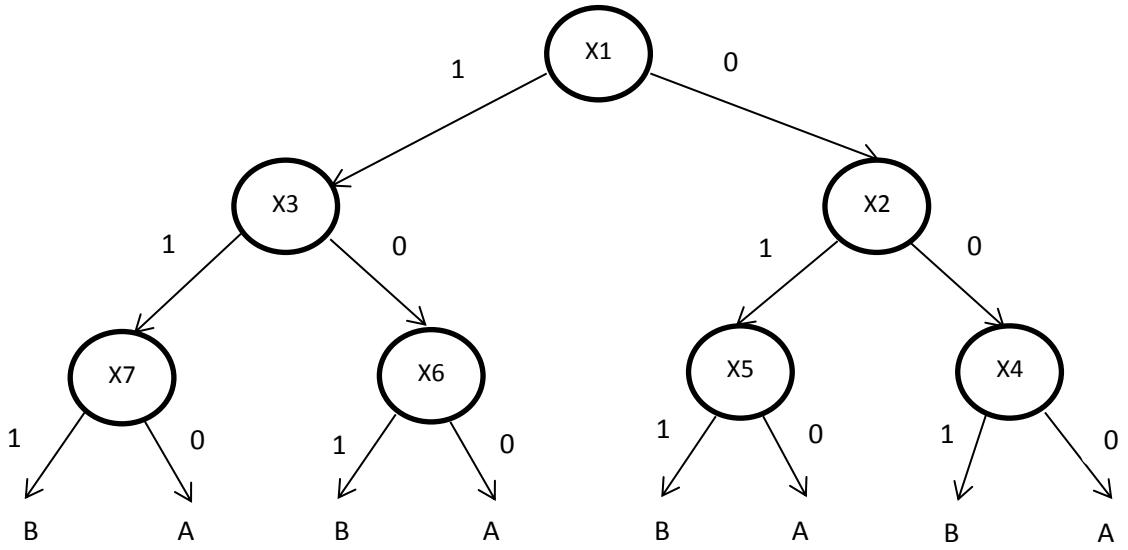

Figure 5. Binary decision tree

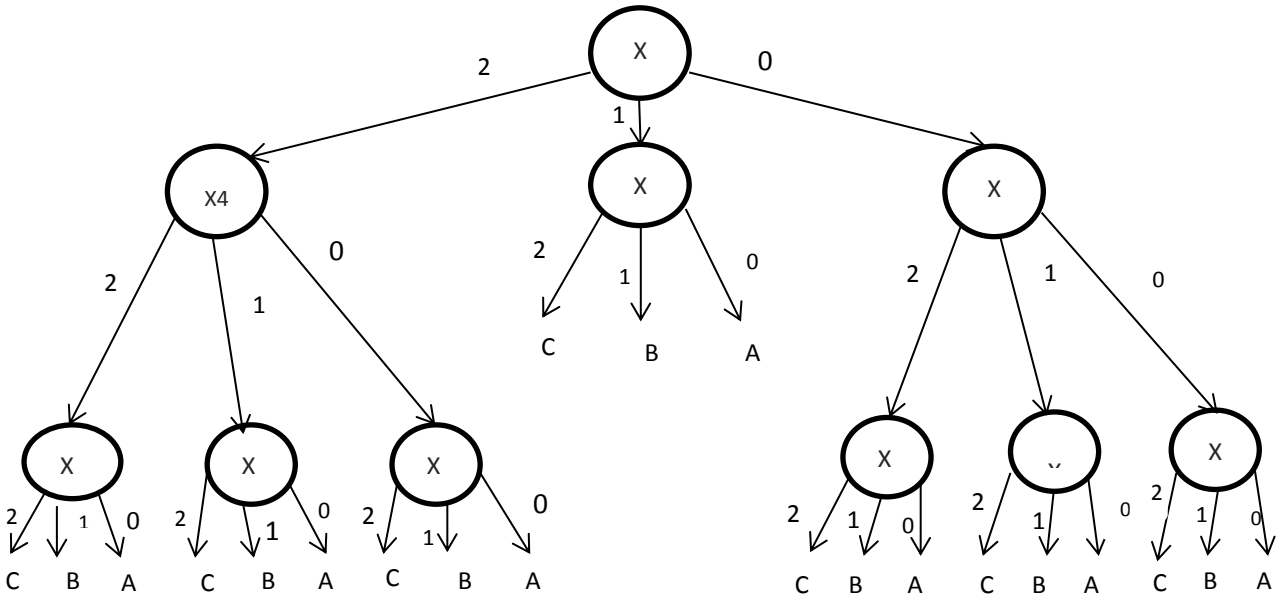

Figure 6. Ternary decision tree

As presented in figures 5 and 6, the decision tree is made of nodes generating a rooted tree. Figures 1 and 2 illustrate how a classification problem is solved by asking a series of questions about the attributes of the test record. Based on an answer, a series of question is asked until we reach a conclusion about the class label of the record. A decision tree

shows these series of questions and their possible answers in an organized hierarchical format. The tree has three types of nodes:

- **Root node:** that has no incoming edges and zero or more outgoing edges.
- **Internal nodes:** each of which has exactly one incoming edge and two or more outgoing edges.
- **Leaf or terminal nodes:** each of which has exactly one incoming edge and no outgoing edges.

J48 algorithm is presented by Quinlan (Quinlan, J. R. 1993) uses greedy algorithm to generate the decision trees in a top-down recursive manner. The algorithm for inducing decision tree is presented in figure 7. The main strategy of the J 48 algorithm is as follows:

- J 48 tree starts with single node representing the training samples.
- If the samples belong to the same class, then the node becomes a leaf and is labeled with that class.
- An entropy-based procedure known as *information gain* is used by J48 algorithm for selecting the most suitable attribute that can classify the data precisely. This attribute becomes the "test" or "decision" attribute at the node.
- For each value of the test feature, a branch is generated.

The recursive partitioning stopping criteria are as follows :

- All the samples in a node belong to the same class

– There are no more features on which the samples may be further partitioned. In this situation, majority voting is used. This includes changing the given node into a leaf and labeling it with the class with has the highest majority among samples.

Algorithm: Develop a decision tree from given training data set (DT)
Samples is the training set
Set of attributes is all of the available attributes
Returns a tree node
DT(*samples, set of attribute*)
Begin
Generate a node *N*;
If all *samples* belong to the same class A **then**
Return *N* as a leaf node labeled with the class **A**;
Else if *set of attribute* is empty **then**
Return *N* as a leaf node labeled with the most common class in *samples.* (Majority voting)
Else Begin
Choose the attribute among *list of attribute* with the highest information gain (*test-attribute)*;
Name node N with *test-attribute*;
Let *si* be the set of samples in samples for which test-attribute = ai;
For each known range of values *ai* of *test-attribute;*
Begin
Generate an out-going branch **K** from node N with *test-attribute = ai;*
If *si* has an element (non-empty) **then**
Attach **K** to the node returned by DT (*si, set of attribute-(test-attribute))*
Else
Attach **K** to a leaf labeled with the most common class in *samples*;
End
Return Decision Node N
End
End

Figure 7. Basic algorithm for inducing a decision tree

3.5.1 Attribute Selection Measure

The information gain measure is a metric which is used to select the test attribute at each node in J48 tree. This metric is known as a feature selection measure or a measure of the goodness of split. The feature with the highest information gain is selected as the test feature for the current node.

Let *S* be a set consisting of *s* data samples and the class label attribute has *m* different values representing *m* different classes, C_i (for i = 1,…,m). Let s_i be the number of samples of S in class C_i. The expected information needed to classify a given sample can be presented as follows:

$$I(s_1, s_2, \dots, s_m) = -\sum_{i=1}^{m} P_i \log(P_i) \quad (1)$$

In (1) P_i is the probability that a sample belongs to class C_i and is estimated by s_i/s .

If attribute *A* have *v* different values, {a1,a2,…,a*v*} then it can divide *S* into *v* subset, {*S1,S2,...,Sv*}, where s_j contains the samples of S that have value aj of A. If *A* is chosen as the test attribute (the best attribute for splitting), then these subsets will correspond to the branches generated from the node containing the set S. Let S_{ij} be the number of samples of class C_i in a subset s_j. The entropy, (expected information based on the partitioning into subsets by A), is calculated by following formula:

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j}+\dots+s_{mj}}{s} I(s_{1j}, s_{2j}, \dots, s_{mj}) \quad (2)$$

The fraction $\frac{s_{1j}+\dots+s_{mj}}{s}$ can also be interpreted as the weight of the j*th* subset. It is actually the number of samples in the subset (having value aj of *A*) divided by total

number of samples in S. The small entropy value shows more pure subset. $I(s_{1j}, s_{2j}, \ldots, s_{mj})$ is calculated as follows for any subset S_j.

$$I(s_{1j}, s_{2j}, \ldots, s_{mj}) = - \sum_{i=1}^{m} P_{ij} \log(P_{ij}) \quad (3)$$

In equation (3), $P_{ij} = \frac{S_{ij}}{S_j}$ and it is the probability that a sample in S_j belongs to C_i. The information that can be gained by branching on *A* is as follows:

$$\textit{Info-Gain (A)} = I(s_1, s_2, \ldots, s_m) - E(A) \quad (4)$$

The algorithm selected the attribute with highest Info-gain as the test attribute for the given set S. A node is generated and labeled with the attribute; branches are generated for each value of the attribute. The samples are partitioned accordingly.

