International Journal of Innovation, Creativity and Change. www.ijicc.net
Volume 14, Issue 8, 2020

Prediction and Protection of Car Driving Accident in Urban Zone

Niyom Sutthaluang[a], Somchai Prakancharoen[b], [a]Faculty of science, ChandrakasemRajabhat University, Bangkok, Thailand. [b]Faculty of applied science, KingMongkut's University of Technology North Bangkok, Thailand. Email: [a]niyom.s@chandra.ac.th,[b]somchai-prakan@hotmail.com

Car driving accidents are unsatisfying incidents that should not be occurring. According to accident statistics of Thailand car accidents are occurring in higher population proportion than many countries. Two thousand and four hundred participants were asked to answer a questionnaire about their experience of car accidents during 2018-2019. An accidental occurring pattern was calculated with Rough Set and Structural Equation Modelling. These patterns were used to predict accident damage for any observation which does not known about the accident damage. Since there were many rules the calculation of Euclidean distance consumed too much retrieval time. The performance of matched rule retrieval computing time was enhanced with weighted Euclidean dissimilarity retrieval computing time. It could decrease weighted, normalised Euclidean distance by about 7.5%. The accuracy of cross validation testing was about 85.91%. Prediction and protection of car accident web application were developed especially on high and highest damage magnitude level. One hundred Chandrakasem Rajabhat University student volunteers participated in the road accident testing. This group of students was pre-tested on vehicle driving. Records were kept on the accident and its damage magnitude that were occurring during the pre-testing.. After that, this group of students was assigned to learn about twenty important causes of road accidents, such as weather, road situation, traffic signals and driver's condition. After that, all of them were secondly assigned to drive a car on a particular testing route. Paired t-test technique result showed that the average road accident, after accident prevention program training, was significantly less than those who did not join the training on accident prevention program.

Key words: *Car Driving Accident, Prediction.*

Introduction

World health organisation (WHO) reported that Thailand's road traffic deaths was ranked no.3 in the world (Thailand parliamentarian seminar, 2019). According to Thai RSC's annual data January–July 2020 accident statistics report, there are 582,008 in total persons were suffered with 572,897 classified as injuries and 9,111 deaths (Technology Information Department, 2019). The traffic accident loss is an important life and economic issue that must be solved. The objective of this research is to find out Thailand traffic accident occurring pattern in order to define accident protection mechanism. Related important accident attributes that were reported in relevant research were considered for the questionnaire. Each attribute was defined a five ordinal data level of severity or attribute magnitude from lower (1) to highest (5) based on the Likert 5 point scale. These conditional attributes were used to construct a questionnaire. The class attribute of this research was the level of accident severity, with 3 levels (lower, moderate and severe). Data collections were conducted from January 2018 to September 2019.

There were about two thousand five hundred observations, from the collected data. Insight pattern of accident patterns was trained and tested for its validation. Rough set, classification techniques were used to classify the severity of class variable, the damage magnitude. These insight patterns of traffic accidents were used to create digital media, which explained about the cause of the accident and the prevention method. The important causes of traffic accident were presented with multimedia to one hundred Chandrakasem Rajabhat University students so that all of them are informed about important traffic accident causes. These students were assigned to the driving test on predefined routes. Based on prediction and protection of car driving accident application, students had to input some data, such as drinking level, sleepiness level, etc. to a computer application before they started the car driving test. The application, based on input data, should forewarn users about level "2" and "3" of severity of damage level on dangerous points which student will drive through. Any accidents occurring were measured in order to test if the traffic accident pattern reliability.

Related theory and research

Rough Set Classification *(JiaweiHan, 2012)*

Rough sets are a mathematical technique that can find out the patterns or rule of conditional attributes that classify the class variable. Some observations (X) should have indiscernibility (IND) or similar value in a set of defined attributes (A).

$$IND_{is}(B) = \{(x, x') \in U^2 | \forall \in B, a(x) = a(x')\} \quad (1)$$

Where:
IS - information table compose of (U, A)
U - non empty finite set of objects
x - an object or observation
x - another x object
A - set of conditional attributes
a - element of set B
B - set of attribute which $B \in A$

These indiscernibility groups or set of observations will be used to set an approximation of x observation to class attribute(D) based on B. Otherwise, some subset of the IND_{is} (B) have some members that might match, to define a class attribute value.

There are two X approximation: B-lower and B-upper. B-lower approximation is a situation that all subset of IND_{is} (B), all members, match to define a class attribute value.

$$\underline{B}X = \{x | [x]_B \subseteq X\} \quad (2)$$

Where $[x]_B$ is B indiscernibility relation B-upper approximation is situation that some subset of INDis (B) all members match to define a class attribute value.

$$\overline{B}X = \{x | [x]_B \cap X \neq \varnothing\} \quad (3)$$

The observation or object which matched B-lower is certainly related to class variable (D) value. While objects in B upper have some objects that match to D value, but some objects do not. Objects in B-lower represent sets of objects that certainly point to class variable D.

The attribute value patterns of each conditional attribute are formed up the group of rules that all of them are pointing to a particular D value. There may be some conditional attributes that have no any effect on pattern generating. These sets of attributes could be eliminated from Rough Sets generating. The criteria of suitable attribute deletion are based on that the stability coefficient must keep its value on 1.00. Reduct attribute sets are the minimal set of attributes that could preserve indiscernibility relation and set approximation.

Gain ratio (Saed Sayad, 2020)

Gain ratio is an improvement of information gain. Gain ratio could be used to measure the amount of conditional attribute information (or importance) of splitting in decision tree generation. Gain ratio can be calculated from equation (4).

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo(A)} \quad (4)$$

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|Dj|}{|D|} \times \log_2\left(\frac{|Dj|}{|D|}\right) \quad (5)$$

$$Gain(A) = Info(D) - Info_A(D) \quad (6)$$

$$Info_A(D) = \sum_{j=1}^{v} \frac{|Dj|}{|D|} \times Info(Dj) \quad (7)$$

$$Info(D) = -\sum_{i=1}^{m} p_i \times \log 2(p_i) \quad (8)$$

Where:
D - data partition (class label j=1,m)
C_i - the class label attribute
P_i - the probability of an arbitrary sample belongs to class C_i
D_j - a subset of a discrete attribute values (I = 1, v).
Info (D) - an entropy of D.
$Info_A$ (D) - expected information that is used to classify tuple from D based on partitioning by attribute A.
Gain (A) - expected information reduction by choosing attribute A.

Gain ratio (A) is widely used in attribute selection in many data mining algorithms (NarasimhaPrasad, 2013) since it reduces the bias toward attributes with large number of values.

