Application of Data Mining Techniques for Classification of Traffic Affecting Environments

Kanokwan Khiewwan[1,*], Phrommate Weeraphan[2], Khumphicha Tantisontisom[2], Jindaporn Ongate[2]

[1]Faculty of Industrial Technology, Kamphaeng Phet Rajabhat University, Kamphaeng Phet 62000, Thailand
[2]Faculty of Science and Technology, Kamphaeng Phet Rajabhat University, Kamphaeng Phet 62000, Thailand
***Corresponding author's email:** kanokwan_kh@kpru.ac.th
Received: 08/05/2020, Accepted: 12/06/2020

Abstract

This research studied different data mining techniques use for classifying data on traffic volume and factors influencing traffic. The data that was analyzed consisted of 31, 147 records from the westbound traffic data volume of MN DOT ATR station 301, which is roughly midway between Minneapolis and St Paul, MN. The data that was retrieved from the UCI Machine Learning Repository were from 2014 to 2018. The studies showed that the Decision Tree (DT) is the technique that provides the highest accuracy for data classification at 79.79 percent, followed by the *k* Nearest Neighbor (KNN), with accuracy at 72.79 percent with $k = 1$, and finally, the Support Vector Machine (SVM) with accuracy at 59.54 percent. Additionally, DT can identify that time data, which is an essential factor affecting the traffic volume.

Keywords:
Traffic, Data Mining, Decision Tree, K Nearest Neighbor, Support Vector Machine

1. Introduction

Smart mobility is one of the objectives to improve quality of life of urban people in a smart city. It provides basic infrastructure and smart problem solutions to traffic infrastructure development which currently focuses on supporting the increasing demand and avoidance of traffic congestion. Solving traffic jam is a major issue in every big city [1]. The main cause of traffic volume is from an unbalanced transportation system because of increasing demand and supply [2]. Therefore, smart mobility development is designed to improve transportation system more effectively. Nowadays, computer technology, together with information and communication technology (ICT) are developed with most accuracy and efficiency, is promptly widely disseminated.

For example, the wireless sensor system installation collects the weather, environment, time data that relate to traffic problems. Data that may influence traffic volume or traffic congestion can be obtained, which can be analyzed and can then be used in designing efficient solutions to traffic problems.

Machine learning and data mining are computer science technologies for collecting and analyzing data and then, classifying data for accuracy. Data classification is an important technique in Machine Learning and Data Mining. There are many studies which have applied this method to generate models that can be applied to new data collection to develop precise automated learning systems.

The article of Bhavsar et al. described machine learning methods for data analytics in transportation [3]. In the present, the researchers are focusing on improving existing Intelligent Transportation Systems (ITS) applications and developing new ITS applications that rely on the quality and size of the data. Machine learning methods can significantly improve data analytics for the transportation system. The research of Zhang et al. used the *k* Nearest Neighbor (KNN) model, which predicted the number of car movements towards highways, and compared this with the regression model [4]. The results showed that the accuracy of the presenting technique is more than 90%. The research of Antonio explored the system

used by city planners to classify locations in the city according to the traffic elements, the Artificial Neural Network and Support Vector Machine [5]. The test results showed that the Support Vector Machine provided a better solution in location classification. Furthermore, it appeared that these methods could solve traffic problems on the road quickly, including collecting data on traffic noise on the streets.

According to these previous researches, classification can identify the problems relating to traffic congestion effectively, as such, the researchers of this proposed study is to apply this technique to classify data to determine whether weather variables or different time period affect the traffic and if there is the correlation between them. The study has two main purposes. The first one is to find a technique that is suitable for traffic classification, and the second one is to explore factors that affect the traffic. This study will be done based on 31,147 records of dataset.

In this study, the researchers compared the behaviors of three data mining techniques, namely; "Decision Tree", "Support Vector Machine" and "*k* Nearest Neighbor". The study compared the accuracy in model development of each technique. The researchers aimed to select the technique that provided the best results for decision making in developing e smart mobility in order to solve traffic problems in the future.