3.5.2 J48 Parameter Setting

Pruning a decision tree is a main step in optimizing the computational efficiency as well as classification accuracy of developed model. Some of the advantages of applying pruning methods to a decision tree are: reduction in the size of the tree (or the number of nodes), reducing unnecessary complexity, avoiding over-fitting of the data set when classifying new data. There are some factors that should be tuned when developing J48 algorithm using WEKA. These factors are as follows:

BinarySplits: False. This will let the tree to split nominal attributes.

ConfidenceFactor: The confidence factor is used for pruning process. Decreasing the confidence factor decreases the amount of pruning.

Unpruned: False. This will let the decision tree to perform pruning process while building the tree.

The *ConfidenceFactor* is the most important factors associated with J48 algorithm using WEKA data mining tool. In or der to develop the most efficient algorithm, the optimal values of these factors need to be determined. Thus we tried to run the experiment with different values of these parameters in order to find the best values of the parameter. The J48 classifier was tested with confidence factor ranging from auxiliary values near zero to 1.0. As presented in Figure 8, performance of the classifier on the testing set increased as the confidence factor increased. The highest value for precision reached at confidence factor of 0.5. After that the precision is constant.

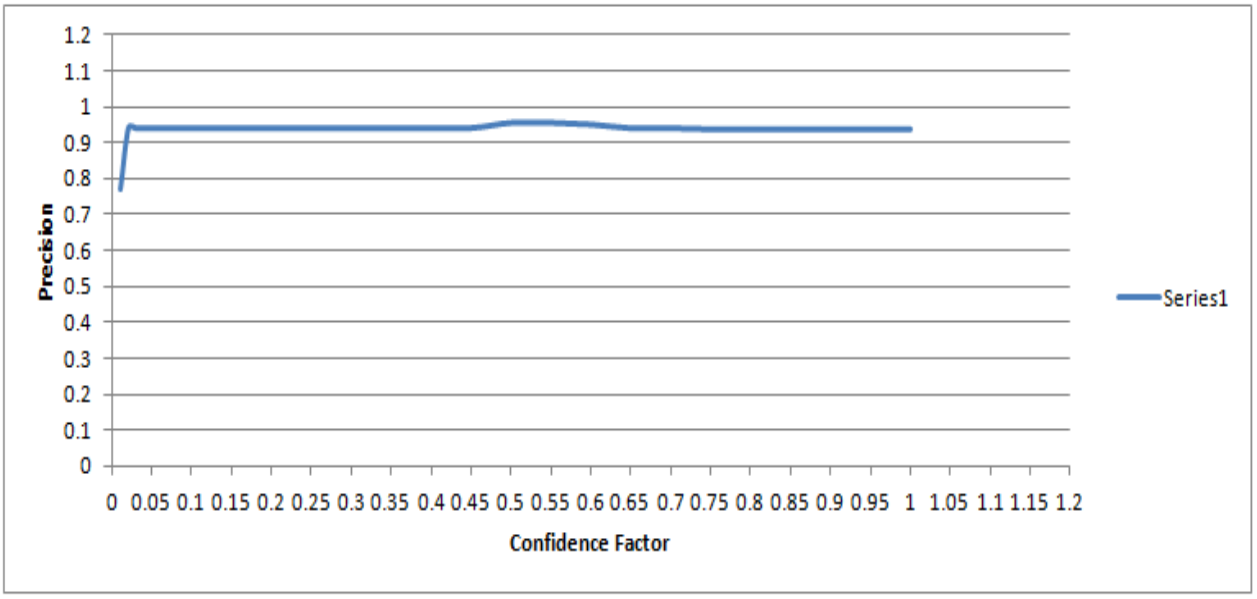

Figure 8. Optimal Value of Confidence Factor

The parameter of the J 48 algorithm are set as presented in the following:

BinarySplits: False

ConfidenceFactor: 0.5

Unpruned: False

As it is presented in figure 9, the J48 algorithm is able to classify future congestion up to 6 minutes ahead of time with very high and considerable quality. And up to 10 minutes with good performance.

TP	horizon	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
historical																
1		0.67	0.64	0.57	0.48	0.42	0.37	0.32	0.24	0.21	0.15					
2		0.78	0.73	0.7	0.63	0.58	0.55	0.45	0.41	0.36	0.27					
3		0.92	0.88	0.86	0.82	0.81	0.78	0.74	0.72	0.69	0.61	0.57	0.55	0.51	0.48	0.41
4		0.96	0.96	0.96	0.95	0.91	0.87	0.85	0.82	0.82	0.78	0.67	0.64	0.62	0.6	0.52
5		0.93	0.91	0.89	0.86	0.84	0.81	0.74	0.75	0.71	0.64	0.61	0.59	0.54	0.47	0.4
6		0.86	0.85	0.81	0.77	0.74	0.71	0.68	0.62	0.57	0.51	0.48	0.43	0.41		
7		0.83	0.8	0.73	0.71	0.68	0.62	0.59	0.55	0.51	0.42	0.4				
8		0.75	0.71	0.67	0.61	0.59	0.56	0.48	0.45	0.42						
9		0.68	0.65	0.61	0.59	0.54	0.51	0.48	0.45	0.41						

Precision	horizon	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
historical																
1		0.92	0.89	0.86	0.83	0.78	0.75	0.71	0.68	0.61	0.56					
2		0.93	0.92	0.89	0.86	0.84	0.82	0.78	0.75	0.7	0.66					
3		0.95	0.93	0.91	0.88	0.86	0.85	0.83	0.73	0.7	0.67	0.64	0.59	0.56	0.51	0.48
4		0.96	0.96	0.96	0.96	0.95	0.95	0.95	0.93	0.9	0.85	0.83	0.82	0.71	0.68	0.63
5		0.96	0.95	0.94	0.93	0.92	0.9	0.9	0.87	0.87	0.8	0.77	0.72	0.72	0.68	0.63
6		0.95	0.93	0.91	0.9	0.88	0.88	0.85	0.83	0.83	0.78	0.71	0.69	0.64		
7		0.93	0.88	0.85	0.82	0.78	0.75	0.71	0.64	0.56	0.51	0.49				
8		0.87	0.84	0.81	0.78	0.73	0.65	0.58	0.52	0.5						
9		0.81	0.78	0.74	0.71	0.65	0.58	0.52	0.49	0.46						

Figure 9. J48 result

Figure 10 represents the decision tree developed after running the J48 algorithm with the above mentioned parameters.

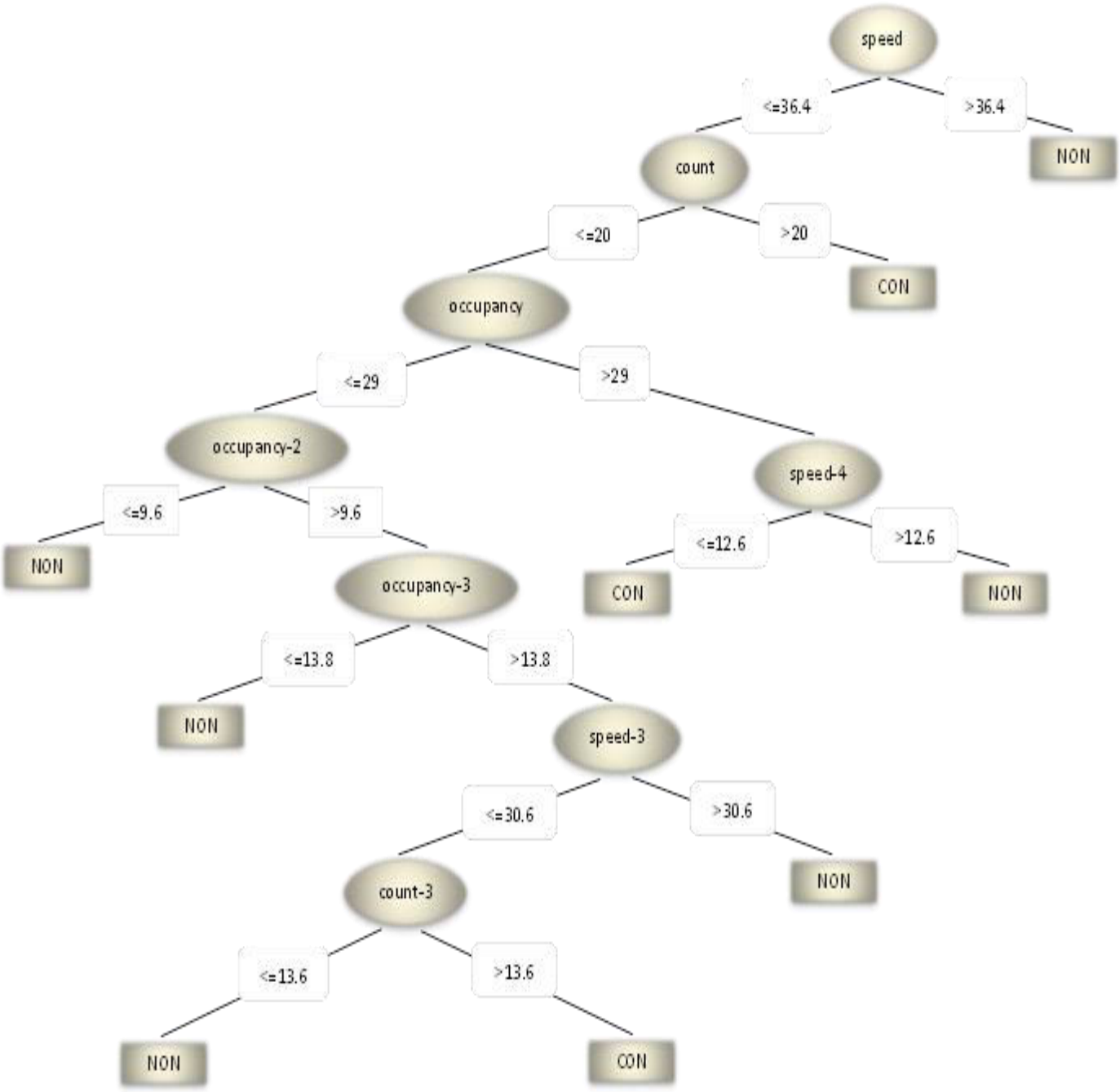

Figure 10. J48 Decision Tree

As it is presented in figure 10 the size of the decision tree and the number of the leaves are 19 and 9 respectively.

