World Academy of Science, Engineering and Technology
International Journal of Computer and Systems Engineering
Vol:17, No:1, 2023

Combining the Deep Neural Network with the K-Means for Traffic Accident Prediction

Celso L. Fernando, Toshio Yoshii, Takahiro Tsubota

***Abstract*—**Understanding the causes of a road accident and predicting their occurrence is key to prevent deaths and serious injuries from road accident events. Traditional statistical methods such as the Poisson and the Logistics regressions have been used to find the association of the traffic environmental factors with the accident occurred; recently, an artificial neural network, ANN, a computational technique that learns from historical data to make a more accurate prediction, has emerged. Although the ability to make accurate predictions, the ANN has difficulty dealing with highly unbalanced attribute patterns distribution in the training dataset; in such circumstances, the ANN treats the minority group as noise. However, in the real world data, the minority group is often the group of interest; e.g., in the road traffic accident data, the events of the accident are the group of interest. This study proposes a combination of the k-means with the ANN to improve the predictive ability of the neural network model by alleviating the effect of the unbalanced distribution of the attribute patterns in the training dataset. The results show that the proposed method improves the ability of the neural network to make a prediction on a highly unbalanced distributed attribute patterns dataset; however, on an even distributed attribute patterns dataset, the proposed method performs almost like a standard neural network.

***Keywords*—**Accident risks estimation, artificial neural network, deep learning, K-mean, road safety.

I. Introduction

A good transport system is core for the development of any country. However, while on the one hand, transport and road infrastructure are seen as determining factors for economic development [1]-[3], on the other hand, road safety is a major concern worldwide. The World Health Organization estimates that 1.35 million people die each year due to traffic accidents, and the economic cost of the road accident is approximately 3% of the gross domestic product, GDP, for most countries [40].

One way to reduce road accidents is to understand their causes and to be able to predict its occurrence so that advanced measures such as information provision, traffic control and speed control can be taken. Much has been done in this respect; especially, understanding the association between the traffic accident occurrences with the traffic environment factors has been studied applying some traditional statistical methods. Reference [4] investigated the effect of three different pavement types on the accident risk; [5] relates the roadway characteristics with the fatal accident on the rural segments of the highway in Oklahoma; [6] explored the impact of traffic congestion on the frequency of road accident on M25 London orbital motorway; [7] explored the relationship between the age of the surface of the porous asphalt pavement with the accident occurrence. All these authors [4]-[7] applied the Poisson/family of the Poisson regression on their studies. Other authors applying the Logistic regression investigate: the relationship between the type of the road accident with the potentiality of the serious injury or death [8]-[11]; the accident hotspot in the road network [12]-[14]; [15] associates the accident risk with three main causes, driver behavior, vehicle mechanical issues, and the road traffic environment; [16] relates the accident risk with factors related with age and gender, speed, traffic control type, time of day.

Traditional statistical methods are the state-of-the-art for accident risk assessment [17]. Although the traditional statistical methods are used to assess the accident frequency and/or their association with traffic environmental factors, these methods suffer from some problems. For example, when applying the Poisson model, if the mean and variance of the samples in the data are not equal, the test statistics derived from the model will not be correct due to the biased standard errors estimated by the maximum likelihood method [18]-[21]. On the other hand, the Logistic regression has been used for classification problem, where the analyst focuses to identify patterns in the attribute factors that lead to the occurrence of the accidents, if the association between the attributes factors and the target response results from the interaction effect among the attribute factors, the Logistic regression, itself, cannot map these interaction effects among the factors [22], [23], it will depend, highly, on the analyst skills to identify such interaction effect and adjust the model accordingly.

Recently, the ANN, an emerging data mining technique is being used for classification and prediction purpose. Moved mainly by the availability of the large amount of data, and data from different sources, ANN is becoming an alternative or complement to the traditional statistical data analysis technique [24], [25]. According to [22], the self-adaptability – ability to adjust the model to any data; the universal approximation function – ability to capture very complex relationship in data; and the nonlinearity modeling – ability to model data from real-world relationship, are some of the important advantages from the ANN over the traditional statistical methods.

Studies such as [26]-[28] comparing the performance of the ANN with some traditional statistical methods found that the ANN has shown a better result.

Despite ANN being known for their ability to map complex patterns in data, they, also, have some difficulties in:

C. L. Fernando was with Ehime University, Matsuyama, Japan, He is now with Lurio University, Nampula, Mozambique (phone: +258 87 0066777; e-mail: cfernando@unilurio.ac.mz).

T. Yoshii and T. Tsubota are with Ehime University, Matsuyama, Japan (e-mail: yoshii@cee.ehime-u.ac.jp, tsubota.takahiro.jl@ehime-u.ac.jp).

- Predicting beyond the range of data used to train the model. It is because, the model can only generalize within the domain of the training data [29], and
- Making classification on unbalanced data. For example, in a binary classification when one class represents the majority on the data samples, the model is more likely to be biased toward the majority class, showing poor classification on the minority class [30], [31].

In general, in the real-world data, the minority class in the database represents the class of interest. An example is the traffic accident data where the accident events (the class of interest) correspond to no more than 1% of the samples, and non-accident events about 99%.

The learning process of the ANN assumes that the classes in the data are distributed in an even manner, and the misclassification costs of any class are also equally weighted [32]; thus, in circumstances such as the traffic accident data are used, the evaluation metric can be compromised [30], [33].