Weighted dissimilarity (Statistical engineering division, 2017)

Euclidean distance is a mathematical technique that is used to calculate the distance value between samples or object.

$$d_{o1,o2} = \sqrt{\sum_{1}^{n} (a_i - b_i)^2} \quad (9)$$

Where:

$d_{o1,o2}$ - a distance value between object o1 and o2

a_i, b_i - two attribute's value of calculating two samples. i is a running number of attribute, i = 1, n

If $d_{o1,o2}$ value is greater than 0 then two samples are dissimilar. Whereas, two samples are identical if the $D_{o1,o2}$ value is zero. Thus the similarity measure is opposed to dissimilar. Since the Euclidean dissimilarity value boundary has no limited upper bound value, then normalised Euclidean distance is proposed in order to set the limit value of distance between 0-1.

$$Nd_{o1,o2} = \frac{\sqrt{\sum_{1}^{n} (a_i - b_i)^2}}{\sqrt{\sum_{i=1}^{n} a_i^2} + \sqrt{\sum_{i=1}^{n} b_i^2}} \quad (10)$$

Where:

$Nd_{o1,o2}$ is the normalised Euclidean distance which its value is bounded in $0 \leq Nd_{o1,o2} \leq 1$.

Michael Greeacre, (2008) states in some situations, some attributes have more effective to class variable than other attributes. Therefore, weighting of all attributes should not have an equivalence weight. Equation (11) represents the normalised weighted Euclidean distance formula.

$$d_{wen}(i, j) = \frac{\sqrt{\sum_{k=1}^{p} w_k (x_{k,i} - x_{k,j})^2}}{\sqrt{\sum_{k=1}^{p} w_k x_{k,i}^2} + \sqrt{\sum_{k=1}^{p} w_k x_{k,j}^2}} \quad (11)$$

Where:

$0 \leq d_{wen}(i, j) \leq 1$

d_{wen} - weighted Euclidean normalised distance or dissimilarity between sample oi and oj.

There are many techniques used to calculate weighting value defining, such as chi-square, standard deviation and correlation. The researcher chose a Gain ratio since it is

widely used in attribute selection in data mining. This weighting method considers each attribute's information gain on splitting partition based on some attribute (class attribute). Since summation of all attributes gain ratio is not 1.00, thus this research has adjusted their value into the scale of 0-1 by using equation (12).

$$W_{ai} = \frac{W_i}{\sum_{i}^{n} W_i} \quad (12)$$

Where:
W_i - a gain ratio of attribute i.
W_{ai} - normalised value of W_i

$$0 \le \sum_{i}^{n} W_{ai} \le 1.00$$

On the other hand, similarity between an object, where its value lays between 0.00 and 1.00, can be found out by equation (13).

$$S_{wen}(i, j) = 1 - d_{wen}(i, j) \quad (13)$$

Structural equation modelling (Diana Suhr, 2020)

Structural equation modelling is a statistical method that was used to explore factors and illustrates or confirms their causal relationship. Factor analysis is used for dimension or attributes reduction. Some attributes are significantly related or interact, thus these attributes are composed to be a new latent variable or factor.

Some attributes have no significant relation to any others therefore this attribute should be eliminated (factor loading $\le$ 0.30). The statistics that represent if factor analysis is acceptable is KMO, KMO > 0.50 (Leonardo Miljko, 2017).

After dimension reduction, all factors are proposed their relationship subject to researcher's hypothesis. This model is tested for model fitting (Cornell statistical consult, 2020) based on many statistics goodness of fit indices as present in Table 1.

Table 1: SEM model fit indices

Measure	Label	Cut-off for good fit
x^2	Chi-square of model	p-value>0.05
GFI	Goodness of fit	GFI>=0.95
RMSEA	Root mean square error of approximation	RMSEA<0.08

The modified model that passed the model fitting test should represent factor relationship.

The relationship may be one direction or covariance between two factors. The relationship will show which factor is cause or effect. Bidirectional or covariance relationship between two factors is presented if there is statistical covariance or correlation between them (Will Monroe, 2017).

Related research

A classification and recognition model for the severity of road traffic accidents in mechanical engineering was developed by Xi Jianfeng et al., (2019). A road traffic accident severity pattern was considered on human, vehicle and environmental factors. Rough Set Theory was used to choose the effective factors in road accidents. A Support Vector Machine (SVM) technique was used to establish road accidental severity.

The comparison of classification accuracy showed that SVM calculation, on effective factors, gave more accuracy classifications than SVM calculation on all factors. The proposed conditional attributes were driver attribute, vehicle attributes, driving environment, road attributes and accident attribute while class attribute was accident attribute (accident type, accident time and month).

The factors that contribute to road accidents (Jonathan J. Rolisona, 2018)

Road accident records from the official road accident department present the major factors that cause an accident. For young drivers, accidents are frequently occurring due to the inexperience factor, lack of skill and risk taking behaviours (fast driving, drinking). While, older driver accidents factors were medical condition, distraction and eye sight.

Profiling High Frequency Accident Locations Using Association Rules (Koen Vanhoof, 2003)

Traffic accident association rule was used to construct sequences of many attributes such as traffic condition, environment condition, road condition, human condition and geographical condition. Based on accident type, these association rules were invented on differing circumstances. Therefore, a particular accident type should be happening if the situations were matched to some association rule.

A Data Mining Framework to Analyse Road Accident Data (Sachin Kumar, 2015)

Road accidents in Dehradun (India) dataset were composed of many attributes such as lighting on the road, roadway feature, road type, area around the road and type of accident. About 11,574 accident observations during 2009-2014 were clustered by k-modes clustering algorithm. These clusters were mining for accident rules by using association rule. The finding rules gave important information to stakeholders to use as a guideline to prevent and reduce accident occurring.

Cause Analysis of Traffic Accidents Based on Degrees of Attribute Importance of Rough Set (Gang Tao, 2015)

There are many attributes that related to accident evidence. Some attributes might have no effect on the class variable classification. Some attribute might cause the inconsistent rule.

Rough Set Theory was used to detect this kind of attributes. The certainty rules were generated on prediction of traffic accidents. There were many conditional attributes used to generate the Rough Set rule such as sex, age, degree of education, mode of transport, vehicle safety condition, weather, visibility, lighting, topography, traffic signals, type of road and road line type.

Vehicle, and Environment Characteristics on Collision Warning System Design (Yong-Seok Kim, 2003)

There are many factors that are the major cause of road accidents such as vehicle, driver and the environment. This research studies the technical detail of vehicle brake efficiency, road situation and driver behaviour. The results of this study should help drivers to prevent car accident occurring or reduce car collision.

Role of Gender in Road Accidents (Adlih Al, 2003)

The traffic accident record in Jordan was studied. The root causes of car accident were different in younger and senior driver.
Evident statistics show that male drivers are a larger proportion than female drivers in car accidents. Impatience and lack of attention, especially in younger driver, are important factors of a car accident happening.

Research methodology
This research was conducted using research methodology as described in Figure 1.