3.6 Artificial Neural Network (ANN)

ANN is one kind of predictive data mining technique which is widely used. It is a graph composed of nodes, which are sometimes referred to as units or neurons, and connections between the nodes (Zeidenberg, 1990). . ANN is a simulation model of the human brain, and it imitates the way that human brains make decision. It tries to learn the knowledge that exists in the data and store the learned knowledge within neuron connection weights (Giudici, 2003). ANN structure consists of following three main layers: input, hidden and output layers. There are some nodes (neurons) in each layer. These nodes are connected together with weighted links. In ANN network, the input nodes represent the input variables, the hidden and the output nodes play more active role in computations (Stalinski and Tuluca, 2006).

3.6.1 ANN Learning Algorithm

Learning process is done in an ANN network through adjusting weights. The ANN network is trained in order to extract the hidden rule that exist between input variables and output variables. This learning process can be used in classification problem. As presented before, there are supervised and unsupervised learning algorithms for any data mining techniques. ANN network takes advantage of supervised learning most of time to extract the hidden rules (Hill and Lewicki, 2007). An error back-propagation (Rumelhart et. al. 1986) is a supervised learning method which is used in ANN. This method lets the ANN to compare the responses of the output values to the desired values and to readjust the weights in the ANN to find the best values of weight. If

the values of the weights are set correctly, the response of the ANN will be closer to desired values when the same input is inserted to the ANN structure. Error back-propagation is the most useful learning method for ANN (Zeidenberg, 1990). ANN algorithm compares its generated output to the actual output from the training data. Then the error in each output neuron is estimated. For each neuron, the correct output is calculated. ANN specifies how much lower or higher the output must be adjusted to match the actual output stored in tested cases. The difference between the generated output and the actual output is presented as local error. The ANN continuously adjusts the weights of each neuron to minimize the local error. The back propagation does this process. It calculates the gradient of the error of ANN considering its modifiable weights. It is actually an iterative gradient algorithm developed to minimize the error between the generated output and the actual output of an ANN (Goh, 2000). In ANN, Back propagation method is used to determine the weights and thresholds between the input and hidden layers and those between the hidden and output layers (Hsiao and Huang, 2002). The sigmoid transfer function is used to modify the output of each neuron. The output of each hidden and output neuron are presented by the sigmoid functions (6) , (7) respectively.

$$\mathrm{F}(x_j) = \frac{1}{1+e^{-(\sum_{i=1}^{n} X_i \cdot W_{ij} - b_{ij})}} \tag{6}$$

$$\mathrm{F}(x_k) = \frac{1}{1+e^{-(\sum_{i=1}^{n} X_j \cdot W_{jk} - b_{jk})}} \tag{7}$$

In the abovementioned formulas X_i is the value of the input variable, W_{ij} and W_{jk} are connection weights between the input and the hidden neuron and between the hidden neuron and the output neuron, respectively, b_{ij} and b_{jk} are thresholds terms for the ith

and kth neuron, respectively; i, j, and k are the number of neurons in each layer (Kim et al., 2004). Figure 11 shows a Artificial Neural Network.

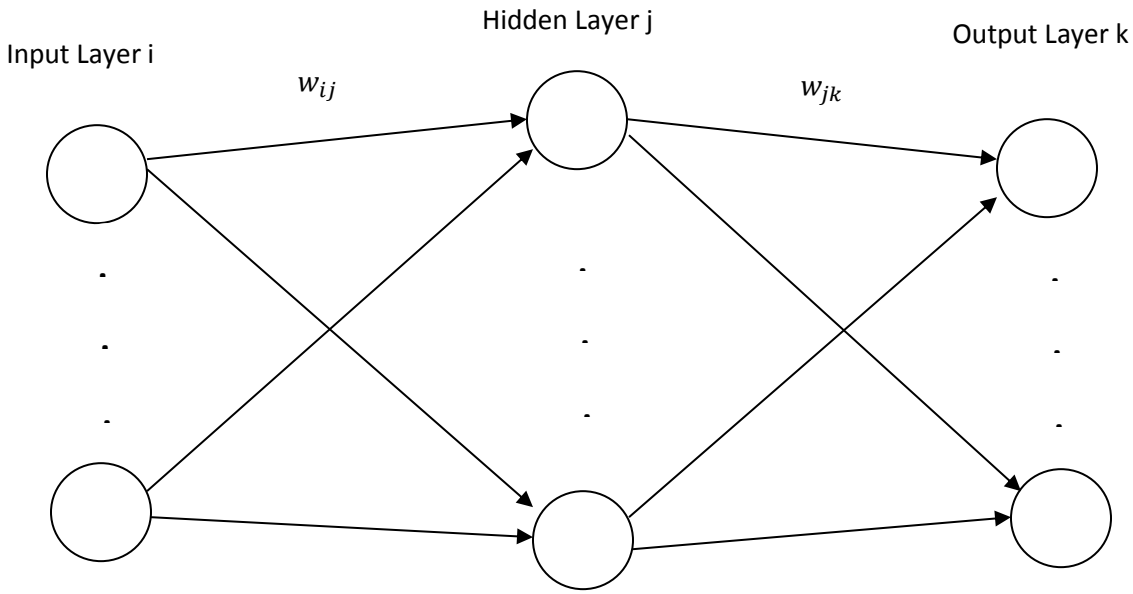

Figure 11. Artificial Neural Network

3.6.2 ANN Parameter Setting

Artificial Neural Network has some parameter that need to be set in order to have the best performance of the ANN network. These parameters are set usually by trial and error procedure. These parameters are as follows:

Parameter Setting:

Hidden layer: number of nodes in hidden layer.

Learning Rate : it is a user-designated parameter that specifies how much the link weights can be changed. The learning rate actually changes the speed at which the ANN arrives at the minimum solution. If it is too high the system might diverge completely and if it is too low it may takes time to converge on the final solution.

applies a greater or lesser portion of the respective adjustment to the old weight.

Momentum : Momentum simply adds a fraction *m* of the previous weight update. It is used to prevent the system from converging to a local minimum.

Training time: It is the number of times the training vectors are used to update the weights.

The ANN classifier was tested with different values for above-mentioned parameters the figures 12 through 15 show the optimal values of the parameters respectively.