This study focuses on improving the predictive ability of the ANN when dealing with highly unbalanced data as it is the traffic accident data. To improve the performance of the ANN, various techniques have been proposed; some consist of the combination between the ANN models, known as ensemble approach [38]; this consists of creating different sets of the neural network to perform the same task in order to solve the problem in hand, and the output from each set are combined to generate the final result. The other is a combination between the ANN model with other techniques, an example is the SCDNN method which combines the spectral clustering with deep neural network [39], known as modular approach; this consists of decomposing the problem in hand into several modular tasks, and for each modular task an appropriate ensemble member is applied to execute the subtask [34]. The modular approach explores the ability of each ensemble member to execute the assigned activity. Although these two approaches have been implemented, still much has to be done.

The ensemble approach, which creates sets of the neural network models still uses the whole training data for each model set; therefore, each set will suffer from the unbalanced samples distribution in the data effect when dealing with such data.

In order to improve the predictive ability of the traffic accident on a highly unbalanced data, this study proposes a combination of the k-means clustering with the neural network (deep neural network). The proposed method applies the k-means to cluster the samples in the training data, and for each group of clustered dataset a neural network is developed. The expectation of the proposed method is that each neural network will be trained with a specific set of data in which the sample distribution problem will be minimized; therefore, each set of the neural network will have better generalization.

This paper is organized into six sections. The second section describes proposed methodology and the evaluation criteria used in this study. The third section describes the data processing method. The fourth and fifth sections present the study, showing the result and discussions. Finally, the sixth section summarizes the finding and makes the conclusion.

II. Methodology

A. Unbalanced Data Definition

Unbalanced data can be distinguished in two distinct ways. One way is concerned about the outputs (target response) distribution. Assuming a classification problem, it is said that a database is unbalanced when the classes in the target response are not evenly distributed. Another way to look at unbalanced data is to focus on the distribution of the attributes pattern in the input data [35].

This study focuses on the unbalanced of the attribute pattern distribution.

According to [33], [35], the training process of the neural network is a mapping of the input vector patterns to the output data; therefore, modeling error may occur if for the same output there are a huge variation of the input vector patterns associated. If some of the input patterns are small in number, the modeling tends to ignore the minority group, treating them as noise.

B. Proposed Method

Based on the assumption that the traffic data may consist of different attributes patterns of unbalanced distribution, this study proposes:

1. The samples from the data are grouped in small sets, applying the K-mean cluster method; these sets of data are called *Clustered Data*.
2. Each Clustered Data are used to train a specific Neural Network; and each of them are called *Clustered Neural Network*.
3. Each clustered neural network is tested with unseen samples from the same domain as its training data.
4. The results of each clustered neural network are combined.

To group the samples from the attribute (input) in small sets, the k-means algorithm is applied. The k-means can map patterns from the data and group them according to their similarity. The similarity is defined based each sample's proximity to a reference point - center of each cluster group – based on (1); and the number of sets to be created by the k-mean is defined based on the Elbow method.

Let C_k be the center of a cluster group; d_{pk} is the distance from the sample p^{th} in the database to the center of group *k*.

$$d_{pk} = \sqrt{\sum_{i=1}^{N} \left(x_{ip} - \bar{x}_{ik} \right)^2} \tag{1}$$

The vector $x_{ip} \in C_j$ if $d_{pj} < d_{pk}$ for $k \neq j$ where $\bar{x}_{ik}$ is the mean of the explanatory variable *i* within the *k* group; x_{ip} is the sample p^{th} in the explanatory variable *i*.

To develop the network system, a deep neural network is used to train each set of the clustered data (clustered neural network). Each clustered neural network is tested with filtered samples; the filtering process of the test data consists of assign each sample from the test data to the appropriate clustered neural network so that the test subset match the domain of the

data subset used to train each clustered neural network model. This is done by calculating the Euclidean distance of each sample from the test data to the centers of each subset from the training data defined with the k-mean, as in (2):

$$d_{pk\in tst} = \sqrt{\sum_{i=1}^{N} \left(x_{ip\in tst} - \bar{x}_{ik} \right)^2} \quad (2)$$

where $x_{ip\in tst}$ is the p^{th} instance from the test dataset, tst in the explanatory variable i; $d_{pk\in tst}$ is the distance from the p^{th} instance to the center of the cluster group C_k defined by the k-mean.

The outputs of each clustered model are combined and the performance of the network system is evaluated. Fig. 1 describes the whole process of the proposed method, starting from clustering up to the combination of the outputs.

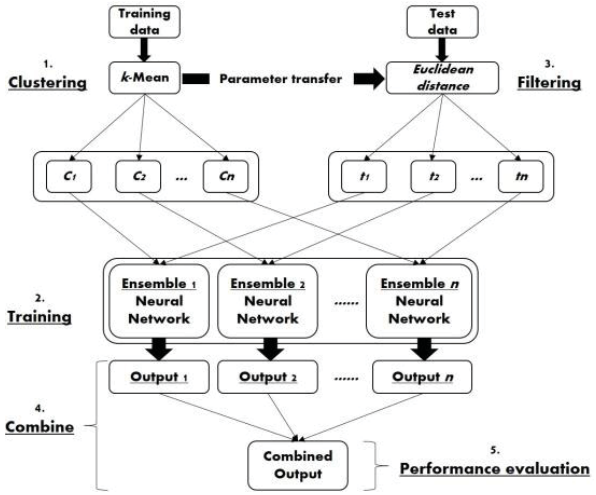

Fig. 1 Clustered Neural Network, flowchart of the proposed method

C. Neural Network Definition and Characteristics

The topology of the neural network (deep neural network) is defined based on trial-and-error process. The one which shows the best result is selected. Since the goal is to predict the accident risk events, the accidents are treated as binary data, therefore, the loss function is the binary cross-entropy, and the activation functions are ReLU and Sigmoid for the hidden layers and the output layer, respectively. The model uses the early stopping to avoid the overfitting and allows the model to have good generalization.