2. Methodology

2.1 Dataset

The Metro Interstate Traffic Volume datasets from UCI Machine Learning Repository was used in this study. The tested data had 31,147 instances from the year 2014 to 2018. The data were divided into five groups. There were 4,502 records in 2014, 3,599 records in 2015, 7,811 records in 2016, 8,689 records in 2017 and 6,546 records in 2018. Each year, the original set was separated d into 80% train data and 20% test data. The dataset had 9 attributes and 3 classes. The detailed attributes are shown in Table 1, and the detailed classes are shown in Table 2.

Table 1: Attributes of Dataset

No.	Attribute	Description	Values
1	temp	Numeric Average temp in kelvin	Actual Average temp in kelvin
2	rain_1h	Numeric Amount in mm of rain that occurred in the hour	Actual Amount of rain
3	snow_1h	Numeric Amount in mm of snow that occurred in the hour	Actual Amount of snow
4	clouds_all	Numeric Percentage of cloud cover	Actual Percentage of cloud cover
5	weather_main	Categorical Short textual description of the current weather	Clear, Clouds, Drizzle, Fog, Haze, Mist, Rain, Smoke, Snow, Squall, Thunderstorm
6	weather_description	Categorical Longer textual description of the current weather	light-snow, broken-clouds, drizzle,few-clouds, fog, freezing-rain, haze, heavy-intensity-drizzle, heavy-intensity-rain, heavy-snow, light-intensity-drizzle, light-intensity-shower-rain, light-rain, light-rain-and-snow, light-shower-snow, lightsnow, mist, moderate-rain, overcast-clouds, proximity-shower-rain,

No.	Attribute	Description	Values
			proximity-thunderstorm, proximity-thunderstorm-with-drizzle, proximity-thunderstorm-with-rain, scattered-clouds, shower-drizzle, shower-snow, sky-is-clear, sleet, smoke, SQUALLS, thunderstorm, thunderstorm-with-drizzle, thunderstorm-with-heavy-rain, thunderstorm-with-light-drizzle, thunderstorm-with-light-rain, thunderstorm-with-rain, very-heavy-rain
7	Day	Day of the data collected	range 1-31
8	Month	The month of the data collected	range 1-12
9	Time	Time of the data collected	range 0-23

Table 2**:** Classes for Prediction

No.	Class	Traffic Volume
1	small	1-2,500
2	medium	2,501-5,000
3	large	5,001-7,500

The data were divided into three sections. The density group was the underlying basis for dividing the data into 3 classes, for various distribution. The minimum traffic volume is 0 and the maximum is 7,280. It was inaccurate to divide them into fewer or more classes.

2.2 Experiment software

All experiments on the classifiers described in this research were conducted using algorithms from Weka. Weka which is a well-defined software applied for real-world problems [6]**.** Weka consists of three data mining techniques such as Classification, Clustering and Association**.** The software generates the model from learning data input and analysis can be done with the model using the test data**.**

2.3 Decision Tree Classification

Decision Tree **(**DT**)** is a type of data mining technique that is used to build classification models in the form of a tree-like structure. It is a structure that includes a root node, branches, and leaf nodes. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label**.** The topmost node in the tree is the root node**.** The challenge of creating a decision tree is creating the smallest decision tree [7]. The most acceptable algorithm is C4.5 that can solve only one learning solution**.** It cannot be used if it led to overfitting**.** In Weka program, this study applied y C4.5 as an algorithm in decision tree classification**.**

2.4 Support Vector Machine Classification

Support Vector Machine is a technique for data classification by analysis and structure recognition**.** The most well-known feature of SVM is the linear classifier, predicting each input's member class between two possible classifications. This more accurate technique builds a hyperplane or set of hyperplanes to classify all inputs in a high-dimensional or even infinite space [8]. The closest values to

the classification margin are known as support vectors. The SVM's goal is to maximize the margin between the hyperplane and the support vectors, which can classify a non-linear efficiently.