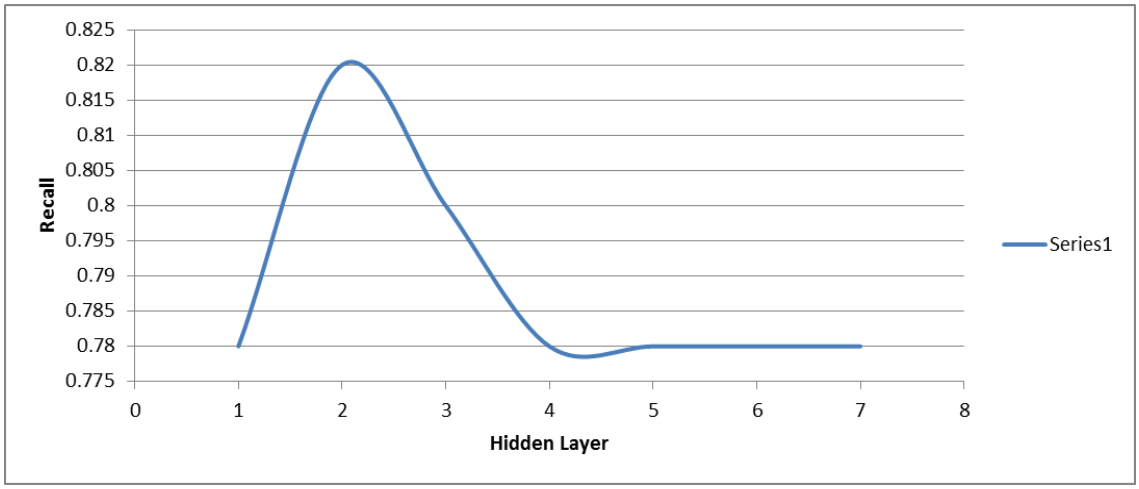

Figure 12. Optimal number of nodes sin hidden layer

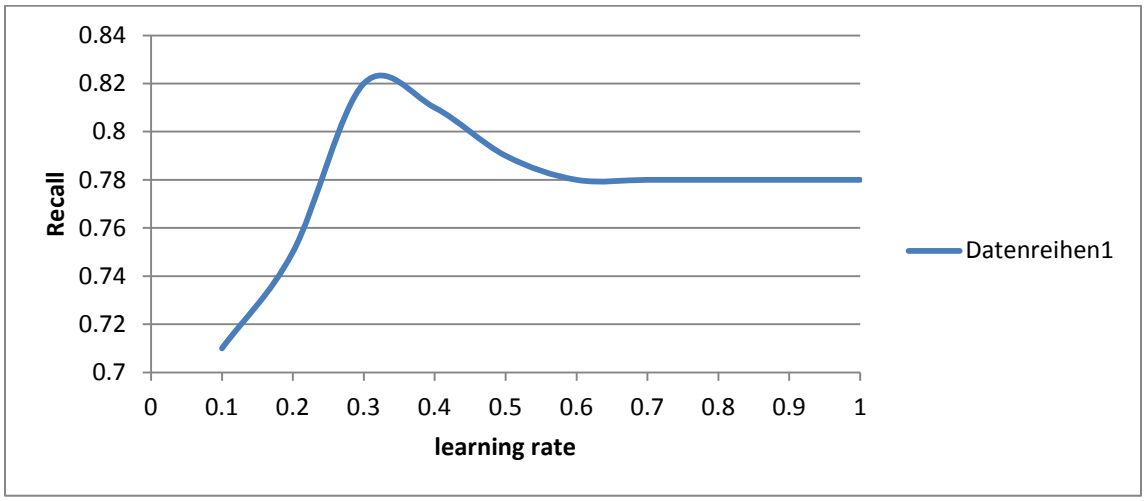

Figure 13. Optimal training rate

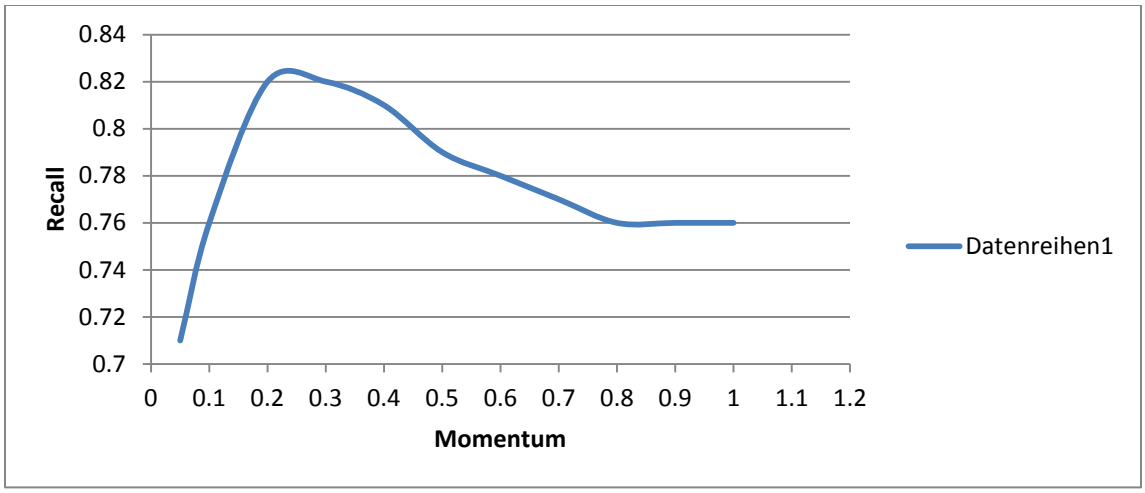

Figure 14. Optimal momentum

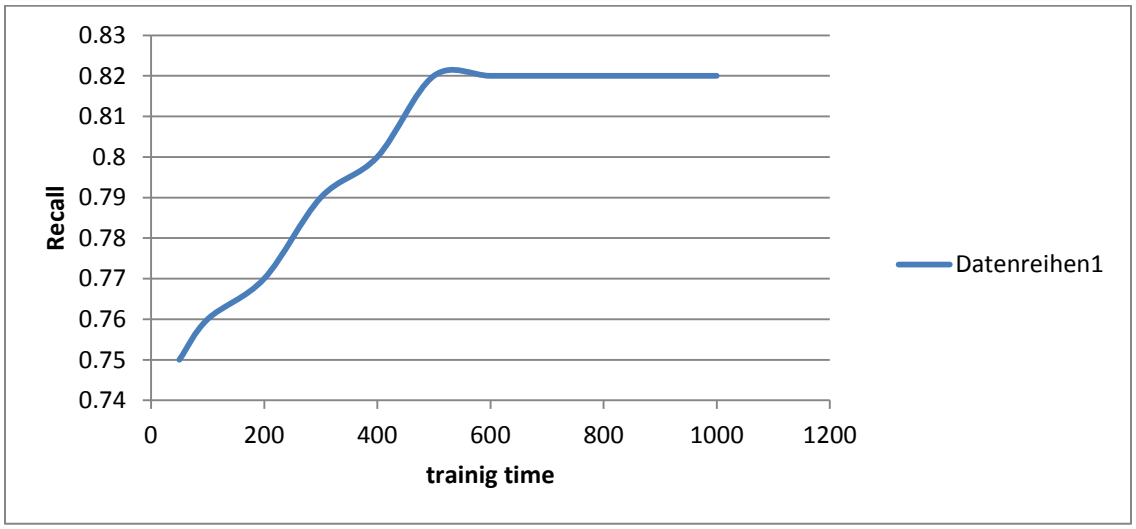

Figure 15. Optimal training time

The experimental result of Artificial Neural Network is presented in figure 16.

TP	horizon	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
historical																
1		0.65	0.63	0.53	0.43	0.42	0.35	0.28	0.22							
2		0.76	0.7	0.64	0.6	0.53	0.48	0.41	0.34							
3		0.91	0.82	0.81	0.74	0.7	0.66	0.65	0.56	0.53	0.46	0.4	0.36			
4		0.95	0.95	0.89	0.85	0.82	0.82	0.78	0.65	0.62	0.6	0.47	0.47	0.47	0.39	0.34
5		0.92	0.9	0.84	0.82	0.78	0.75	0.69	0.6	0.57	0.54	0.42	0.4	0.38	0.36	0.31
6		0.85	0.82	0.79	0.75	0.71	0.68	0.61	0.56	0.51	0.47	0.39	0.31			
7		0.77	0.75	0.71	0.67	0.61	0.57	0.53	0.46	0.41	0.34					
8		0.7	0.69	0.64	0.6	0.53	0.51	0.43	0.37							
9		0.64	0.61	0.56	0.52	0.44	0.41	0.35								

Precision	horizon	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
historical																
1		0.9	0.86	0.82	0.78	0.73	0.68	0.62	0.58							
2		0.92	0.9	0.84	0.81	0.77	0.74	0.7	0.69							
3		0.95	0.88	0.88	0.85	0.81	0.74	0.71	0.7	0.7	0.67	0.61	0.56			
4		0.96	0.92	0.9	0.9	0.86	0.79	0.78	0.78	0.77	0.74	0.69	0.65	0.63	0.6	0.58
5		0.95	0.9	0.87	0.85	0.81	0.74	0.72	0.7	0.67	0.65	0.61	0.58	0.52	0.5	0.47
6		0.95	0.84	0.82	0.8	0.78	0.71	0.69	0.65	0.61	0.6	0.54	0.53			
7		0.91	0.82	0.81	0.79	0.74	0.7	0.65	0.61	0.55	0.5					
8		0.85	0.8	0.8	0.78	0.73	0.65	0.58	0.52							
9		0.79	0.77	0.76	0.72	0.68	0.61	0.54								

Figure 16. ANN result

As presented in figure 16 the ANN algorithm is able to classify future congestion up to 7 minute ahead of time with very good performance. The developed ANN network is presented in figure 17.