D. Model Performance Evaluation Metrics

The model performance is evaluated based on two metrics:

- Receiver Operator Characteristic Area Under the Curve (ROC AUC), which describes the ability of the model to distinguish two different conditions, in this case accident risk-prone traffic condition from safe traffic condition.
- Precision Recall Area Under the Curve (PRAUC), which describes the ability to predict the phenomenon of interest, in this case the traffic accident events.

III. Data Description and Processing

A. Data Description

The data were collected from 2009 to 2018 in two expressways in Japan; the Toumei Expressway and Syutokou Expressway. The data were collected every 5 minutes at the segment level of each section in the respective expressway. An example of road section is shown in Fig. 16, and the details are summarized in Table IV, see in the Appendix; they refer only to Toumei Expressway. The data consist of traffic volume, travelling speed, occupancy, and the accident events.

Considering the time interval of 5 minutes, the traffic volume is defined as the number of vehicles passing through the sensor point on the road segment. Travelling speed is the average speed of all vehicles passing through a sensor point within the time interval. Occupancy is the proportion of time that the sensor is occupied by a vehicle in the time interval. The response variable, accident events, was collected as an aggregate number of all accident events that occurred within the section in the time resolution.

B. Data Processing

Each expressway was processed separately. It means, two different databases are available for this study. However, they consist of same variables and same data collection method.

While the explanatory variables were collected at the segment level, the response variable was collected at the section level. Therefore, to allow all variables (explanatory and the response) in the same space resolution, each explanatory variable was averaged in the section level based on (3), see also Fig. 2, resulting in the averaged traffic volume, averaged traveling speed, and averaged occupancy in the section within the 5-minute time interval.

$$B_t^{avg} = \frac{1}{j} \sum_{j=1}^{J} b_{j,t} \quad (3)$$

where, $b_{j,t}$ is the explanatory variable at the segment level j and time interval t; B_t^{avg} is the average of the explanatory variable in the section within the time interval t.

Since the averaging process of the explanatory variables leads to a loss of information in the explanatory variables, to alleviate this loss, the standard deviation of each explanatory variable within the section was calculated based on (4):

$$B_t^{Std} = \sqrt{\frac{1}{j} \sum_{j}^{J} (b_{j.t} - B_t^{avg})^2} \quad (4)$$

The response variable (traffic accident) is treated as binary data, taking 1 if at least one accident was observed in the section within the time interval, zero otherwise.

C. Data Split for Neural Network Analysis

To perform the analysis, each database was divided into two sets:

- *Training dataset* consists of observations from day 1 to day 20 of each month within the 10 years (2009 to 2018).
- *Test dataset* consists of observations from day 21 to 31.

Table I shows the size of the dataset after splitting, only for Toumei Expressway.

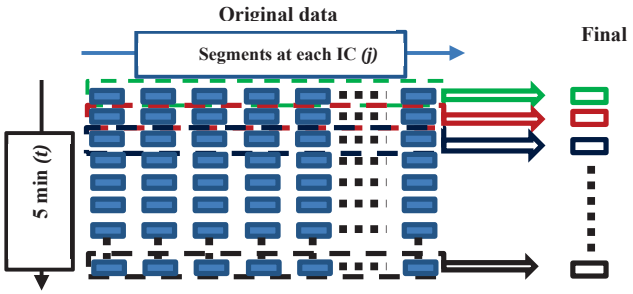

Fig. 2 Attributes transformation from the segment resolution to the section resolution

TABLE I
TOUMEI EXPRESSWAY TRAINING AND TEST DATA SET

Data Division	Data Volume	Accident (Binary)	
		Number	Ratio (%)
Training dataset	2,511,094	4,073	0.16
Test dataset	1,315,745	2,322	0.18
TOTAL (Data)	3,826,839	6,395	0.17

IV. K-MEANS CLUSTERING

A. Number of Clusters Groups: Attribute Patterns in the Training Dataset

Cluster analysis is an unsupervised technique used to group samples based on their intrinsic characteristic. The data clustering principal is to maximize intra-class similarity and minimize inter-class similarity [36]. The optimal number of clustered groups to be generated is defined based on the Elbow method, which consists of the evaluation of the decreasing monotonic curve of dispersion within the cluster. The more the number of groups, the smaller the dispersion; however, the point beyond which the increases in number of the groups implies a smaller reduction in the dispersion is the optimal number of cluster groups [37]. For each training dataset, the Elbow method was applied and three group - *attribute patterns* - were identified, as shown in Fig. 3.

B. Characteristics and Distribution of the Attribute Patterns: Toumei Expressway

Table II and Fig. 4 show an unbalanced distribution of the patterns in the attribute variable of the Toumei Expressway. And the minority group "Pattern C2" accounts for the large number of all accident events.