1. Feature or related conditional attributes were gathered from related research.
These conditional attributes were considered by five road accident experts if these attributes were possible or real road accident causes.

2. Questionnaire, which was composed of related conditional attributes, and used to collect the data about road accident during 2018-2019.
Two thousand four hundred observations were collected from Bangkok capital and six nearby provinces of the Kingdom of Thailand.

3. After data cleaning, reduct set of attributes was extracted under stability coefficient preservation criteria (1.0).
The certain rules (not inconsistent) were created, under a reduct set of attributes, by Rough Set technique.

4. Attributes Gain ratio was calculated in order to use them as each attributes weighting.
Weighted normalised Euclidean distance was applied to retrieve most similarity training observation with a particular observation where decision or class attributes were unknown. Since there were a large amount of generating rules then information retrieval consumed too much computational time. This research suggests a solution to reduce calculation time. In brief, the largest gain ratio value attribute, weight, was assigned its value in largest value, 5, thus the weighted normalised Euclidean distance was calculated with smallest value 1 attribute. This process was iteratively performed by the next below large gain ratio value attribute. Saturated accumulated distance of a group of attributes was considered as cutting point of training observation retrieval purging. Training observation that has a distance value smaller than cutting point must be calculated weighted, normalised Euclidean distance for all the attributes left.

5.Causal model of all conditional attributes are illustrated by using Structural Equation Modelling.
All conditional attributes were dimension reduced to form new factors by using Factor Analysis technique.

This factor relationship model was purposed and tried out for a suitable causal relationship. Structural equation models should inform direct effect and interaction between factors.

6. Decision tree algorithms were used to extract classification of decision attributes under conditional attributes.
Accuracy of correct classification was compared to Rough Set classification.

7. Twenty road or traffic accident scenarios (rule) and high value of damage magnitude, were chosen for preparing traffic accident evidence prevention multimedia and application program.
The media was used to train the project volunteers about accident important causes.

After that, these volunteers were pre-tested and post-tested on application program usage testing in real vehicle driving. Accident damage magnitude of all volunteers on pre-test and post-test was compared with paired t-test, if the average of the accident damage magnitude was equal.

8. Research result and further research were discussed.

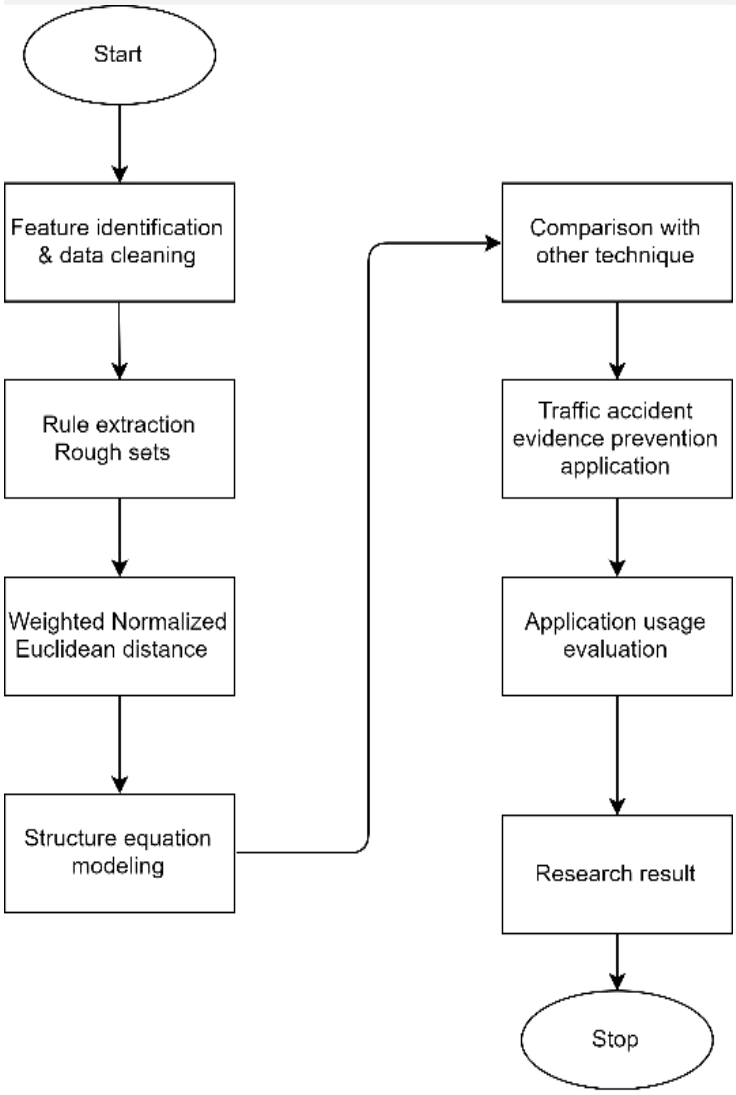

Figure 1: Research methodology

Research result

Related features identification

The research conditional attributes were collected from reviewing papers and Thai royal safety culture.

The attributes were considered as cause attributes of road accidents by five road accident experts. The acceptance of all attributes was proven by 0.8 score of IOC: Index of item objective congruence test. There were four groups (Human, Vehicle, Road and Environment). Each group was composed of their related attributes. For example, the Human group was composed of driver's drinking level, fast driving sleepiness level and visibility level. Damage magnitude level was a decision attribute.

Table 2: Summary of research traffic accident attributes

Group	Attribute	Label	Data type	Categories	Type
Human	DL	Drinking level	Ordinal	1= no, 2-3 a little bit	conditional
				4-5 = much more	
Human	FD	Fast driving level	Ordinal	1=1-50 km/HR, 2 = 51-80 km/HR	conditional
				3 = more than 81km/HR	
Human	SL	Sleepiness level	Ordinal	1 = fresh, 5 = very sleepy	conditional
Human	VL	Visibility level	Ordinal	1= see well, 5 = can barely see	conditional
Vehicle	VC	Vehicle condition	Ordinal	1= good condition, 5 = most incomplete	conditional
Road	RS	Road signal	Ordinal	1= available or fully installed	conditional
				2 = not available or not installed	
Road	SLT	Street lighting	Ordinal	1= available or fully installed	conditional
				2 = not available or not installed	
Road	BR	Broken road	Ordinal	1= good condition, 5 = most incomplete	conditional
Road	RR	Road repairing	Ordinal	1= no road repairing	conditional
				5 = highest road repairing	
Environment	HR	Heavy rain	Ordinal	1-2 = bright, 3-5= heavy rain	conditional
Environment	SRB	Roadside burning	Ordinal	1= not any roadside burning	conditional
				5= highest road repairing	
Decision	DM	Damage magnitude	Nominal	1= no injuries, 2 = slightly injured	class
				3= severely injured	

Data collection and cleaning

Questionnaires were sent to the urban area of Bangkok capital with a one thousand sample size, while one thousand five hundred were sent to five provinces around Bangkok. Two thousand four hundred questionnaires were sent back.