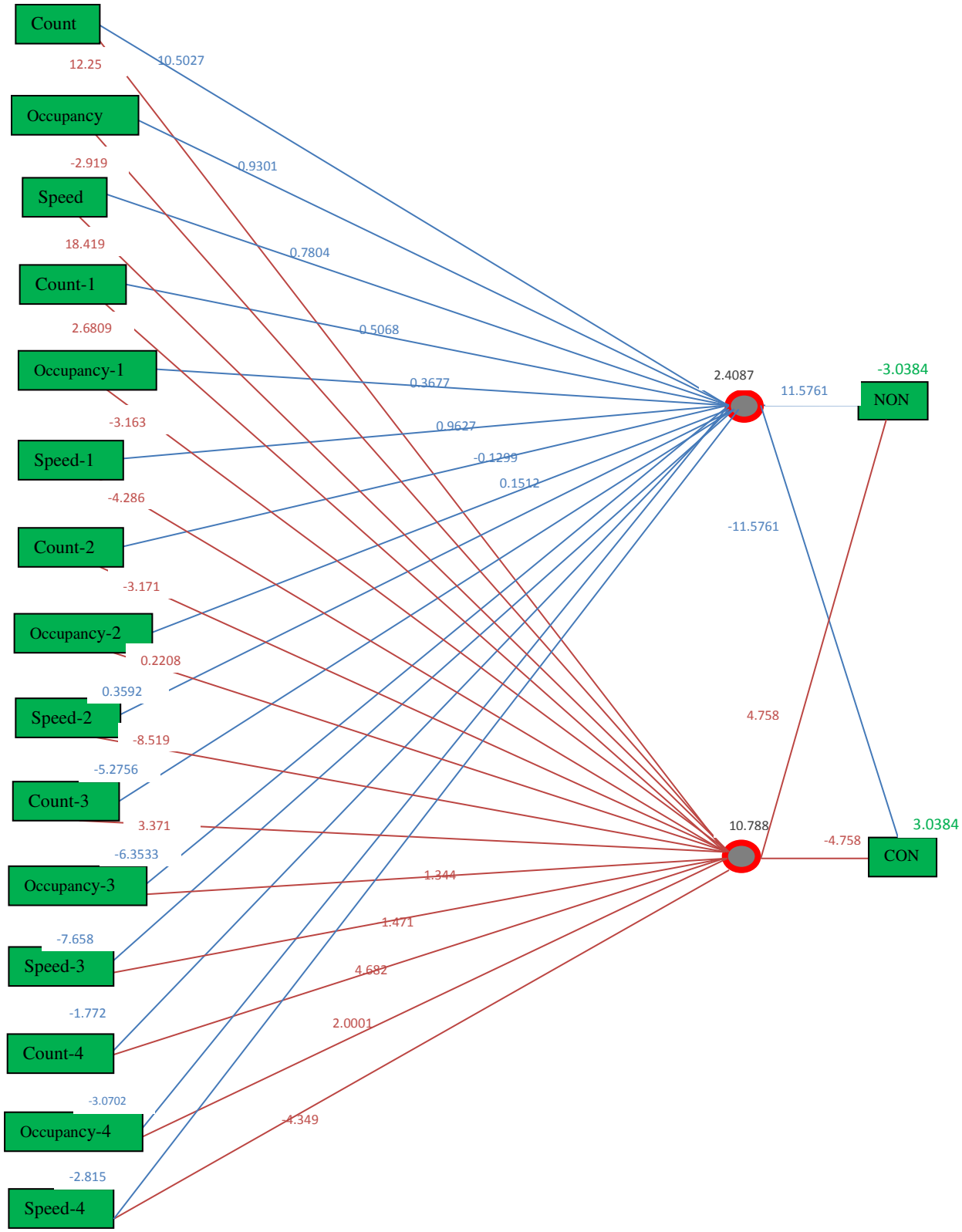

Figure17. ANN network

3.7 Support Vector Machine (SVM)

The Support Vector Machine (SVM) is a classification method introduced in 1992 by Boser et al, (Boser et al., 1999). A N-dimensional hyper-plane is generated by this algorithm to optimally classify the data into categories. A SVM finds a line (or, in general, hyperplane) that maximized the margin between the support vectors as presented in figure 18.

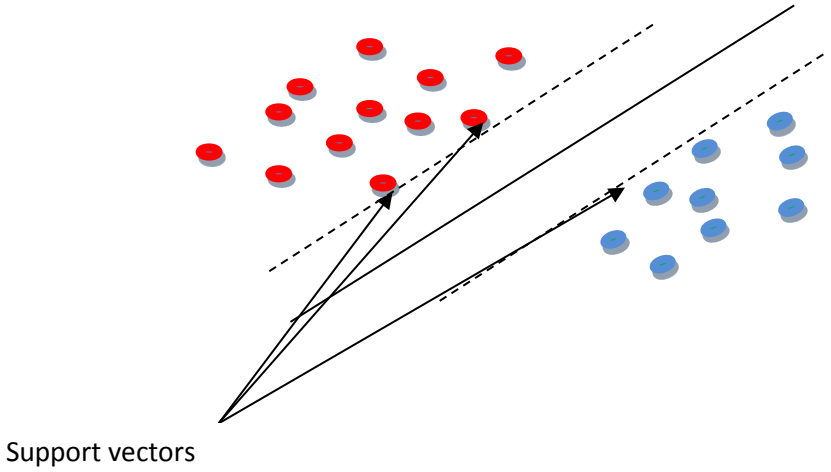

Figure 18. Support Vector Machine

SVM algorithms are associated with kernel methods (Scholkopf, Smola 2002 ; Shawe-Taylor, J ; Cristianini 2004). But in real cases it might be needed to classify complicated objects that are not linearly classifiable in their current space. So the SVM take advantage of Kernel methods to map the data to a space with higher dimension. The figure 18 shows that the data which are not linearly classifiable in 2 dimensional space can be linearly classified when are mapped to 3-dimentional space.

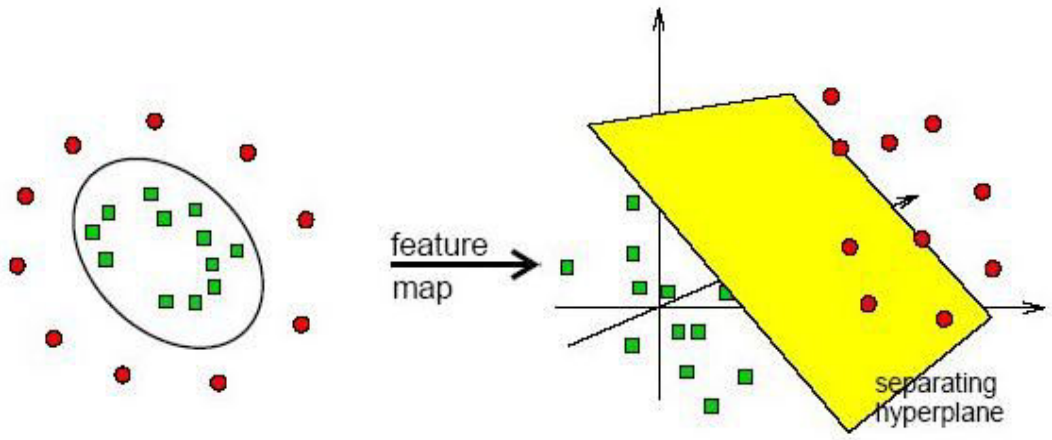

Figure 19. Mapping process

3.7.1 SVM Parameter Setting

There are different types of kernels that SVM take advantage of(Steinwart & Christmann, 2008). Some of these kernels are: Linear, Polynomial, Radial Basis Function (RBF) and Pearson VII Universal Kernel (PUK). In this study different kernels has been tested on this problem and Polynomial kernel has the best performance. The general form of the Polynomial kernel is as follows:

$$K(x,y) = (x^T * y + r)^n \quad (8)$$

In equation (8) x and y are vectors of features in training data and r is a constant number.

Table 8. SVM Kernel Selection

Kernel	Recall	Precision
Polynomial	0.95	0.95
PUK	0.91	0.81
RBF	0.75	0.78

When applying polynomial kernel, the optimal value of n should be specified. In this problem the optimal value of n is equal to 3 as presented in figure 20.

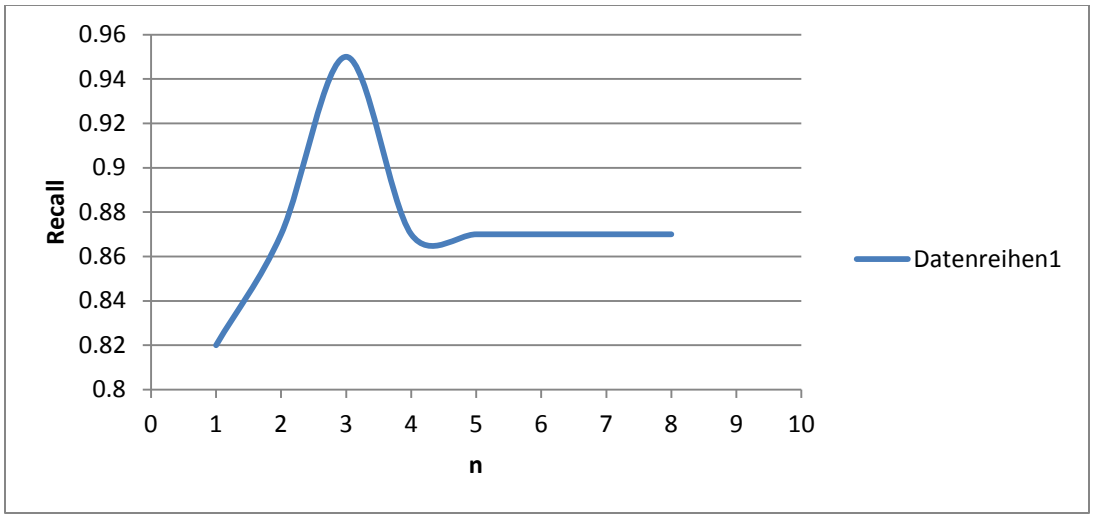

Figure 20. Optimal value of exponents(n)

The performance of the SVM on this problem is presented in figure 21.

TP	horizon	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
historical																
1		0.65	0.61	0.52	0.4	0.38	0.33	0.25	0.18							
2		0.75	0.7	0.61	0.58	0.51	0.45	0.41	0.35	0.27	0.23					
3		0.9	0.81	0.8	0.74	0.72	0.64	0.62	0.54	0.5	0.42	0.31	0.24	0.22	0.12	0.1
4		0.95	0.87	0.87	0.82	0.79	0.73	0.69	0.65	0.6	0.56	0.39	0.34	0.3	0.17	0.13
5		0.92	0.85	0.82	0.76	0.74	0.7	0.64	0.61	0.57	0.5	0.31	0.24			
6		0.84	0.83	0.74	0.7	0.66	0.63	0.52	0.47	0.4	0.37					
7		0.74	0.7	0.64	0.57	0.52	0.45	0.36	0.32							
8		0.71	0.65	0.61	0.54	0.48	0.41	0.33								
9		0.62	0.57	0.51	0.43	0.39	0.32									

Precision	horizon	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
historical																
1		0.87	0.87	0.8	0.75	0.71	0.67	0.6	0.54							
2		0.91	0.87	0.82	0.8	0.75	0.72	0.69	0.65	0.62	0.61					
3		0.94	0.86	0.84	0.83	0.81	0.73	0.7	0.67	0.63	0.6	0.57	0.52	0.43	0.4	0.36
4		0.95	0.9	0.87	0.85	0.81	0.79	0.77	0.75	0.75	0.73	0.71	0.71	0.7	0.6	0.54
5		0.92	0.86	0.84	0.82	0.79	0.75	0.72	0.71	0.7	0.69	0.64	0.63			
6		0.91	0.9	0.8	0.78	0.76	0.7	0.65	0.6	0.54	0.52					
7		0.87	0.8	0.76	0.74	0.71	0.7	0.63	0.58							
8		0.82	0.8	0.74	0.71	0.65	0.61	0.54								
9		0.76	0.71	0.69	0.64	0.61	0.58									

Figure 21. SVM result

The presented SVM is consisting of 65 support vectors. As presented in Figure20, The SVM is capable of classifying traffic status up to 6 minute ahead of time.