C. Characteristics and Distribution of the Attribute Patterns: Syutokou Expressway

Different than the Toumei Expressway, in Syutokou Expressway, the attribute patterns are evenly distributed, as in Table III and Fig. 5.

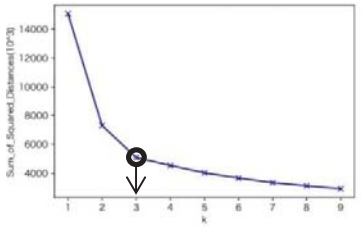

Fig. 3 Number of clustered groups defined based on Elbow method

TABLE II
ATTRIBUTE PATTERNS DISTRIBUTION IN TOUMEI TRAINING DATASET

Models (Patterns)	Sample volume		Target (Accident binary)
	Volume	Ratio	
Pattern C1	866,042	0.345	837
Pattern C2	226,984	0.090	2,279
Pattern C3	1,418,068	0.565	957
Full data	2,511,094	1	4,073

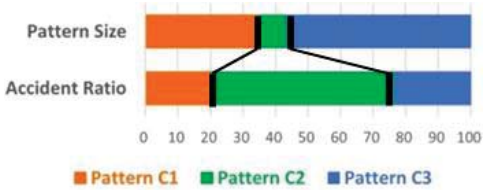

Fig. 4 Attributes pattern distribution in the training dataset

TABLE III
ATTRIBUTE PATTERNS DISTRIBUTION IN SYUTOKOU TRAINING DATASET

Models (Patterns)	Sample		Target
	Volume	Ratio	
Pattern dt-0	226,136	0.327	996
Pattern dt-1	221,191	0.320	2,322
Pattern dt-2	243,873	0.353	896
Full data	691,200	1	4,214

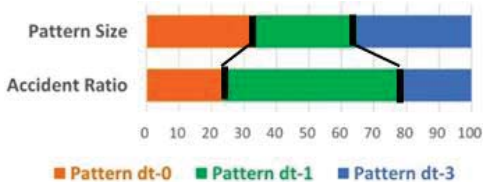

Fig. 5 Attributes pattern distribution in the training dataset

D. Attribute Patterns vs. Traffic Stream Relationship

An important aspect that can be observed is that the groups of the attribute patterns generated by the k-means match the traffic stream characteristics (traffic condition), see Figs. 6 and 7. The Pattern C3 and Pattern dt-2 correspond to the free-flow traffic condition; Pattern C2 and Pattern dt-1 correspond to the congested condition; and Pattern C1 and Pattern dt-0 correspond to transition from the free-flow to congested condition, respectively.

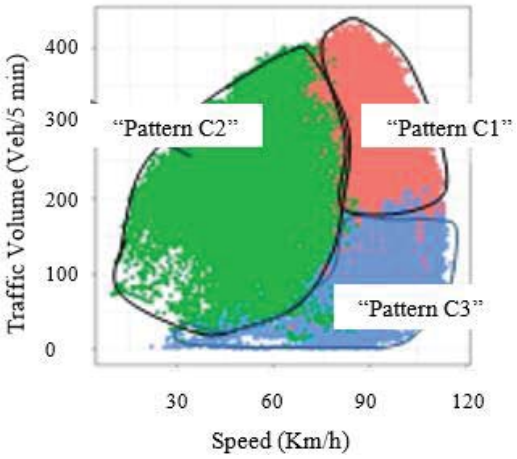

Fig. 6 Relationship between the attributes pattern distribution and the traffic stream condition – Toumei Expressway

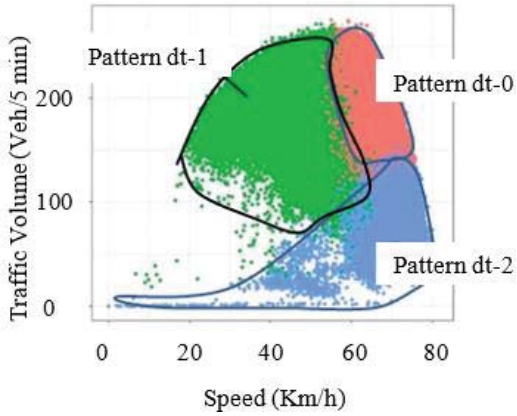

Fig. 7 Relationship between the attributes pattern distribution and the traffic condition – Syutokou Expressway

V. Traffic Accident Prediction Model Development

A. Neural Network Model Architecture

To develop the traffic accident prediction model, ensemble neural networks were developed for each clustered data, and the outputs from the ensemble neural networks were combined.

The neural network used for modeling Toumei Expressway data is a deep neural network (deep learning) with 6-16-6-1 nodes in the input layer, first and second hidden layers, and output layer, respectively. The model was trained with 100 epochs, the early stopping is considered to terminate the training process when the overfitting starts to occur.

For the second experiment, Syutokou Expressway traffic data, the topology of neural network consists of 6-11-6-1 nodes, in the input layer, first and second layers, and output layer, respectively. The model was trained with 100 epochs, the early stopping is considered to terminate the training process when the overfitting starts to occur.