Table 3: Amount of sending and receiving of questionnaire collection

	Sent	Received
Urban	1,000	897
Rural	1,500	1,503
Total	2,500	2,400

Car driving accident pattern classification

The patterns of car driving accidents were generated by using Rough Set and Decision Tree classification tree.

1. Rough set

All conditional attributes, eleven, were calculated for their importance on rule generating by Rough Set algorithm. Two attributes, SRB-RR, were not important or effective on rule generating. These attributes should be deleted. The stability coefficient was still equal to 1 thus deletion of two attributes was acceptable since the generated rules were not altered.

Table 4: Reduct set with stability coefficient

	size	Pos-reg	Stability coefficient	Reduct set
1	10	0.914	1	{DL,FD,SL,VL,VC,RS,SLT,BR,HR,RR}
2	9	0.904	1	{DL,FD,SL,VL,VC,RS,SLT,BR,HR}

RULE_SET
ATTRIBUTES 12
DL symbolic: Drinking level
FD symbolic: Fast driving
SL symbolic: Sleepiness level
VL symbolic: Visibility level
VC symbolic: Visibility capability
RS symbolic: Signal - road
SLT symbolic: Street lighting
BR symbolic: Broken road level
HR symbolic: Heavy rain
SRB symbolic: Side road burn
RR symbolic: Road repairing
class symbolic: Damage magnitude level
DECISION_VALUES 3

Rough Set has generated 1,870 certain rules as shown in table 5.

Table 5: Partial generated certain rules under reduct set

#1=1144, #2=603,#3=142 RULES Total 1870	Frequency
(DL=2)&(FD=1)&(SL=4)&(VL=2)&(VC=3)&(RS=3)&(SLT=4)&(BR=2)&(HR=1)=>(class=1[1])	1
(DL=1)&(FD=1)&(SL=3)&(VL=2)&(VC=3)&(RS=4)&(SLT=4)&(BR=2)&(HR=1)=>(class=1[2])	2
(DL=2)&(FD=1)&(SL=4)&(VL=3)&(VC=3)&(RS=4)&(SLT=5)&(BR=1)&(HR=1)=>(class=1[1])	1
(DL=1)&(FD=1)&(SL=3)&(VL=2)&(VC=2)&(RS=3)&(SLT=3)&(BR=1)&(HR=1)=>(class=1[7])	7
(DL=1)&(FD=1)&(SL=2)&(VL=2)&(VC=2)&(RS=4)&(SLT=3)&(BR=1)&(HR=1)=>(class=1[5])	5
(DL=1)&(FD=1)&(SL=2)&(VL=2)&(VC=2)&(RS=3)&(SLT=3)&(BR=1)&(HR=1)=>(class=1[19])	19
(DL=1)&(FD=2)&(SL=3)&(VL=2)&(VC=2)&(RS=2)&(SLT=2)&(BR=2)&(HR=1)=>(class=1[1])	1
.	
.	
(DL=3)&(FD=1)&(SL=4)&(VL=3)&(VC=2)&(RS=5)&(SLT=4)&(BR=1)&(HR=1)=>(class=1[1])	1
(DL=1)&(FD=1)&(SL=1)&(VL=2)&(VC=2)&(RS=3)&(SLT=2)&(BR=1)&(HR=1)=>(class={1[12],2[1]})	13
(DL=1)&(FD=1)&(SL=3)&(VL=2)&(VC=2)&(RS=3)&(SLT=3)&(BR=2)&(HR=1)=>(class=1[3])	3
.	
.	
(DL=2)&(FD=1)&(SL=4)&(VL=2)&(VC=3)&(RS=5)&(SLT=3)&(BR=3)&(HR=1)=>(class=2[1])	1
(DL=1)&(FD=1)&(SL=3)&(VL=3)&(VC=3)&(RS=4)&(SLT=3)&(BR=1)&(HR=1)=>(class=2[1])	1
(DL=2)&(FD=1)&(SL=1)&(VL=2)&(VC=3)&(RS=4)&(SLT=5)&(BR=1)&(HR=1)=>(class=2[1])	1
.	
.	
(DL=5)&(FD=1)&(SL=5)&(VL=5)&(VC=2)&(RS=2)&(SLT=5)&(BR=1)&(HR=1)=>(class=3[1])	1
(DL=4)&(FD=3)&(SL=1)&(VL=2)&(VC=3)&(RS=5)&(SLT=5)&(BR=1)&(HR=2)=>(class=3[1])	1
(DL=2)&(FD=2)&(SL=1)&(VL=3)&(VC=3)&(RS=5)&(SLT=4)&(BR=2)&(HR=1)=>(class=3[1])	1

2.Weighted Euclidean distance

2.1 Gain ratio

Each attribute was calculated for its Gain ratio based on decision variable, damage magnitude level. Heavy raining (HR) attribute gain ratio value was the largest value at 0.64 while road signal (RS) was the smallest Gain ratio value. All of Gain ratio values were normalised to a new value, called weighting. The summation of all weighting value is 1.00.

Table 6: Gain ratio and normalised weighting of all conditional attributes

Number	Attribute	Gain ratio	Weighting
1	HR	0.64	0.55
2	VL	0.10	0.08
3	VC	0.09	0.08
4	SL	0.08	0.07
5	SLT	0.06	0.05
6	BR	0.06	0.05
7	DL	0.05	0.04
8	FD	0.04	0.04
9	RS	0.03	0.03

2.2 Euclidean distance or dissimilarity

Based on generating rule (Rough set), new observation (with unknown decision attribute) could predict its damage magnitude level by finding out the most similarity with training observations by using normalised Euclidean similarity.

2.3 Euclidean distance calculation time reduction

There were a large amount of rules which were generated from the Rough Set method.

Therefore, amount of rule retrieval computation time should reduce data processing performance. In order to reduce the waiting time, we have to consider on Euclidean weighted normalised dissimilarity. The gain ratio of each reduct attribute, except core attributes, was transformed to have a value of 0-1, a weighting value of each attribute.

Step1: Firstly, choose the attribute that has largest Gain ratio value. Then observation attribute value was set with most difference value. For example, observation #1 was defined all attribute value to 5 and 1 for observation #2. The objective of this attribute value setting was to maximise distance or dissimilarity between two observations. Thus,

the distance summation of these attributes should present a group of attributes that maximise dissimilarity.

Step 2: Calculate the Euclidean distance with weighting and normalised formula between two observations.

1. Start with weighted normalised Euclidean distance calculation between observation #1 and observation#2 on the most gain ratio value attribute, HL (0.55) was calculated. The dissimilarity value was 0.78, as shown in Table 7.