2.5 k Nearest Neighbor Classification

k Nearest Neighbor is a technique for classification by finding the closest value to k. In this technique, the output is a class membership. The member is assigned to the class most common along with its *k* Nearest Neighbor [4] In each classification, there might be a movement because the membership is very close to other neighbors**.**

2.6 Research method/Experiment Process

From the study of the original data (from daily sensors collected from 2012 to 2018), factors were classified to receive good information. The researchers separated data into each year. The test data was chosen for 5 years, from 2014 to 2018. The results showed which variables and correlations affect the traffic volume for each year. This is shown in Fig. 1.

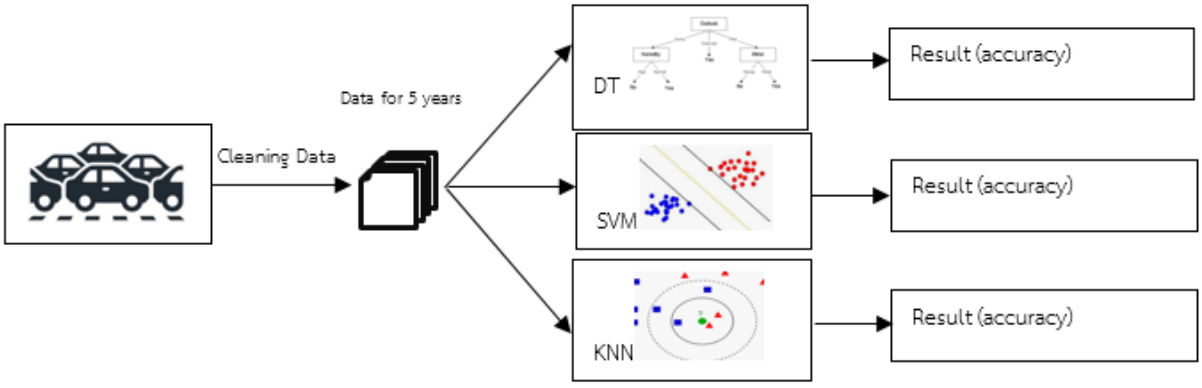


Fig. 1 Overall experiment.

Fig. 1 shows the overall experiment. For the first step, the researchers did the cleaning of data to identify missing data and then converted the data into the format that can be applied to Weka**.** Then, the data were divided into data sets**.** Each set consisted of data on temperature, rain measurement, snow measurement, and car measurement, for every hour, for every day, for each year. Then, the data were imported into the software to create a classification model**.** The results indicated the accuracy values of the data classification of the different techniques. The results are as shown in Table 3 and Table 4.

3. Results and Discussion

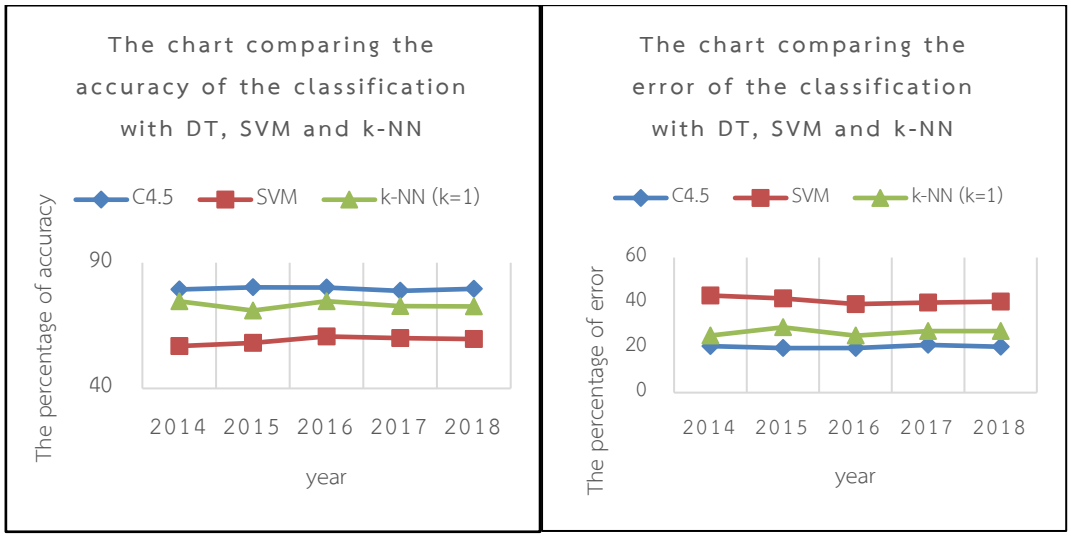
The results of the experiments were analyzed in terms of effectiveness and efficiency. The effectiveness of all classifiers was evaluated in term of accuracy and error. The results are shown in Table 3-4 and Fig. 2.