3.8 PART Algorithm

A PART algorithm is actually a combination of C4.5 Decision tree and RIPPER algorithm (Frank & Witten). The C4.5 tries to learn the rule based on decision tree and RIPPER tries to learn the rule based on separate-and-conquer algorithm. Both of C 4.5 and RIPPER algorithms perform global optimization procedures on the initially produced set of rules. Both C4.5 and RIPPER algorithms start with an initial model and then iteratively improve it using heuristic techniques. PART algorithm is a rule-induction process that avoids global optimization procedure that the two above-mentioned algorithms do, but nevertheless produces accurate, compact set of rules. The C4.5 algorithm presents a rule in decision tree format. It tries to generate one rule for each path from the root to the leaf. Based on (Pagallo and Haussler 1990), it is possible to simplify the rules generated with this procedure without losing their predictive performance. Moreover, the optimization process also takes a lot of time. The Part algorithm combines the C4.5 RIPPER algorithm to take advantage of the positive advantages of both of the algorithms while disregarding the negative pints of them. The simplicity of PART is the main advantage of it. Combining separate-and-conquer methodology with decision tree adds flexibility and speed to PART algorithm. The PART algorithm differs from standard approach in the way that each rule is created. To make a rule, a pruned decision tree is generated for the current set of instances and then the leaf with the largest coverage is set to the rule and the tree is discarded. The main idea of PART algorithm is to build partial trees instead of fully explored one. In order to generate a sub tree, The PART algorithm

tries to find the sub tree that cannot be simplified further. When the sub tree is found the tree building algorithm starts and a rule is generated. The tree building algorithm is presented in figure 22.

Procedure expand subsets

While there are subsets that have not been expanded **and**

All the subsets that are expanded so far are leaves

Choose next subset to be expanded and expand it

If all the subset that are expanded are leaves **and**

Estimated error for sub tree >= estimated error for node

Undo expansion into subsets and make node a leaf.

Figure 22. The tree building algorithm

3.8.1 PART Parameter setting

There are some factors that should be tuned when developing J48 algorithm using WEKA. These factors are as follows:

ConfidenceFactor: The confidence factor is used for pruning process. Decreasing the confidence factor decreases the amount of pruning.

MinNumObj: The number of minimum instances per node. In most case it is equal to 2 (if a split yields a child leaf with less than a minimum number of instances from the data set, the parent node and its children are combined into a single node)

Figure 23 and 24 shows the optimal values for these two parameters respectively.

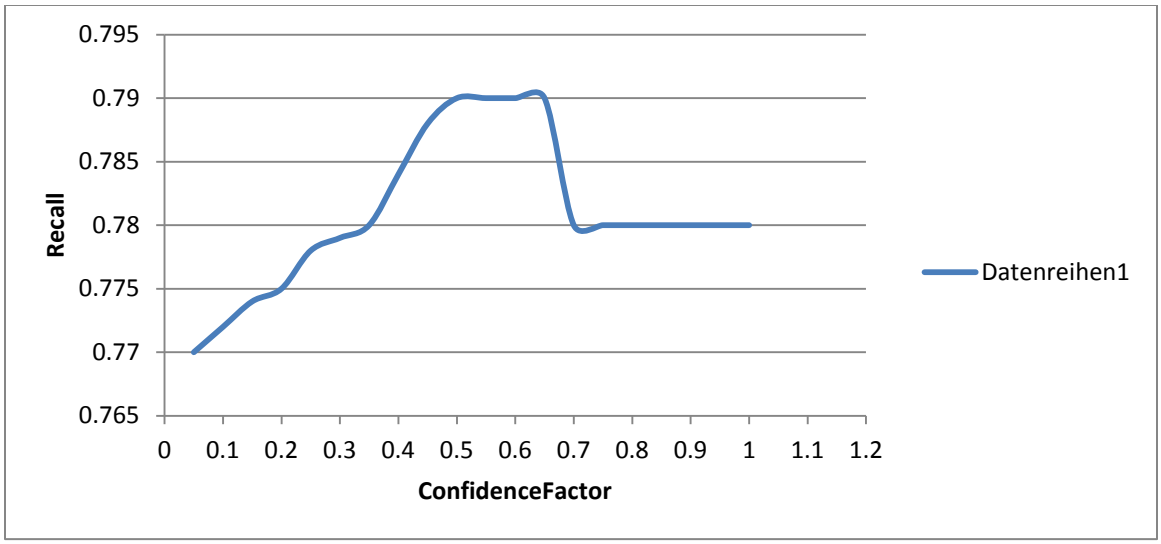

Figure 23. PART's Confidence Factor

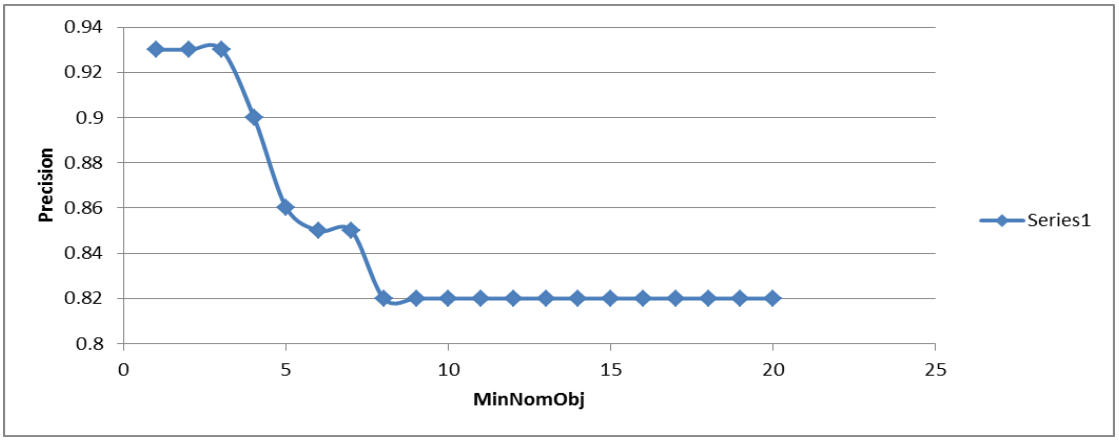

Figure 24. Min Number of objects

The experimental results for PART algorithm to classify congestion are presented in figure 25.

TP	horizon	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
historical																
1		0.64	0.6	0.51	0.4	0.34	0.3									
2		0.72	0.64	0.6	0.55	0.48	0.43	0.4	0.31	0.24						
3		0.89	0.8	0.76	0.7	0.66	0.61	0.58	0.51	0.44	0.41					
4		0.95	0.86	0.84	0.82	0.78	0.72	0.72	0.63	0.58	0.51	0.47	0.3	0.23		
5		0.92	0.84	0.78	0.74	0.72	0.68	0.62	0.57	0.51	0.48	0.42	0.28			
6		0.81	0.71	0.68	0.61	0.55	0.51	0.45	0.39							
7		0.73	0.68	0.62	0.54	0.51	0.42	0.31	0.22							
8		0.7	0.63	0.57	0.5	0.47	0.39	0.24								
9		0.61	0.56	0.52	0.45	0.35	0.26									

Precision	horizon	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
historical																
1		0.86	0.84	0.8	0.71	0.68	0.65									
2		0.9	0.86	0.81	0.78	0.74	0.7	0.67	0.61	0.56						
3		0.92	0.86	0.82	0.82	0.78	0.72	0.68	0.65	0.58	0.54					
4		0.95	0.9	0.89	0.81	0.8	0.79	0.77	0.76	0.75	0.7	0.63	0.51	0.4		
5		0.91	0.84	0.81	0.8	0.77	0.73	0.7	0.67	0.65	0.62	0.6	0.57			
6		0.9	0.85	0.83	0.8	0.76	0.75	0.73	0.7							
7		0.85	0.82	0.8	0.75	0.7	0.67	0.61	0.58							
8		0.81	0.75	0.72	0.7	0.67	0.62	0.54								
9		0.74	0.68	0.65	0.59	0.55	0.52									

Figure 25. PART result

As presented above the PART algorithm is capable of classifying congestion up to 7 minute with good performance. The set of rules that have been developed by PART algorithm are presented in figure 26.