B. Results

- Toumei Expressway Results of Each Clustered Data

The clustered neural network trained with data set "Pattern C1" – which correspond to Free-Flow traffic condition, showed the best performance when the ROC-AUC is the evaluation metric, see Fig. 8. On the other hand, the ability to predict the accident events is higher on the ensemble member trained with the "Pattern C2" – the Congested traffic condition, see Fig. 9

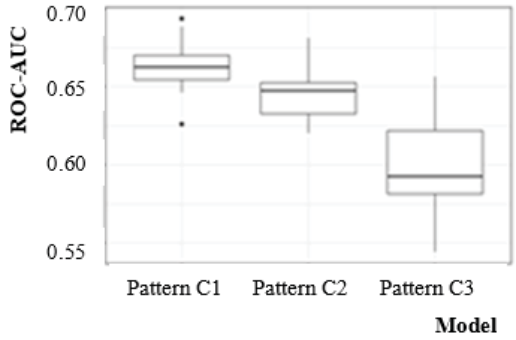

Fig. 8 Ability to distinguish between the accident-prone traffic conditions from safe traffic condition (ROCAUC) for each clustered dataset (ensemble member)

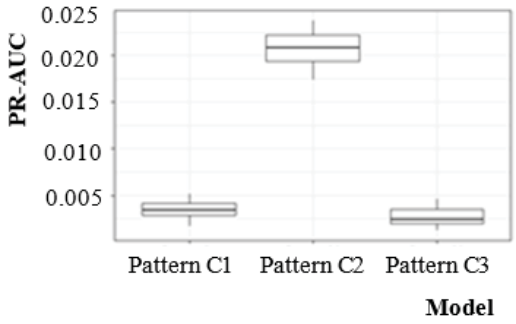

Fig. 9 Ability to predict accident occurrence for each clustered dataset (ensemble member)

- Toumei Expressway Result of the Combined Output

Combining the output of each ensemble neural network, the result is compared with the standard neural network model. The standard neural network model consists of training the neural network with "Full data", it is, without clustering.

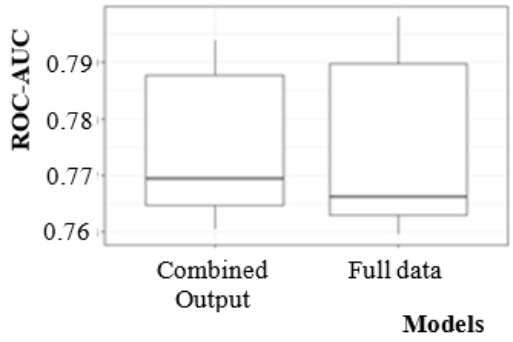

Fig. 10 Comparison between the proposed method and the standard method on the ability to distinguish between the accident-prone traffic conditions from safe traffic condition

The result of the comparison shows: The ability of the both methods, in regard to distinguishing between accident-prone traffic conditions from the safe traffic conditions (ROCAUC),

is comparable; no notable difference can be observed between them, see Fig. 10. On the other hand, the proposed method, combined output of the ensemble neural networks, outperforms the standard neural network with respect to the ability to predict the accident (PRAUC), as in Fig. 11.

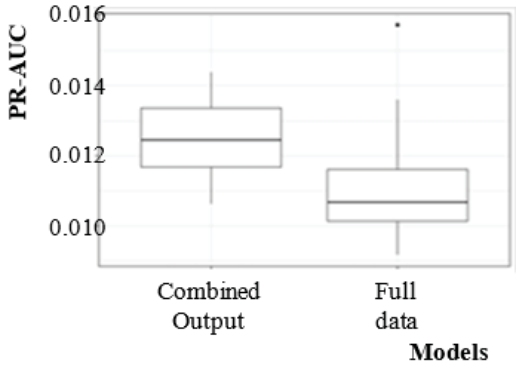

Fig. 11 Comparison between the proposed method and the standard method on the ability to predict accident occurrence

- Syutokou Expressway Results of Each Clustered Data

The result of each ensemble neural network shows that, with respect to ROCAUC, the ensemble member trained with "Pattern dt-0" – free-flow traffic – had better performance, see Fig. 12. On the other hand, with respect to PRAUC, the ensemble member trained with "Pattern dt-1", congested traffic condition, had better performance compared to others ensemble neural networks, as shown in Fig. 13.

- Syutokou Expressway Result of the Combined Output

Comparing the combined output method with the standard method, the results from both metrics, ROCAUC and PRAUC, show almost the same performance, see Figs. 13 and 14

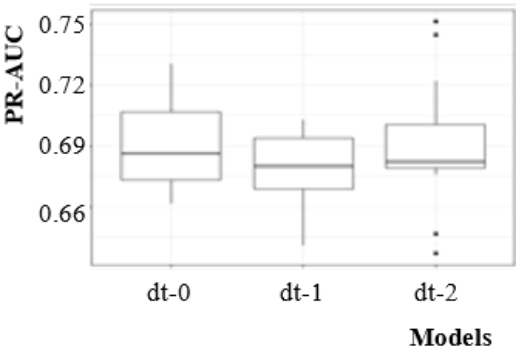

Fig. 12 Ability to distinguish between the accident-prone traffic conditions from safe traffic condition from each clustered dataset

C. Discussion of the Results

When the performance of the model is evaluated based on ROC AUC, the results show that there is no considerable difference between the proposed method (combined output from the ensemble neural networks) and the standard method (neural network trained with the full data).