Table 3 The accuracy of data classification by using DT, SVM and KNN.

Data mining Techniques	**Accuracy Value (%)**				
	2014	**2015**	**2016**	**2017**	**2018**
DT (C4.5)	78.42	80.35	81.44	79.10	79.63
SVM	59.58	59.27	60.27	60.11	58.49
KNN (*k*=1)	74.11	71.60	73.88	71.85	72.52

Table 4 The error values of data classification by using DT, SVM and KNN.

Data mining Techniques	Error Value (%)				
	2014	**2015**	**2016**	**2017**	**2018**
DT(C4.5)	21.58	19.65	18.56	20.9	20.37
SVM	40.42	40.73	39.73	39.89	41.51
KNN (*k*=1)	25.89	28.4	26.12	28.15	27.48



(a) (b)

Fig. 2 (a) The chart comparing the accuracy of the data classification using three technique (b) The chart comparing the error of the data classification using three techniques.

Fig. 2 (a) shows a comparison of the percentages of the accuracy of classification of data from 2014 to 2018 using DT (Decision Tree), SVM (Support Vector Machine**)** and KNN (*k* Nearest Neighbors). It shows that DT has the highest accuracy at 79.79%, followed by KNN with an average accuracy of 72.79% (k=1) and, finally, SVM with accuracy at 59.54%.

Fig. 2 (b) shows the comparison of errors in the data classification. It shows that SVM contains the highest errors, following by KNN and finally DT. The reason that DT has the most accurate data classification, where the tested data have both discontinuous data such as weather, date, time, and continuous data such as temperature. The expected result is the number of cars divided into three groups**.** The decision tree can classify both continuous variables and discrete variables [9].

The researchers also tested the k-value of KNN. A k value equals to 1 indicates that data mining is the most accurate. It means that the data is highly fragmented or not related to the closest neighbors [10]. SVM is the least accurate for data mining because SVM can only classify linear and non-linear problems and theoretically separate the dataset into two groups [11]. However, the researchers divided the data into three datasets, so it was not suitable to use SVM.

Once the predictive model was built, the measurement of accuracy was compared based on precision and recall, which are indications of the efficiency of the model. The results of the comparison are shown in Table 5.

Table 5 Confusion matrix of predicted with a classification model.