RULE 1 :
occupancy <= 20.4 AND
speed > 60.4: NON

RULE 2 :
speed > 32.2 AND
occupancy <= 19.2 AND
speed > 47.8: NON (3299.0/14.0)

RULE 3:
occupancy-3 > 7.8 AND
occupancy > 29 AND
occupancy-3 > 25.2 AND
speed-1 < 35.4 AND
speed-4 < 21.4: CON (46.0/9.0)

RULE 4:
count-4 > 12.8 AND
occupancy > 29 AND
occupancy-3 >= 26.6: CON (22.0)

RULE 5:
occupancy-1 > 11.2 AND
occupancy-4 >= 33.8 AND
count-1 >= 20.8 AND
speed-4 < 41.4: CON (13.0)

RULE 6:
occupancy-4 >= 33.8 AND
occupancy-1 > 11.2 AND
count-1 >= 24.8 AND
occupancy >= 25.6: CON (15.0)

RULE 7:
occupancy > 26 AND
count >= 18.8 AND
speed-3 > 13.2 AND
speed-3 <= 24.6: CON (9.0)

Figure 26. PART rules

3.9 K-Nearest Neighborhood Algorithm (K-NN)

The KNN algorithm performs classification process by comparing the attributes of the test object with K object in the training set that are closest to the test object and chooses a label for the testing object based on the predominance of a particular class in this neighborhood. To classify an unlabeled new object, the distance of this testing object to the labeled objects is computed, its k-nearest neighbors are identified, and then the class of the testing item is set based on the majority class of its nearest neighbors (Larose, 2005). Figure 6 presents the nearest-neighbor classification method. Given a training set TR and a test object O = ($\acute{x}$,$\acute{y}$), the K-NN algorithm computes the distance (or similarity) between O and all the training objects (x, y) ∈ TR to determine its K nearest-neighbors. (y is the label of the training data (x). and, $\acute{y}$ is the label of the test data ($\acute{x}$)) Once the K nearest neighbors are specified, the test object is classified based on the majority class of its nearest neighbors.

Majority voting: $\acute{y} = \text{argmax}_v \sum_{(x_i,y_i)\in D_z} I\,(v = y_i)$ (9)

In equation (9) v is a class label, y_i is the class label for the ith nearest neighbors, and I (·) is an indicator function that returns the value 1 if its argument is true and 0 otherwise.

3.9.1 K-NN Parameter Setting

The number of nearest neighbors is parameter that needs to be set for K-NN algorithm. The linear search is used to find the nearest neighbors. The distance is taken in to account by 1-distance weighting method. The optimal number of neighbors is presented in figure 27.

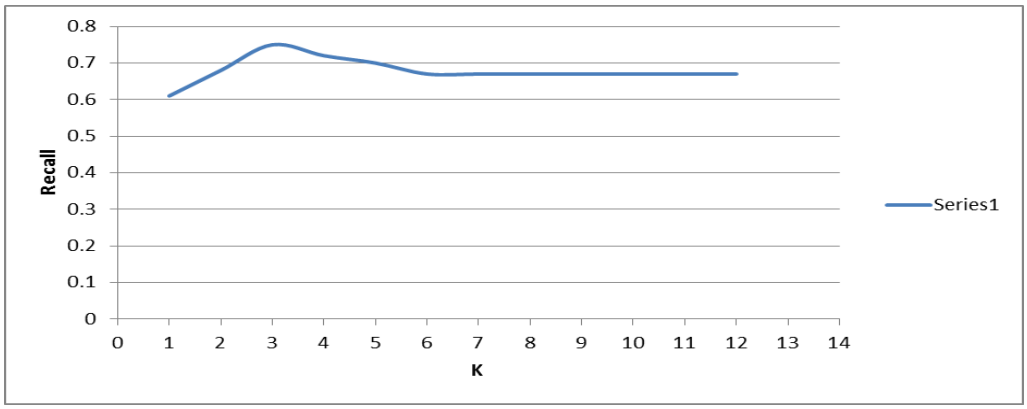

Figure 27. Optimal number of neighbors

The performance of K-NN method for traffic classification is presented in figure 28.

TP	horizon	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
historical																
1		0.62	0.58	0.55	0.42	0.32	0.24	0.2								
2		0.71	0.63	0.6	0.51	0.47	0.41	0.37	0.34							
3		0.84	0.79	0.73	0.67	0.64	0.61	0.54	0.52	0.46	0.41	0.37				
4		0.95	0.85	0.82	0.8	0.75	0.73	0.71	0.61	0.55	0.5	0.41	0.26	0.16		
5		0.91	0.83	0.76	0.72	0.71	0.66	0.6	0.55	0.5	0.42	0.37	0.21			
6		0.8	0.7	0.64	0.56	0.5	0.43	0.4	0.34							
7		0.71	0.64	0.6	0.53	0.4	0.44	0.39								
8		0.69	0.62	0.55	0.48	0.4	0.31	0.22								
9		0.6	0.54	0.5	0.42	0.31	0.31	0.2								

Precision	horizon	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
historical																
1		0.84	0.82	0.81	0.74	0.66	0.65	0.59								
2		0.9	0.85	0.84	0.75	0.71	0.67	0.61	0.57							
3		0.92	0.85	0.81	0.81	0.74	0.7	0.66	0.62	0.55	0.51	0.42				
4		0.95	0.9	0.86	0.8	0.77	0.76	0.72	0.71	0.63	0.59	0.56	0.56	0.51		
5		0.9	0.82	0.78	0.76	0.71	0.66	0.62	0.58	0.55	0.51	0.48	0.43			
6		0.9	0.84	0.81	0.76	0.72	0.65	0.6	0.53							
7		0.84	0.82	0.76	0.71	0.65	0.61	0.54								
8		0.8	0.76	0.71	0.67	0.63	0.58	0.54								
9		0.72	0.66	0.62	0.53	0.5	0.47	0.42								

Figure 28. K-NN reslt.

As presented in figure 28 the K-NN algorithm classify traffic status up to 7 minute ahead of time with good performance.

3.10 Comparative Result

The comparison of the L48, ANN, SVM , PART and K-NN is presented in figure 29. As presented in figure 29, J48 algorithm has better performance compared with other algorithms.

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Recall	J48	0.96	0.96	0.96	0.95	0.91	0.87	0.85	0.82	0.82	0.78	0.67	0.64	0.62	0.6	0.52
	NN	0.95	0.95	0.89	0.85	0.82	0.82	0.78	0.65	0.62	0.6	0.47	0.47	0.47	0.39	0.34
	SVM	0.95	0.87	0.87	0.82	0.79	0.73	0.69	0.65	0.6	0.56	0.39	0.34	0.3	0.17	0.13
	PART	0.95	0.86	0.84	0.82	0.78	0.72	0.72	0.63	0.58	0.51	0.47	0.3	0.23		
	KNN	0.95	0.85	0.82	0.8	0.75	0.73	0.71	0.61	0.55	0.5	0.41	0.26	0.16		
Precision	J48	0.96	0.96	0.96	0.96	0.95	0.95	0.95	0.93	0.9	0.85	0.83	0.82	0.71	0.68	0.63
	NN	0.96	0.92	0.9	0.9	0.86	0.79	0.78	0.78	0.77	0.74	0.69	0.65	0.63	0.6	0.58
	SVM	0.95	0.9	0.87	0.85	0.81	0.79	0.77	0.75	0.75	0.73	0.71	0.71	0.7	0.6	0.54
	PART	0.95	0.9	0.89	0.81	0.8	0.79	0.77	0.76	0.75	0.7	0.63	0.51	0.4		
	KNN	0.95	0.9	0.86	0.8	0.77	0.76	0.72	0.71	0.63	0.59	0.56	0.56	0.51		

Figure 29. Comparative result

The J48 is able to classify future traffic status up to 10 minute ahead of time with good performance while the performance of other classifiers presented here will decrease dramatically after 6 or 7 minute.

CHAPTER 4

CONCLUSION AND FUTURE RESEARCH

4.1 Conclusion

This study presents a model for classifying the next state of traffic congestion using data mining techniques. Data mining techniques usually lead to good results when dealing with the abundant amounts of data. Intelligent Transportation Systems (ITS) technology collects large amount of historical traffic flow data that will provide researcher with information for improvement of traffic control and predicting the next state of traffic congestion. The comparative study using J48 Decision Tree, Artificial Neural Network, Support Vector Machine, PART, K-Nearest Neighborhood algorithms is done and the result shows that the J48 algorithm has a better performance compared with other algorithms. Given the historical speed, occupancy and vehicle counts data the classification algorithm is able to classify the future status of traffic to congested or non-congested. The proposed J48 algorithm provides a very promising RECALL and PRECISION when applied to data from the northbound Interstate I-15 Northbound from I-215 up to Desert Inn, Las Vegas, NV. The historical record versus time horizon analysis conducted to shows that how much historical data we need to classify the future congestion status as far as possible. The Developed algorithm is able to classify the future congestion status up to 6 minutes ahead of time with very good performance.

4.2 Future Research

There are a lot of research gaps in classification and prediction of the congestion status. The research that has been done before were able to predict the real-time status of

traffic congestion. This research presented in this study is able to predict the future state of traffic congestion. But there are still many issues that can be considered in congestion prediction. Some of the research options are as follows:

- Using ensemble classifier.
- Developing fuzzy classifier or fuzzy models.
- Extending the model to arterials and street.

REFRENCES

Aday, L. A., & Andersen,R. (1974). A framework for the study of access to medical care. *HealthServices Research, 9*, 208-220.

Aftabuzzaman. Md. (2007). *Measuring Traffic Congestion- A Critical Review*. Institute of Transport Studies. Retrieved, Retrieved from http://www.atrf.info//papers/2007/2007_Aftabuzzaman.pdf.

Boser, B.E., Guyon, L.M. & Vapnik. V.N. (1992). A training algorithm for optimal margin classi_ers. In D. Haussler, editor, 5th Annual ACM Workshop on COLT, pages 144-152, Pittsburgh, PA,. ACM Press.

Bertini,R.L., Leal, M., & Lovell, D.J. (2002). Generating Performance Measures from Portland's Archived Advanced Traffic Management System Data. *Transportation Research Board Annual meeting*.