2. The second highest value of gain ratio was VL (0.08).Weighted normalised Euclidean distance calculation between observation #1 and observation#2 on attribute HL and VL was calculated. The accumulated dissimilarity value was 0.67, as shown in Table 7.

3. The third highest value of gain ratio was VC (0.08). Weighted normalised Euclidean distance calculation between observation #1 and observation#2 on attribute HL, VL and VC was calculated. The accumulated dissimilarity value was 0.67, as shown in Table 7.

4. Try out weight normalised Euclidean distance calculation until all attributes are covered. The result of accumulated weighted normalised Euclidean distances is shown in Table 7.

Table 7: Weighted normalised Euclidean distance accumulated between two highest-lowest all attributes observation

Weighting	0.55	0.08	0.08	0.07	0.05	0.05	0.04	0.04	0.03
Variable-value	HL	VL	VC	SL	SLT	BR	DL	FD	RS
5 vs 5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5 vs 1	0.78	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67

Step 3: Consideration of dissimilarity effect attributes.

According to Table 7 attribute HL and VL were two attributes which were the important attributes of dissimilarity between two extreme different value observations. Therefore, Euclidean distance calculation computing time of one observation to all observations (rule) should be reduced by calculating a normalised Euclidean distance between unknown class variable and the high and highest damage magnitude, 2 and 3 first observation. Firstly, Euclidean distance will consider just two variables, HL and VL. If weighted normalised Euclidean distance value is greater than 0.67 then the first training rule or observation was marked as skipped rule. If Euclidean distance value is less than

0.67 then weighted normalised Euclidean distance must continue to be calculated on all conditional weighted attributes left. The rule number and their weighted normalised Euclidean distance were kept in table lookup in order to be considered for the most similar later referencing rule.

Step 4:
After weighted normalised Euclidean distance calculating between an unknown class attribute observation and all training observations, the smallest weighted normalised Euclidean distance of any training observation was chosen. The matched rule's class variable was assigned to an unknown class attribute, request for retrieving, observation. If there are not any rule dissimilarity met to 0.20 dissimilarity threshold, then class variable damage magnitude level 1 (not any damage) will be assigned to an unknown class attribute observation.

The suggested technique was tried out on one hundred observation retrievals. The computational time was reduced by about 7.5% but the accuracy was about 1.63% less than all attribute WEN calculation.

Table 8: Comparison about performance, accuracy between two WEN calculation solutions

	A. All attributes calculation WEN	B. Calculation WEN under accumulate dissimilarity threshold	Compare B to A.
Average time	4 time unit	3.7 time unit	7.5% faster
Correct classify (%)	87.34	85.91	1.63% less accuracy

3. Decision tree

Decision tree j48 algorithm was used to classify class variable, accident damage magnitude, by using Weka Data mining tools (Remco R. Bouckaert, 2013). The accuracy of correct classification value was about 78.29 %.

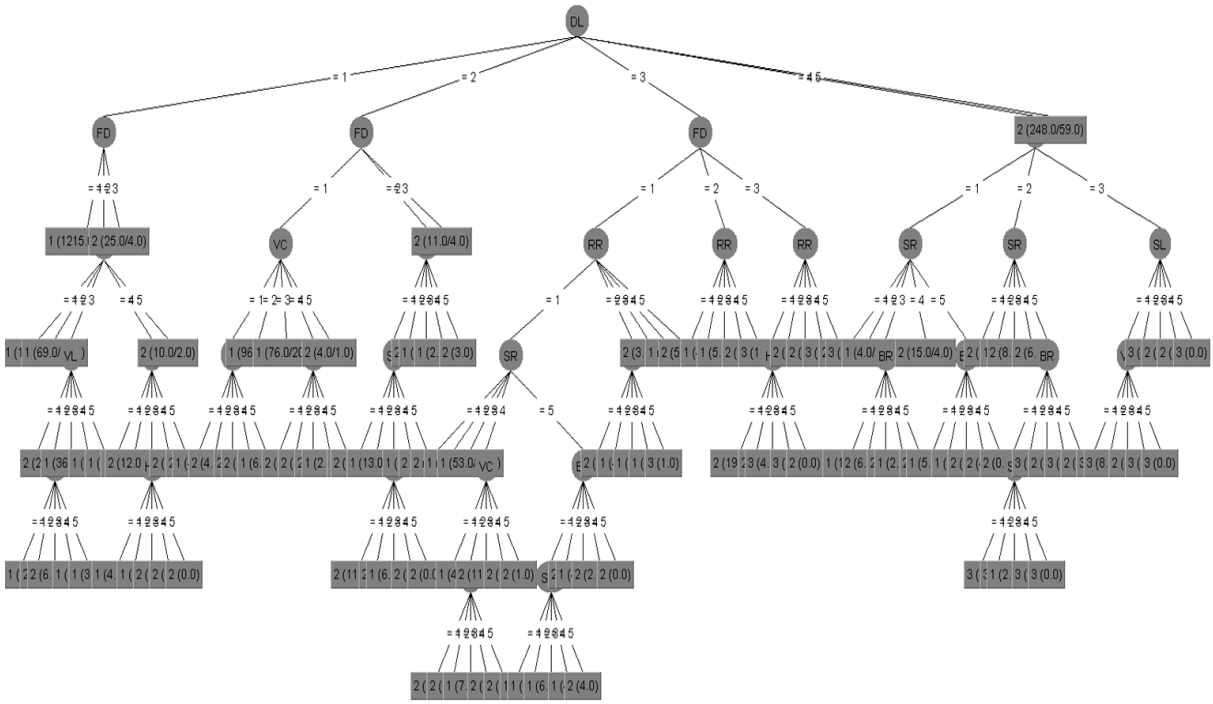

Figure 2: Decision tree J48 classification

Therefore, Rough set, classification was a better algorithm for classification in this dataset.

Table 9: Accuracy and computation time comparison

	A. Rough set + WEN threshold	B. Decision tree j48	Compare A to B.
Average time	3.70 time unit	3.77 time unit	1.89% faster
Correct classify (%)	85.91	78.29	8.87% more accuracy

4.Structural Equation Modelling (SEM)

All attributes were checked for their normal distribution. All attributes skewness value were laid in normal distribution criteria, skewness test (-1 < skewness < +1). After that, these attributes were calculated for their correlation with other attributes in order to form up the new latent variable or factor by using factor analysis method. The factor analysis calculation for new factors was successful since the KMO value was larger than 0.50 and

the Bartlett's test passed significance at α0.05. Variamax Rotation technique and Principle Component analysis were used to construct the new factors.

Table 10: Factor analysis statistics

KMO-Bartlett's test	
Kaiser-Meyer-Olkin measure of sampling adequacy	0.54
Bartlett's test of sphericity-sig	0

There were five factors. All attributes were still present. All attributes were composed to form five factors as shown in Table 11.