		DT				SVM				KNN			
		small	normal	large	Recall	small	normal	large	Recall	small	normal	large	Recall
2014	small	238	27	22	82.93%	219	68	0	76.31%	211	50	26	73.52%
	normal	18	275	60	77.90%	79	274	0	77.62%	39	207	107	58.64%
	large	17	87	158	60.31%	102	160	0	0.00%	17	72	173	66.03%
	Precision	87.18%	70.69%	65.83%		54.75%	54.58%	0.00%		79.03%	62.92%	56.54%	
	f-measure	85.00%	74.12%	62.95%		63.76%	64.09%	0.00%		76.17%	60.70%	60.92%	
2015	small	229	38	30	77.10%	228	56	13	76.77%	207	51	39	69.70%
	normal	17	194	60	71.59%	164	89	18	32.84%	42	138	91	50.92%
	large	8	28	115	76.16%	91	58	2	1.32%	18	65	68	45.03%
	Precision	90.16%	74.62%	56.10%		47.20%	43.84%	6.06%		77.53%	54.33%	34.34%	
	f-measure	83.12%	73.07%	64.61%		58.46%	37.55%	2.17%		73.40%	52.57%	38.97%	
2016	small	462	111	50	74.16%	433	190	0	69.50%	398	158	67	63.88%
	normal	28	528	99	80.61%	220	435	0	66.41%	109	369	177	56.34%
	large	4	81	199	70.07%	138	146	0	0.00%	38	129	117	41.20%
	Precision	93.52%	73.33%	57.18%		54.74%	56.42%	0.00%		73.03%	56.25%	32.41%	
	f-measure	82.72%	76.80%	62.97%		61.24%	61.01%	0.00%		68.15%	56.29%	36.28%	
2017	small	465	83	73	74.88%	464	157	0	74.72%	432	88	101	69.57%
	normal	36	532	135	75.68%	336	367	0	52.20%	152	323	228	45.95%
	large	2	92	313	76.90%	250	163	0	0.00%	30	172	211	51.09%
	Precision	92.45%	75.25%	60.08%		44.19%	53.42%	0.00%		70.36%	55.40%	39.07%	
	f-measure	82.74%	75.46%	67.46%		55.54%	52.81%	0.00%		69.96%	50.23%	44.28%	
2018	small	372	32	29	85.91%	334	99	0	77.14%	321	65	47	74.13%
	normal	27	498	53	86.16%	130	448	0	77.51%	92	364	122	62.98%
	large	26	100	172	57.72%	136	162	0	0.00%	35	135	128	42.95%
	Precision	87.53%	79.05%	67.72%		55.67%	63.19%	0.00%		71.65%	64.54%	43.10%	
	f-measure	86.71%	82.45%	62.32%		64.67%	69.62%	0.00%		72.87%	63.75%	43.03%	

The DT technique can classify the most important factor in data mining, which is used to determine the Root Node. From the experiment, time is the most important factor influencing the traffic volume, as shown by the use of DT on the 2016 dataset (see Fig. 3).

Fig. 3 shows that the decision tree demonstrates that time is the most important factor that affects the traffic volume. It has the leaf node as the target result to classify the traffic volume data. The example of the decision tree for the 2016 dataset shows that the "5.00 AM - 8.00 AM" time slot has densest traffic volume.

The three models for data classification techniques have different classification methods. The decision tree determines the information to be collected to select the classification model. The Support Vector Machine is a linear model for classification and regression analysis. It can solve linear and non-linear problems and work well for many practical problems. The algorithm creates a line or a hyperplane which separates the data into classes. The *k* Nearest Neighbor, each sample should be classified similarly to its surrounding samples. Therefore, if the classification of a sample is unknown, then it could be predicted by considering the classification of its nearest neighbor samples.

Table 5 shows the simple metric movements derived from high frequency predicted test data. From the table, the decision tree shows higher precision and recall than other techniques. Due to highly dispersed data, it is very difficult to classify the data into 3 groups with the linear SVM technique, just like KNN. Finding similar vectors is difficult, but the decision tree can show the relationship between the factors from the information gain, making the prediction more accurate than other techniques.

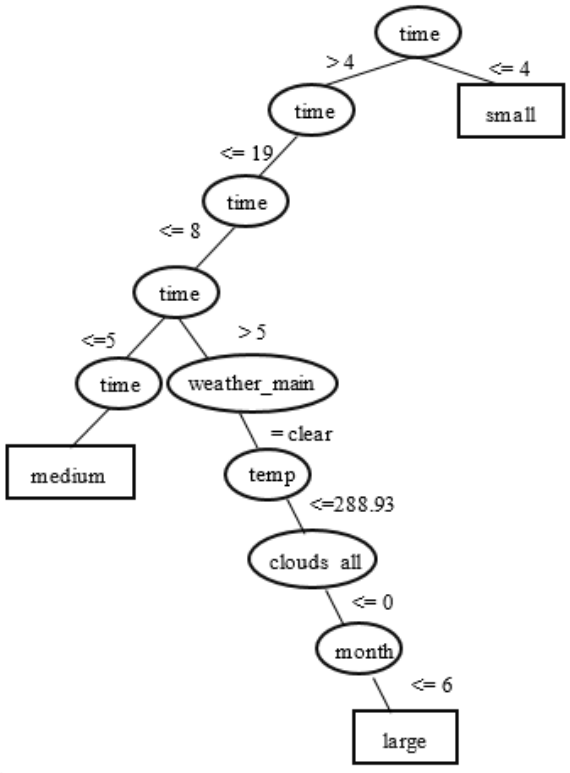


Fig. 3 Example of the decision tree of 2016 dataset.