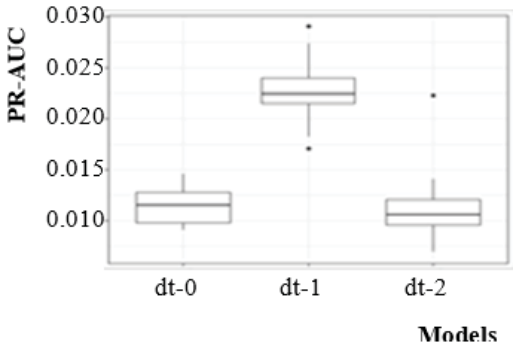

Fig. 13 Ability to predict accident occurrence from each clustered dataset

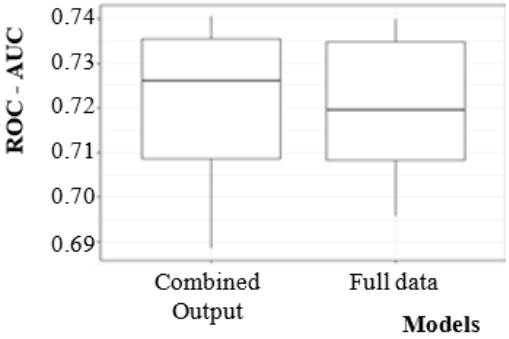

Fig. 14 Comparison between the proposed method and the standard method on the ability to distinguish between the accident-prone traffic conditions from safe traffic condition (ROCAUC)

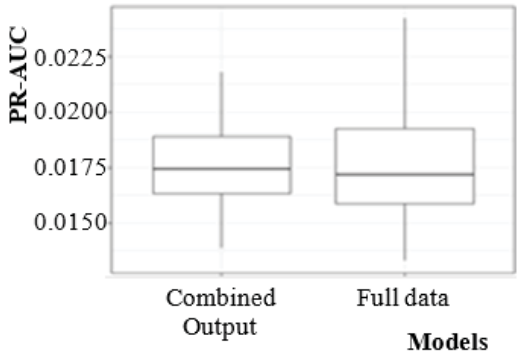

Fig. 15 Comparison between the proposed method and the standard method on the ability to predict accident occurrence (PRAUC)

When the performance is evaluated based on PRAUC, the results show that the proposed method outperforms the standard method. This suggests that, the proposed method can correctly predict a large number of accident events than it can be done applying the standard method, as in Fig. 11. The proposed method (ensemble neural networks) has the advantage to develop different sets of the neural models, each for an appropriate data domain (group of attribute patterns); this allows each ensemble member to minimize the generalization error caused by the unbalanced distribution of the input patterns attributes which a unique neural network would face if the whole training dataset was used to train a model. However, Fig. 15 shows a comparable performance between the proposed

method and the standard method, with respect to the ability to predict accident occurrence, PRAUC, this result is contrary to the one on Fig. 11; the reason of this is because the input patterns attributes (dt-0, dt-1 and dt-2) in training dataset of Syutokou Expressway are evenly distributed, see Fig. 5; therefore, since there is no attribute pattern which is dominant in the whole training dataset, the samples from each pattern are treated equally even when the standard neural network is used to develop the model; thus, the generalization error caused due to the unbalanced distribution of the input patterns is minimum. This fact makes both methods' (combined output of the ensemble neural networks and standard neural network) performance almost the same.

VI. Conclusion

This study argues that unbalance distribution in the input of the attribute patterns in the training dataset hinders the neural network ability to make accurate predictions. To improve the predictive ability of the model, a combination of k-means with the neural network has been proposed. The k-means is applied to cluster the training data into small training sets. Each set corresponds to a specific traffic stream characteristic. For each training set a neural network is developed; the study's major findings are:

- The ability of the neural network to distinguish between the accident-prone traffic condition and safe traffic condition increases when the traffic stream tends toward road capacity (e.g., Pattern C1 for Toumei Expressway and Pattern dt-0 for Syutokou Expressway); however, under the free-flow traffic condition (e.g., Pattern C3 for Toumei Expressway and Pattern dt-2 for Syutokou Expressway), this ability deteriorates.
- The ability to predict the accident events increases when the traffic stream tends toward congested condition, and is lower when the traffic stream is under the free-flow condition.
- The combined output of the ensemble neural networks improves the accident occurrence predictive ability of the neural model only when the attribute patterns in the training dataset are highly unbalanced, otherwise, the combined output of the ensemble neural network will have almost the same performance as the standard neural network.

Appendix

TABLE IV
Summary of the Toumei Expressway

ID	Sections	Length (km)	Number of segments in the section
①	Tokyo IC—Yokohama IC	19.7	12
②	Yokohama IC－ Atsugi IC	15.3	8
③	Atsugi IC－Ooimatsuda IC	22.9	12
④	Ooimatsuda IC－Gotenba IC	25.8	12

Fig. 16 Section in Toumei Expressway, Tokyo – Japan

Acknowledgment

Celso L. Fernando thanks to Architecture and Land Planning Faculty from Lurio University (FAPF-UniLurio) for the financial support to submit this paper for conference and publication.