Table 11: Factors and their components

Component/Factor	Attribute	Factor loading
Driver	Sleepiness level – SL	0.765
	Visibility level -VL	0.743
	Vehicle condition - VC	0.355
Road	Broken road - BR	0.812
	Heavy rain - HR	0.799
Signal	Road signal - RS	0.774
	Street lighting - SLT	0.731
Obstacle	Side road burning - SB	0.786
	Road repairing - RR	0.654
Behaviour-driver	Fast driving - FD	0.808
	Drinking level - DL	0.717

Based on research hypothesis, all factors were assigned as the direct cause of factor "Damage magnitude level". Therefore, proposed SEM model was illustrated as shown in Figure 3.

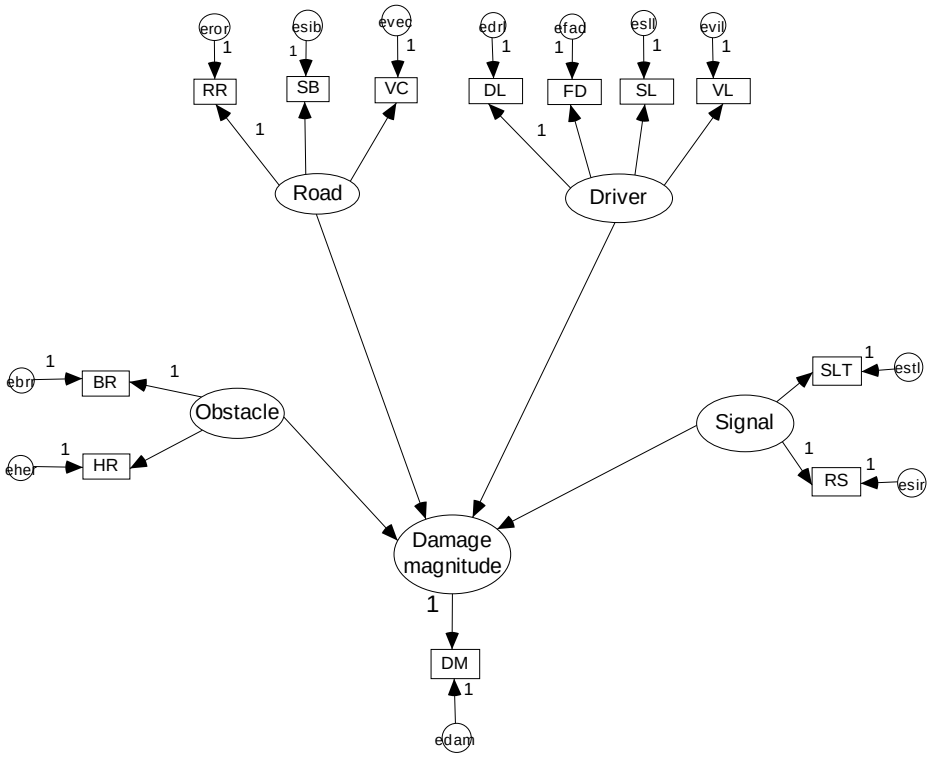

Figure 3: Proposed Structural Equation Modelling

This proposed SEM was adjusted their relationship to meet the model best fit criteria.

The saturated or best fitted model was shown in figure 3.

Table 12: Model fit statistics of Saturated Model

Indices	Value	Criteria
Chi-square	0.06	p-value > 0.05
AGFI	0.98	GFI > = 0.95
RMSEA	0.77	RMSEA < 0.08

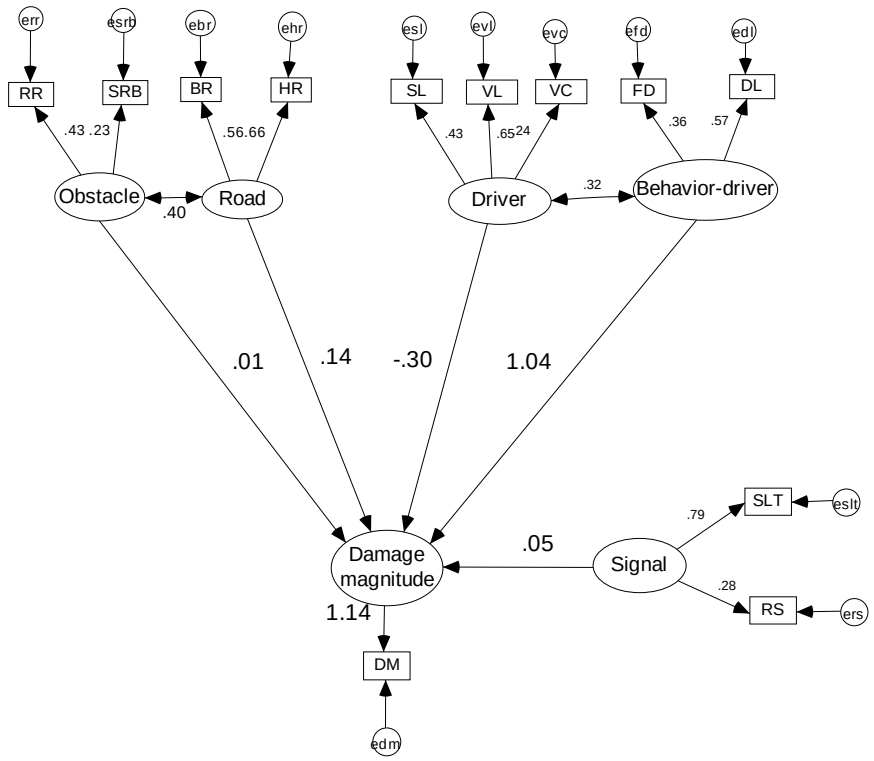

Figure 4: Saturated or fitted Structural Equation Model (standardised model)

The result of SEM presents that factor "Behaviour-driver has bi direction interaction with factor "Driver". Factor "Obstacle" has bi direction interaction with factor "Road". For example, Cause factor "Behaviour-driver" value increase should increase the effect on factor "Driver" (+0.32).Factor "Behaviour-driver" has a biggest effect to factor "Damage magnitude level".

5 Traffic accident evidence prevention application

In practical application development for prediction and prevention, road accidents occurring development, severe accident evidences were selected from Rough Set rules and SEM accidents damage prediction. There were many rules which were generated from Rough Set method thus structural equation model was considered together in order to choose the most flexible attribute value changing rule. For example, some severe damage magnitude 3 rule was collected from (3.3.1) as shown in Table 5.

(DL=4)&(FD=3)&(SL=1)&(VL=2)&(VC=3)&(SR=5)&(SLT=5)&(BR=1)&(HR=2)=>(class=3[1]).

Nine situations (attributes) were the cause of traffic accident damage level three.