4. Conclusion

This paper presents the use of application data mining techniques such as DT, SVM and KNN, which can be used to identify the factors affecting traffic jams. Studies done in this work include h dataset gathered from the West streets between Minneapolis and St. Paul, Minnesota. Data on traffic volume were collected hourly every day from 2014 to 2018. Data were accessed from UCI under the file title is Metro_Interstate_Traffic_Volumn.csv [12]. The studies here identified the appropriate to determine the prediction factors affecting traffic volume.

Fig. 3 shows that DT can be used to develop a decision tree model that provides the highest information gain, which is the time. The patterns extracted from the analyses have been used to predict traffic jams early enough to avoid retention at congested points, minimizing response time to these events and providing alternatives to traffic circulation.

To achieve this, the will-known DT algorithm had been successfully applied to obtain results with an accuracy of more than 75%, depending on the anticipation time of the prediction made. Other algorithms, such as SVM and KNN were also applied, but results were less accurate.

The information generated by these studies can be useful for smart mobility development for traffic management. The studies show that that “time” is the most important factor affecting traffic volume. In future research, studies should aim to develop the GPS data and information of vehicles in transportation planning to reduce cost.

References

[1] Khaimook, S., Yoh, K., Inoi, H., & Doi, K. (2019). Mobility as a service for road traffic safety in a high use of motorcycle environment. *IATSS Research, 43*(4), 235 - 241.

[2] Sun, Y. (2012). Research on urban road traffic congestion charging based on sustainable development. *Physics Procedia, 24*, 1567 - 1572.

[3] Bhavsar, P., Safro, I., Bouaynaya, N., Polikar, R., & Dera, D. (2017). Machine learning in transportation data analytics. *Data Analytics for Intelligent Transportation Systems*. 283-307.

[4] Zhang, L., Liu, Q., Yang, W., Nai, W., & Dong, D. (2013). An improved K-nearest neighbor model for short-term traffic flow prediction. *Procedia - Social and Behavioral Sciences, 96,* 653 - 662.

[5] Antonio, J. (2016). Automated classification of urban locations for environmental noise impact assessment on the basis of road-traffic content. *Expert Systems with Applications, 53,* 1 - 13.

[6] Dash, R. (2013). Selection of the best classifier from different datasets using WEKA. *International Journal of Engineering Research & Technology (IJERT), 2*(3), 1 - 7.

[7] Tan, L. (2015). Code comment analysis for improving software quality. *The Art and Science of Analyzing Software Data,* 493 - 517.

[8] Gove, R. (2012). Machine learning and event-based software testing: classifiers for identifying infeasible GUI event sequences. *Advances in Computers, 86,* 109 - 135.

[9] Yan-yan, S., & Ying, L. (2015). Decision tree methods: applications for classification and prediction. *Shanghai Archives of Psychiatry*, *27*(2), 130 - 135.

[10] Imandoust, S. B., & Bolandraftar, M. (2013). Application of K-nearest neighbor (KNN) approach for predicting economic events: theoretical background. *Journal of Engineering Research and Applications, 3*(5), 605 - 610.

[11] Entezari-Maleki, R., Rezaei, A., & Minaei-Bidgoli, B. (2009). Comparison of classification methods based on the type of attributes and sample size. *Journal of Convergence Information Technology, 4*(3), 94 - 102.

[12] Center for Machine Learning and Intelligent Systems. (2019). *Metro interstate traffic volume data set*. Retrieved February 20, 2019, from https://archive.ics.uci.edu/ml/datasets/metro+interstate+traffic+volume