Botha, G. (2005). Measuring road traffic safety performance. *24th Annual Southern African Transport Conference.*

Dumbaugh, E., & Meyer, M.D. (2003). Exploring the Relationship between Agency Performance Measures and Operations Investments in a Metropolitan Area. *Transportation Research Board*.

Dunham, M. (2003). *Data mining: introductory and advanced topics*. Upper Saddle River, NJ: Prentice Hall.

Eisele, W. L., Schrank, D.L., & Lomax, T.J. (2005). *Incorporating Access Management into the Texas Transportation Institute.Urban Mobility Report 84th Meeting.*

Elhenawy, M., & Rakha, A.H. (2014). Congestion prediction using adaptive boosting machine learning classifiers. Transportation *Research Board 93rd Annual Meeting*.

Frank, E., Witten, H.I. (1998). Generating Accurate Rule Sets Without Global Optimization. *In: Fifteenth International Conference on Machine Learning*, 144-151.

Farrington, J., & Farrington, C. (2005). Rural accessibility, social inclusion and social justice: Towards conceptualization. *Journal of Transport Geography,* 13(1), 1-12.

Frieden, L. (2005). *The Current State of Transportation for People with Disabilities in the United States*. National Council on Disability website.

Global Environmental Management Initiative. (1998). *Measuring environmental Performance. A Primer and Survey of Metrics in use.*Retrieved from http://www.gemi.org/resources/met_101.pdf

Gudmundsson, H., (2000). *Indicators for performance measures for transportation, environment and sustainability in North America.* Ministry of Environment and Energy, National Environmental Research Institute, Denmark.

Geurs, K.T., & Ritsema Van Eck, J.R. (2001). *Accessibility measures: review and application.Evaluation of accessibility impact of land-use transport scenarios and related social and economic impact*. Retrieved from http://hdl.handle.net/10029/9487

Gulliford, M., Figueroa-Munoz, J., Morgan, M., Hughes, D., Gibson, R., & Beech, R. (2002). What does access to health care mean? *Journal of Health Services Research and Policy,*7(3), 186-188.

Giudici, P. (2003), Applied Data Mining: Statistical Methods for Business and Industry, Wiley, New York, NY.

Hane, J., Kamber, M., & Pei, j. (2006). *Data Mining Concepts and Techniques*. Third

Edition. Published by Morgan Kaufmann.

Hill, T. & Lewicki, P. (2007), STATISTICS Methods and Applications. StatSoft, Tulsa, OK.

Hongsakham, W., Pattara-atikom, W., & Peachavanich, R. (2007). Estimating road traffic congestion from cellular handoff information using cell-based neural networks and K-means clustering. *ECTI-CON 2008. 5th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information.*

Hedlund, J. (2008). Traffic Safety Performance Measures for States and Federal Agencies. *U.S. Department of Transportation/NHTSA Office of Behavioral Safety Research.*

Herbel, S., Meyer, M.D., Kleiner, B., & Gaines,. D. (2011). *A Primer on Safety Performance Measures for the Transportation Planning Process*. Economic Development Research Group.

Koenig, M. (1978). Accessibility and Individual Behavior: Accessibility Indicators as a Determinant of Trip Rate and Urban Development. *Paper presented at the PTRC summer meeting.*

Lomax, T., Schrank, D., Turner, S., & Margiotta, R. (2003). *Selecting travel time reliability measures.* Texas Transportation Institute. Cambridge systematics.

Litman, T. (2003). Measuring Transportation: Traffic, Mobility and Accessibility. *Institute of Transportation Engineers,* 73(10),28-32.

Lu, J., & Cao, L. (2003). *Congestion evaluation from traffic flow information based on*

fuzzy logic. IEEE Intell. Transport. Syst. 1,pp50–33.

Larose, D. T. (2005). Discovering knowledge in data: An introduction to data mining. Hoboken, NJ: Wiley.

Morris, J.M., Dumble, P.L., & Wigan, M.R. (1978). Accessibility Indicators for Transport Planning. *Transportation Research A ,* 13, 91-109.

Medley, S. B., & Demetsky, M. J. (2003). Development of congestion performance measures using ITS information, *Virginia Transportation Research Council*, VTRC 03-R1.

Murray, A.T., & Wu, X. (2003). Accessibility trade off in public transit planning. *Journal of Geographical System,* (5), 93-107.

Mingzhou, J., & Haiyuan, W. (2004). *A Study of System Performance Measures for Intermodal Transportation*. A Thesis Submitted to Faculty of Industrial Engineering, Mississippi State University.

Porikli, F., & Li, X. (2004). Traffic congestion estimation using HMM models without vehicle tracking. *IEEE Intelligent Vehicles Symposium*, pp. 188–193.

Pongpaibool, P., Tangamchit, P., & Noodwong, K. (2007): Evaluation of road traffic congestion using fuzzy techniques. *Proceedings of IEEE TENCON 2007*, Taipei, Taiwan.

Pirie, G. (1981). The possibility and potential of public policy on accessibility. *Transportation Research Part : A,* 15(4), 377-381.

Goh, B.H. (2000), "Evaluating the performance of combining neural networks and genetic algorithms to forecast construction demand: the case of the Singapore residential sector", Construction Management and Economics, Vol. 18, pp. 209-217.

Quiroga, C. A. (2000). Performance measures and data requirements for congestion management systems. *Transportation Research Part C*, 8, 287-306.

Quinlan, J. R. (1993). *Programs for machine learning*. Morgan Kaufmann series in machine learning.

Simon, R. (1997). *Integrating Americans with Disabilities Act Para transit Services and Health and Human Services Transportation (as part of TCRP Project J-6. Quick Response for Special Needs)*. Research Results Digest.

Scholkopf, B. & Smola.B.(2002). Learning with Kernels. MIT Press, Cambridge, MA.

Shawe-Taylor, J. & Cristianini*, N. (2004). Kernel Methods for Pattern Analysis. Cambridge UP,Cambridge, UK.*

Stalinski, P. and Tuluca, S.A. (2006). "The Determinants of Foreign Listing Decision: Neural Networks Versus Traditional Approaches", International Research Journals of Finance and Economics, No. 4. pp.220-231.

Steinwart, L. & Christmann, A. (2008). *Support Vector Machines*. Springer.

Sen, L., Majumdar, S.R., Highsmith, M, Cherrington, L, & Weatherby, C. (2011). *Determine Performance Measures for Public Transit Mobility Management*. Texas department of transportation.

Searcy, C. (2012). Corporate Sustainability Performance Measurement Systems: A Review and Research Agenda. *Journal of Business Ethics,* (107), 239-253.

Transportation Research Board. National Research Council. (1997). *Quantifying*

congestion. Retrieved from http://onlinepubs.trb.org/onlinepubs/nchrp/nchrp_rpt_398.pdf.

Texas Transportation Institute. (2005). *The Keys to Estimating Mobility in Urban Areas Applying Definitions and Measures That Everyone Understands*. Retrieved from http://d2dtl5nnlpfr0r.cloudfront.net/tti.tamu.edu/documents/TTI-2005-2.pdf

Tsai, L. W., Chean, Y. C., Ho. C. P., Gu, H. Z., & Lee, S. Y. (2011). Multi-Lane detection and road traffic congestion classification for intelligent transportation system. *2011 3rd International Conference on Machine Learning and Computing (ICMLC 2011).*

U.S. Environmental Protection Agency (EPA 231-K-10-004). (2011). *Guide to Sustainable transportation performance measures.* Retrieved from http://www.epa.gov/dced/pdf/Sustainable_Transpo_Performance.pdf

U.S. Department of Transportation. (2005). *Travel Time reliability. Make it There on Time all the time*. Federal High Way Administration. Retrieved from http://ops.fhwa.dot.gov/publications/tt_reliability/brochure/

U.S. Department of Transportation. (2003). *Strategic Planning and Decision Making in State Departments of Transportation*. Retrieved from http://www.dot.gov/PerfPlan2004/intro_strategic.html

Wang, Y., Chen, Y., Qin, M., & Zhu, Y. (2006). Dynamic traffic prediction based on traffic flow mining. *6th World Congress on Intelligent Control and Automation*.

Yu, L., Liu, M., Shi, Q., Song, G.(2010). Macroscopic Congestion Intensity Measurement Model Based on Cumulative Logistic Regression. *The Open Transportation Journal*. 4,44-51.

Zeidenberg, M. (1990), Neural Network in Artificial Intelligence, Ellis Horwood, New York, NY.

Zhan-quan, S., Jin-qiao, F., Wei, L., & Xiao-min, Z. (2012). Traffic congestion identification based on parallel SVM. *8th International Conference on Natural Computation (ICNC 2012).*

Zheng, J., Garrick, J.N.W., Palombo, C.A., McCahill, C., & Marshall, M. (2013). Guidelines on developing performance metrics for evaluating transportation sustainability. *Research in Transportation Business & Management,* (7), 4-13.

VITAE

Graduate College

University of Nevada, Las Vegas

Abbas Mirakhorli

Degrees:

Industrial Engineering, 2008

Faculty of Industrial Safety and Health, Shaheed Beheshti University of Medical Sciences and Health services, Tehran, Iran

Thesis Title:

A COMPARATIVE STUDY:

Utilizing Data Mining Techniques to Classify Congestion Status

Thesis Examination Committee:

Committee Co-Chair, Alexander Paz, Ph.D.

Committee Co-Chair, Brendan Morris, Ph.D.

Committee Member, Mohamed Kaseko, Ph.D.

Committee Member, Pramen P. Shrestha, Ph.D.

Graduate Faculty Representative, Venkatesan Muthukumar, PhD.