The unity game engine program was used to construct multimedia about the driving scenarios. These scenarios were presented in two VDO clips. Their aim was vehicle driving about attention and inattention. The inattention VDO clip presented adjacent and sequence of related situations based on the level of each attribute in choosing rule. In severe accidental damage level 3, the story board assigned driver character level or acting as mentioned in the rule. Drinking level 4 (high), fast driving above 81 km/hr, Sleepiness level 1, Visibility level 2, Visibility capability level 3", Street light level 5, Broken road situation level 1 and Heavy rain level 2. Otherwise, in an attentional driving situation, the driver character of some important cause factors reduced their level.

Attribute value reduction, firstly considered on large amounts of effectiveness on damage magnitude and were viewed to reduce their value in level 1. For example, drinking level 4 was reduced to level 3. All of the new attribute values were sent to SEM in order to calculate for damage magnitude. If the damage magnitude value was still larger than 3 then new attributes, second effective, was considered to perform attribute level reduction in the same way. There are only two factors, driver and driver behaviour, which could manage a value level reduction since they were directly related to driver.

The reason the attribute value level reduction is 1 is because more comfortable managing by driver than leaving it out. They should have some attribute value reduction that could reliably reduce attribute damage magnitude from 3 to fine level 1.

This research chooses the situation that not too tightly forced the driver. Afterwards, these scenarios which were composed of possible actions by the driver were used as a storyboard in the VDO clip creation with Unity. The Figure 5 VDO clip presents some severe bicycle driving accident evidence. This VDO clip was firstly presented to experimental volunteers. The VDO clip presents the car accident causes which could affect accidents occurring. In VDO clip demonstration, VDO presenter and Audient volunteer discussed accident causes and prevention. Volunteers suggested how to reduce the level value of some attributes in order to escape a car accident. These new attribute values were sent to calculate the damage level by SEM. The previous VDO clip was shown again if the damage level is still on 2 or 3. If damage level value is about 1 then the second VDO clip, fine driving, was instead shown.

Table 13: Twenty VDO clips of traffic accident, conditional attributes and damage magnitude level 2 and 3.

#	Damage magnitude level	Rule
1	2	(DL=5)&(FD=1)&(SL=3)&(VL=2)&(VC=2)&(SR=3)&(SLT=1)&(BR=2)&(HR=1)=>(class=2)
2	2	(DL=5)&(FD=1)&(SL=4)&(VL=2)&(VC=3)&(SR=4)&(SLT=4)&(BR=3)&(HR=4)=>(class=2)
3	2	(DL=3)&(FD=1)&(SL=2)&(VL=1)&(VC=2)&(SR=4)&(SLT=3)&(BR=1)&(HR=1)=>(class=2)
4	2	(DL=4)&(FD=1)&(SL=1)&(VL=3)&(VC=3)&(SR=3)&(SLT=3)&(BR=2)&(HR=1)=>(class=2)
5	2	(DL=3)&(FD=3)&(SL=1)&(VL=2)&(VC=2)&(SR=5)&(SLT=5)&(BR=1)&(HR=1)=>(class=2)
6	2	(DL=2)&(FD=1)&(SL=2)&(VL=3)&(VC=1)&(SR=5)&(SLT=5)&(BR=1)&(HR=1)=>(class=2)
7	2	(DL=4)&(FD=1)&(SL=3)&(VL=3)&(VC=3)&(SR=5)&(SLT=5)&(BR=4)&(HR=1)=>(class=2)
8	2	(DL=3)&(FD=3)&(SL=1)&(VL=3)&(VC=3)&(SR=5)&(SLT=5)&(BR=1)&(HR=1)=>(class=2)
9	2	(DL=5)&(FD=3)&(SL=1)&(VL=3)&(VC=3)&(SR=3)&(SLT=4)&(BR=1)&(HR=1)=>(class=2)
10	2	(DL=4)&(FD=3)&(SL=1)&(VL=3)&(VC=2)&(SR=3)&(SLT=2)&(BR=2)&(HR=1)=>(class=2)
11	3	(DL=5)&(FD=1)&(SL=2)&(VL=2)&(VC=3)&(SR=2)&(SLT=5)&(BR=1)&(HR=1)=>(class=3)
12	3	(DL=4)&(FD=3)&(SL=1)&(VL=2)&(VC=3)&(SR=4)&(SLT=4)&(BR=1)&(HR=2)=>(class=3)
13	3	(DL=5)&(FD=2)&(SL=1)&(VL=2)&(VC=2)&(SR=5)&(SLT=5)&(BR=2)&(HR=3)=>(class=3)
14	3	(DL=3)&(FD=2)&(SL=1)&(VL=2)&(VC=3)&(SR=5)&(SLT=2)&(BR=2)&(HR=2)=>(class=3)
15	3	(DL=3)&(FD=3)&(SL=1)&(VL=2)&(VC=2)&(SR=2)&(SLT=3)&(BR=3)&(HR=1)=>(class=3)
16	3	(DL=5)&(FD=2)&(SL=1)&(VL=2)&(VC=2)&(SR=5)&(SLT=3)&(BR=4)&(HR=2)=>(class=3)
17	3	(DL=4)&(FD=3)&(SL=1)&(VL=2)&(VC=3)&(SR=4)&(SLT=3)&(BR=1)&(HR=4)=>(class=3)
18	3	(DL=4)&(FD=3)&(SL=1)&(VL=1)&(VC=2)&(SR=3)&(SLT=3)&(BR=1)&(HR=1)=>(class=3)
19	3	(DL=5)&(FD=2)&(SL=2)&(VL=2)&(VC=1)&(SR=3)&(SLT=4)&(BR=2)&(HR=2)=>(class=3)
20	3	(DL=5)&(FD=3)&(SL=1)&(VL=2)&(VC=2)&(SR=3)&(SLT=3)&(BR=1)&(HR=2)=>(class=3)

All of the 20 VDO clips were kept in the archive that everyone could visit and try out with data entry of traffic accidents occurring prevention application in order to adapt their behaviour.

Figure 5: Sample snap shot clip (inattention) on traffic, accidental prevention application based on some 3 damage magnitude level rule

6. Prediction and protection of car driving accident application design

With the mobile application usage the car drivers have to enter data about their destination, current data and behaviour. All driver's data, road situation and weather are sent to be calculated for accident damage occurring. The classification algorithm will present the point on the route that has the highest car accident damage magnitude. Since there are many points on a particular route, data about road (BR-RR-SLT-RS) or data about road and obstacle will be considered previously and given only at the road intersection, curve road, bridge, along a railway-canal-river road and main side street entrance.