References

[1] W. Owen, "Transportation and Economic Development," Proc. Seventy-first Annu. Meet. Am. Econ., vol. 49, no. 2, pp. 179–187, 1959.
[2] E. Sorupia, "Rethinking the role of transportation in tourism," East. Asia Soc. Transp. Stud., vol. 5, pp. 1767–1777, 2005.
[3] J. Khadaroo and B. Seetanah, "The role of transport infrastructure in FDI evidence from Africa using gmm estimates," J. Transp. Econ. Policy, vol. 43, no. 3, pp. 365–384, 2009.
[4] T. Tsubota, C. Fernando, T. Yoshii, and H. Shirayanagi, "Effect of Road Pavement Types and Ages on Traffic Accident Risks Effect of Road Pavement Types and Ages on a Traffic Accident Risks," Transp. Res. Procedia, vol. 34, pp. 211–218, 2018, doi: 10.1016/j.trpro.2018.11.034.
[5] J. C. Comer, N. J. Rose, and L. S. Bombom, "Poisson regression analysis of highway fatality accident data in Oklahoma," Int. J. Appl. Geospatial Res., vol. 5, no. 4, pp. 72–86, 2014, doi: 10.4018/ijagr.2014100105.
[6] C. Wang, M. A. Quddus, and S. G. Ison, "Impact of traffic congestion on road accidents: A spatial analysis of the M25 motorway in England," Accid. Anal. Prev., vol. 41, no. 4, pp. 798–808, 2009, doi: 10.1016/j.aap.2009.04.002.
[7] C. Fernando, T. Yoshii, T. Tsubota, and H. Shirayanagi, "Analysis of the Safety Performance of Drainage Pavement focusing on Pavement Age," East. Asia Soc. Transp. Stud., vol. 13, pp. 2016–2026, 2019, [Online]. Available: https://doi.org/10.11175/easts.13.2016
[8] S. Dissanayake and J. Lu, "Analysis of Severity of Young Driver Crashes," Transp. Res. Rec. 1784, no. 02, pp. 108–114.
[9] A. S. Al-ghamdi, "Using logistic regression to estimate the influence of accident factors on accident severity," Accid. Anal. Prev., vol. 34, pp. 729–741, 2002.
[10] E. T. Donnell and J. M. Mason, "Predicting the Severity of Median-Related Crashes in Pennsylvania by Using Logistic Regression," Transp. Res. Rec. 1784, no. 1897, pp. 55–63, 2004.
[11] P. Taylor, S. Y. Sohn, and H. Shin, "Pattern recognition for road traYc accident severity in Korea," Ergonomics, no. April 2013, pp. 37–41, 2010.
[12] M. M. Ahmed, M. Abdel-Aty, J. Lee, and R. Yu, "Real-time assessment of fog-related crashes using airport weather data: A feasibility analysis," Accid. Anal. Prev., vol. 72, pp. 309–317, 2014, doi: 10.1016/j.aap.2014.07.004.
[13] L. Tao, D. Zhu, L. Yan, and P. Zhang, "The traffic accident hotspot prediction: Based on the logistic regression method," ICTIS 2015 - 3rd Int. Conf. Transp. Inf. Safety, Proc., pp. 107–110, 2015, doi: 10.1109/ICTIS.2015.7232194.
[14] M. B. Ulak, A. Kocatepe, E. E. Ozguven, M. W. Horner, and L. Spainhour, "Geographic information system–based spatial and statistical analysis of severe crash hotspot accessibility to hospitals," Transp. Res. Rec., vol. 2635, no. 1, pp. 90–97, 2017, doi: 10.3141/2635-11.
[15] M. Híjar, C. Carrillo, M. Flores, R. Anaya, and V. Lopez, "Risk factors in highway traffic accidents: A case control study," Accid. Anal. Prev., vol. 32, no. 5, pp. 703–709, 2000, doi: 10.1016/S0001-4575(99)00116-5.
[16] [16] H. Chen, L. Cao, and D. B. Logan, "Analysis of Risk Factors Affecting the Severity of Intersection Crashes by Logistic Regression,"

