Effect of the additional input using a data mining technique on the interpretability of the deep neural network

Ehime University ○Fernando, Celso; Yoshii, Toshio; Tsubota, Takahiro; Shirayanagi, Hirotoshi

1. Introduction

Traffic accident risk analysis has mainly been dominated by traditional statistical methods [1][2][3]. Recently, as a result of the advance in data collection and storage, a large volume of accident data has become available, and statistical methods have revealed that cannot cope well with this huge volume of data [4]. This occurs because of the highly non-linear relationship among the attributes, worsened with the need for the interaction terms. To overcome such barriers the deep neural network (NN) has been used, however the black-box nature in the NN output as becomes a concerning topic for many researchers [5][6][7][8][9]. This study evaluates the effect of the additional input generated from data mining technique for the interpretability of the deep neural network. The interpretability of the NN is fundamental for its adoption in decisions having a deep impact on the life of people, such as health diagnostic, military decision and traffic management.

2. Methodology

Factor extraction applying association rule mining

Association rule mining was first proposed by Aggarwal, it was developed to identify items that are often found together in a market basket [10][11][12]. The itemset that are found together in a minimum required frequency are considered frequent itemset or frequent pattern and they define the association rules [10][11]. The frequent patterns will vary in number depending on the user defined threshold. Therefore, one of the major challenges in association rule mining is to avoid large number of patterns without losing of their interesting. The maximal itemset is one of the approaches in association rule mining techniques that can solve the above-mentioned issue [10]. Let call the maximal itemset extracted as *interaction terms*.

Neural network models

The models examined in this paper are feed-forward neural networks (NN). The architecture of the model comprises three hidden layers, with 22, 10 and 5 nodes respectively Fig.(1). The three hidden layers takes ReLU as the activation function and the output layer takes the sigmoid function. To avoid overfitting, neuron dropout is applied in training phase in the first hidden layer, with 25%. The loss function is a binary cross-entropy, the model was trained over 50 epochs and the accuracy is evaluated based on receiver operator characteristics area under the curve (ROC-AUC).

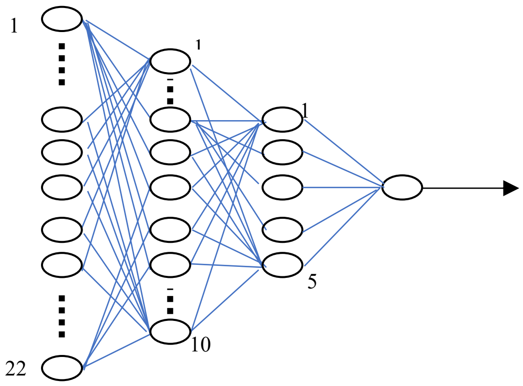

Fig. 1. Neural network architecture

For each interaction term extracted by the association rule mining, an NN model is developed. The ROC-AUC generated by the model that contains the respective interaction terms is compared with the ROC-AUC of the basic model Fig. (2). The basic model is one that does not have an interaction term as part of the input.

Experiment

The data used in this study was collected from an expressway route in Osaka Japan. The length of the route is 23.2km and, the study includes data of one-year April 2010 to March 2011. The attributes in the data consist of weather conditions, vertical and horizontal alignment of each 100 meters road segment, pavement material, and daily traffic volume. Five interaction terms 'maximal itemset' were extracted from the database, they are represented as follow: X_1 = {Straight, High traffic}; X_2 = {Straight, DPSA}; X_3 = {Dry, High traffic}; X_4 = {Dry, DPSA} and, X_5 = {Dry, Straight}. The training data correspond to 90% and validation data 10%. The sample size of the database is 1,073,380 where, 70 observations were classified as accidents events. The Fig (2) shows the result of the experiment, the horizontal line is the ROC of the basic model. The models with the interaction terms are in the x-axis, and they show that the interaction terms defined by X_1, X_3 and X_4 are the interaction terms with high ability to distinguish the traffic safety state (accident risk). While the interaction terms defined by X_2 and X_5 have low ability.

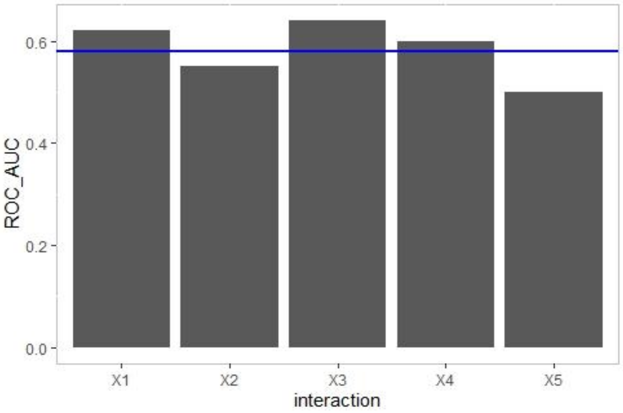

Fig. 2. Performance of the NN models

Conclusion

In this paper, the interpretability of the neural network based on the addition of the interaction term in the input of the NN model was examined. The goal was to understand the effect of each interaction term, by evaluating the performance of the NN in the classification task. The results show that different interaction terms will have different contribution for the ability of the model to discriminate against the real state of the traffic safety condition.

The next step of the research will focus in the understanding the proprieties of the interaction terms for interpretability of the NN.

References

[1] S. Hyodo and T. Yoshii, 'Analysis of the impact of the traffic states on traffic accident risk', *22nd ITS World Congr. Proc. Sci. Pap. ITS-2863, Bordeaux*, p. 2863, 2015.

[2] L. Lin, Q. Wang, and A. Sadek, 'Data mining and complex network algorithms for traffic accident analysis', *Transp. Res. Rec.*, vol. 2460, no. 1, pp. 128–136, 2014.

[3] T. Tsubota, C. Fernando, T. Yoshii, and H. Shirayanagi, Effect of Road Pavement Types and Ages on Traffic Accident Risk, Transp. Res. Procedia, vol. 34, pp. 211-218, 2018.

[4] 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.

[5] I. N. da Silva, D. H. Spatti, R. A. Flauzino, L. H. B. Liboni, and S. F. dos R. Alves, Artificial Neural Networks. Switzerland: Springer, 2017.

[6] M. Tsang, D. Cheng, and Y. Liu, 'Detecting Statistical Interactions from Neural Network weights', *ICLR*, 2018.

[7] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen & K. R. MÃžller, "How to explain individual classification decisions," Journal of Machine Learning Research, pp. 1803-1831, 2010.

[8] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, et al. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI," Information Fusion, 2019.

[9] M. Ancona, C. Öztireli and Gross, "Explaining Deep Neural Networks with a Polynomial Time Algorithm for Shapley Values Approximation," In ICML, 2019.

[10] C. C. Aggarwal and J. Han, Frequent Pattern Mining. New York: Sringer International Publication, 2014.

[11] C. C. Aggarwal, Data Mining. New York: Springer International Publication, 2015.