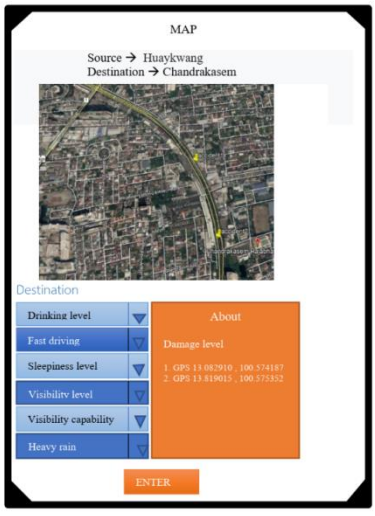

Figure 6: Mobile application – illustrate of map and dangerous point (damage magnitude level 2 or 3)

7. Application usage evaluation

One thousand volunteers were assigned to drive a car on predefine routes. The condition of road, obstacle and traffic signals of this route were already surveyed and kept their current value in the database. The volunteer driver has to enter their situation about behaviour (DL-SL-FD-VL) into the application input screen then the application predicts the level of accident damage magnitude. Car driving volunteers try to reduce some attribute values in order to reduce or escape the accident occurring. Under application program prediction and prevention suggestion, the car driving testing started. After a car driving test finished, the value of the accident damage magnitude level was inputted to the application. The testing was set up in two situations. Firstly, all volunteers were not joined to car driving accident prediction and prevention. All of them were assigned to drive their car in research's predefined routes. Secondly, all volunteers undertook the predicting and prevention program. They learnt how to use the application and adapt their behaviour in order to prevent themselves from road accidents. Afterwards, all of them were again assigned to drive their car on the research's predefined routes.

Research result

1. Car driving accident pattern classification method

In this research, Rough Set and Decision Tree Classification algorithm were used to predict the accident damage magnitude. The classification crossed validation on two thousand test cases. The accuracy of classification was presented in Table 13. Rough Set classification algorithm testing has more accurate than Decision Tree. However, Decision Tree algorithm classification was a little bit greater in class attribute retrieval time. This research choose Rough set classification since the Rough Set rule more accurate prediction and reference rule retrieval time could be reduced by weighted normalised Euclidean saturated weighting threshold.

2. Accidental occurring rate comparison

One hundred volunteer's car driving test database application usage evaluation were compared with the car driving accident damage magnitude mean of the pre-test and post-test by paired t-test (S. Massa, 2013).

Table 14: Damage magnitude level of the same observation before watching road accident cause and prevention VDO clip paired t-test (partial observation)

Damage magnitude level		
Observation	Pre-test	Post-test
1	1	1
2	2	1
3	2	1
.		
.		
.		
99	3	1
100	1	1
Mean	2.43	1.17
SD	0.77	0.46

Table 15: Paired t-test result at α0.05 significant test

t	8.38
df	99.00
sig(2 tail)	0.00

Discussion and further research

1. Discussion

This research training data was gathered from only six provinces in and five provinces around Bangkok, since this area has a very high density of cars travelling and road accident evidence. Road or traffic accident cases from other areas may not be matched to this research's insight road accident occurring rules. Nevertheless, data from new interest road accident area must be collected based on the same suite of prior conditional attributes. Moreover, the classification algorithm that was used to the predict decision variable may be tried out again in order to search for the most accurate classifier.

2. Further research

Since the road situations are dynamically changed, the data of some factors such as weather, road obstacle and road signals must be periodically updated as soon as possible. In order to increase most current road accident and road situation data, this application data entry shall be adapted so that registered members could directly enter some data to prediction and protection of car accident application (World road association, 2020).

There are some favourite social media, such as Twitter - Line, and many car drivers often use them for public data broadcasting (Michael K., 2020). Text mining from these kinds of social media could rapidly increase timely road or traffic accident information (Nayak R., 2010).

REFERENCES

Thailand parliamentarian seminar. (2019). The situation of road safety at global and local levels and the urgent need to reduce road traffic deaths and injuries in Thailand, World health organization, Thailand.

Technology Information Department.(2019). Road accidental statistics, Thai Royal Safety Culture, Thailand.

JiaweiHan.(2012). Data Mining: Concepts and Techniques, A volume in The Morgan Kaufmann Series in Data Management Systems 3rd Edition.

Saed Sayad.(2020). An Introduction to Data Science. Department of Computer Science, Rutgers, The State University of New Jersey, USA.

NarasimhaPrasad. (2013). Gain Ratio as Attribute Selection Measure in Elegant Decision Tree to Predict Precipitation, 8th EUROSIM Congress on Modelling and Simulation.

Statistical engineering division. (2017). EUCLIDEAN DISTANCE, National institute of standard and technology: Nist, USA.

Michael Greeacre.(2008). Correspondence Analysis and Related Methods, UniversitatPompeuFabra, Barcelona.

Diana Suhr.(2020).The Basics of Structural Equation Modeling, University of Northern Colorado, USA.

Leonardo Miljko. (2017). Exploratory Factor Analysis - KMO and Bartlett's Test, Bosnia and Herzegovina.

Cornell statistical consultant.(2020). Fit Indices commonly reported for CFA and SEM, Cornell University, USA,.

Will Monroe.(2017). Covariance and Correlation, Stanford University USA.

Xi Jianfeng et al. (2019).A classification and recognition model for the severity of road traffic accident, Advances in Mechanical Engineering, Vol. 11(5) 1–8.

Jonathan J. Rolisona.(2018).What are the factors that contribute to road accidents? An assessment of law enforcement views, ordinary drivers' opinions, and road accident records, Department of Psychology, University of Essex, UK.

Koen Vanhoof.(2003).Profiling High Frequency Accident Locations Using Association Rules, proceedings of the 82nd Annual Transportation Research Board, Washington DC. (USA).

Sachin Kumar.(2015). A data mining framework to analyze road accident data, Kumar and Toshniwal Journal of Big Data.

Gang Tao.(2015).Cause Analysis of Traffic Accidents Based on Degrees of Attribute Importance of Rough Set, IEEE 12th International Conference on Ubiquitous Intelligence and Computing.

Yong-Seok Kim.(2003).Effects of Driver, Vehicle, and Environment Characteristics on Collision Warning System Design, Linköping Institute of Technology, Sweden.

Adlih. Al.(2003).Role of Gender in Road Accidents, Civil Engineering Department, Jordan University, Amman, Jordan.

Remco R. Bouckaert.(2013).Weka manual 3-7-8, Waikato university, New Zealand.

S. Massa.(2013). t-Test, University of Oxford, UK.

World road association (2020) Establishing and maintaining crash data system, PIARC, available online at URL: https://roadsafety.piarc.org/en/introduction on 31 August.

Michael K. Hepworth.2020. Posting About Your Car Accident On Social Media, Hepworth &Associates, USA.

Nayak R., Piyatrapoomi N., Weligamage J. (2010) Application of text mining in analysing road crashes for road asset management. In: Kiritsis D., Emmanouilidis C., Koronios A., Mathew J. (eds) Engineering Asset Lifecycle Management. Springer, London.