Traffic Inj. Prev., vol. 13, no. 3, pp. 300–307, 2012, doi: 10.1080/15389588.2011.653841.
[17] F. Crocco, S. De Marco, and D. W. E. Mongelli, "An integrated approach for studying the safety of road networks: Logistic regression models between traffic accident occurrence and behavioural, environmental and infrastructure parameters," WIT Trans. Ecol. Environ., vol. 142, pp. 525–536, 2010, doi: 10.2495/SW100481.
[18] B. Debrabant, U. Halekoh, W. H. Bonat, D. L. Hansen, J. Hjelmborg, and J. Lauritsen, "Identifying traffic accident black spots with Poisson-Tweedie models," Accid. Anal. Prev., vol. 111, no. November 2017, pp. 147–154, 2018, doi: 10.1016/j.aap.2017.11.021.
[19] D. Lord and F. Mannering, "The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives," Transp. Res. Part A Policy Pract., vol. 44, no. 5, pp. 291–305, 2010, doi: 10.1016/j.tra.2010.02.001.
[20] D. Lord, S. R. Geedipally, and S. D. Guikema, "Extension of the application of conway-maxwell-poisson models: Analyzing traffic crash data exhibiting underdispersion," Risk Anal., vol. 30, no. 8, pp. 1268–1276, 2010, doi: 10.1111/j.1539-6924.2010.01417.x.
[21] A. Abdulhafedh, "Crash Frequency Analysis," J. Transp. Technol., vol. 06, no. 04, pp. 169–180, 2016, doi: 10.4236/jtts.2016.64017.
[22] G. P. Zhang, "Neural networks for classification: A survey," IEEE Trans. Syst. Man Cybern. Part C Appl. Rev., vol. 30, no. 4, pp. 451–462, 2000, doi: 10.1109/5326.897072.
[23] G. Fürst and P. Ghisletta, "Statistical Interaction between Two Continuous (Latent) Variables," 11th Congr. Swiss Psychol. Soc. August, pp. 1–12, 2009.
[24] M. G. Karlaftis and E. I. Vlahogianni, "Statistical methods versus neural networks in transportation research: Differences, similarities and some insights," Transp. Res. Part C Emerg. Technol., vol. 19, no. 3, pp. 387–399, 2011, doi: 10.1016/j.trc.2010.10.004.
[25] Y. Zhang and P. Lorenz, "AI for Network Traffic Control," IEEE Netw., vol. 32, no. 6, pp. 6–7, 2018, doi: 10.1109/MNET.2018.8553647.
[26] H. T. Abdelwahab and M. A. Abdel-aty, "Development of Artificial Neural Network Models to Predict Driver Injury Severity in Traffic Accidents at Signalized Intersections," Transp. Res., no. 01, pp. 6–13, 1997.
[27] L. Chang, "Analysis of freeway accident frequencies: Negative binomial regression versus artificial neural network," Elsevier, Science DirectSafety Sci., vol. 43, pp. 541–557, 2005, doi: 10.1016/j.ssci.2005.04.004.
[28] M. De Luca, "A Comparison between Prediction Power of Artificial Neural Networks and Multivariate Analysis in Road Safety Management," Transport, vol. 32, no. 4, pp. 379–385, 2017, doi: 10.3846/16484142.2014.995702.
[29] S. Araghinejad, M. Azmi, and M. Kholghi, "Application of artificial neural network ensembles in probabilistic hydrological forecasting," J. Hydrol., vol. 407, no. 1–4, pp. 94–104, 2011, doi: 10.1016/j.jhydrol.2011.07.011.
[30] M. Bekkar, H. K. Djemaa, and T. A. Alitouche, "Evaluation Measures for Models Assessment over Imbalanced Data Sets," J. Inf. Eng. Appl., vol. 3, no. 10, pp. 27–38, 2013, [Online]. Available: http://www.iiste.org/Journals/index.php/JIEA/article/view/7633
[31] F. He, X. Yan, Y. Liu, and L. Ma, "A Traffic Congestion Assessment Method for Urban Road Networks Based on Speed Performance Index," Procedia Eng., vol. 137, pp. 425–433, 2016, doi: 10.1016/j.proeng.2016.01.277.
[32] Y. Sun, M. S. Kamel, A. K. C. Wong, and Y. Wang, "Cost-sensitive boosting for classification of imbalanced data," Sci. Pattern Recognit., vol. 40, no. 12, pp. 3358–3378, 2007, doi: 10.1016/j.patcog.2007.04.009.
[33] G. Menardi and N. Torelli, Training and assessing classification rules with imbalanced data, vol. 28, no. 1. 2014. doi: 10.1007/s10618-012-0295-5.
[34] A. J. C. Sharkey, "On Combining Artificial Neural Nets," Conn. Sci., vol. 8, no. 3–4, pp. 299–314, 1996, doi: 10.1080/095400996116785.
[35] S. E. Kim and I. W. Seo, "Artificial Neural Network ensemble modeling with conjunctive data clustering for water quality prediction in rivers," J. Hydro-Environment Res., vol. 9, no. 3, pp. 325–339, 2015, doi: 10.1016/j.jher.2014.09.006.
[36] J. Han, M. Kamber, and J. Pei, "Third Edition: Data Mining Concepts and Techniques," J. Chem. Inf. Model., vol. 53, no. 9, pp. 1689–1699, 2012, [Online]. Available: http://library.books24x7.com/toc.aspx?bkid=44712
[37] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," J. R. Stat. Soc. Ser. B Stat. Methodol., vol. 63, no. 2, pp. 411–423, 2001, doi: 10.1111/1467-9868.00293.
[38] S. Chen, W. Wang, G. Qu and J. Lu, "Application of Neural Network Ensembles to Incident Detection", IEEE International Conference on Integration Technology, 388-393, 2007.
[39] T. Ma, F. Wang, J. Cheng, Y. Yu and X. Chen, "A Hybrid Spectral Clustering and Deep Neural Network Ensemble Algorithm for Intrusion Detection", Sensors, 2016.
[40] UNECE, "Road Safety for All", 2019

Celso L. Fernando Was born in Maputo Mozambican, in October 1st, 1984; in 2009 received his Bachelor degree in Architecture and Urban Planning from Eduardo Mondlane University in Maputo, Mozambique; 2018 received a Master degree in Engineering from Ehime University in Matsuyama city, Japan; 2021 earned the Doctoral degree in Engineering from the same university in Japan. The research fields of interest are traffic safety, travel behavior, artificial intelligence.

He worked as Assistant Lecturer for five years at Lurio University, Nampula-Mozambique, right after he received his bachelor degree; after that, moved to Japan to continue with his studies, under the sponsorship of the Japanese International Cooperation Agency (JICA) in ABE Initiative Program. Currently, he works as Lecturer at Lurio University Nampula-Mozambique. He has published three articles, one as a journal paper (Analysis of the Safety Performance of the Drainage Pavement focusing on Pavement Age - 2019), and the others as proceeding paper (Effect of Road Pavement Types and Ages on Traffic Accident Risk, and Effect of the Multicollinearity of Interaction Terms on the Performance of the ANN Model).