Multiple Classifier of Traffic Accident Based on Matter-Element Analysis

Chao Sun\textsuperscript{1a}, Wei Quan\textsuperscript{2,b*}

\textsuperscript{1}Harbin Institute of Technology Department of traffic science and engineering Harbin China 15000
\textsuperscript{2}Harbin Institute of Technology Department of traffic science and engineering Harbin China 15000
\textsuperscript{a}Email:19s132096@stu.hit.edu.cn
\textsuperscript{b*}Correspondence: weiquan@hit.edu.cn

Abstract—The paper establishes a new multiple classifier for grading the traffic accident based on the Hard Decision Model and the parallel topological structure. Logistic Regression, decision tree(DT) and BP Neural Network are selected as the base classifiers and matter-element analysis is employed as the fusion algorithm to improve the traditional topological structure. Matter-Element analysis has potential to increase the objectivity and dynamic in the criterion setting of the multiple classifier by combining the confidence coefficient of the base classification results. The accuracies of different base classifiers are used for weight factors calculation instead of the analytic hierarchy process (AHP) or the fuzzy synthetic evaluation (FSE) to avoid the influence of human factors. 200 recordings of the traffic accident are selected as a case study for methodology verification. The results show that compared with the base classifiers, the capacity to identify the class of traffic accident can be strengthened apparently by the multiple classifier based on confidence coefficient and Matter-Element Analysis. Meanwhile, this classifier provides the best opportunity to avoid many bad conditions of single model, such as overfitting and under-fitting.

1. INTRODUCTION

With the rapid development of cities and the surge of vehicle, the issues of traffic safety in the city have become increasingly prominent. The evaluation of urban traffic safety, especially classification of traffic accident, can not only quantitatively grasp the overall development level of urban transport, but also accurately locate the security problems in the city. And with the further development of the machine learning, the various identification and classification algorithms are becoming more and more mature, based on which linear regression\cite{1}, logistics regression\cite{2}, neural networks\cite{3} and decision tree\cite{4} are widely employed in classification of traffic accident. However, there are still certain space for the improvement in accuracy and stability of the classification achieved by a single model\cite{5}. Therefore, the multiple classifier consisting of the several base classifiers is playing a more and more important role in the issues of classification\cite{6}. This is because, compared with a single base classifier, the multiple classifier can improve the generalization ability of classification and avoid overfitting\cite{8}\cite{9}.

In existing papers, additionally, multiple classifiers can be divided into two types: SDM(\text{Soft Decision Model}) and HDM(\text{Hard Decision Model})\cite{10}. The SDF emphasizes more on the special base classifiers and curb of their weaknesses, whereas the latter pays more attention on the fusion
algorithms in the form of hard decision and some general base classifiers, such as Decision Tree, Logistic Regression and BP Neural Network[11]. This dissertation mainly studies on the improvement of the multiple classifier based on HDM. Generally, the training set is used to train the general base classifiers, and the fusion algorithm in the form of hard decision is used to fuse the results from general base classifiers. Therefore, apart from the selection of base classifiers, the most important component for a multiple classifier based on the HDM is topological structure, including CS (the cascade structure), PS (the parallel structure) and HS (hybrid structure)[12]. Choosing the right structure can not only improve the accuracy and generalization ability of classification, but also increase the efficiency of time and space. We have found, however, it still has potential to improve the topological structure of current multiple classifier because of its lack of consideration in correlations between the base classifiers, which has the bad impact on the accuracy and stability.

In response to the above-mentioned shortcomings, the overarching theme of this paper is to establish a new multiple classifier for grading the traffic accident based on the hard decision model and the parallel structure, which is achieved by Matter-Element Analysis. The main contributions of this work lie in three aspects: 1) How to use the confidence coefficient to establish the traffic accident grade recognition model when the traffic accident belongs to a certain class. 2) Matter-element theory is firstly selected to establish a multiple classifier based on the parallel topological structure because it systematically considers the correlations between general base classifiers, which aims to solve the incompatibility problems and dynamically ensure the implementation of the multiple classifier. 3) The accuracies of different general base classifiers are employed to calculate weight factors rather than analytic hierarchy process (AHP) or fuzzy synthetic evaluation (FSE) to avoid the influence of human factors on results classification.

2. METHODOLOGY

2.1. Fitting of Base classifier
First of all, the data of traffic accident, including class of traffic accident, characteristic of driver, road condition for driving and crossing condition for pedestrian, are randomly divided into two parts: the train set and the test set. Next, the training set is used to fit all the base classifiers in SPSS 23 in order to evaluate all relevant parameters as well as accuracy of classification. And then these base classifiers are employed to identify the class of traffic accident in the test set, which lays the foundation for estimation of the confidence coefficient where the traffic accident belongs to one class.

2.2. Estimation of Confidence Coefficient
In general, the confidence coefficient is the probability of guaranteeing that the error between the predicted value and the actual value is limited in one certain range. For example, if we point out that the accuracy of the predicted result is 95%, the corresponding percentage is called confidence coefficient, also known as the confidence level, which refers to the reliability where people approve of the predicted results. By means of Estimation of Confidence Coefficient and Confidence Interval in SPSS 23, we can get the exact values of confidence coefficient corresponding to the different classes.

2.3. Multiple classifier based on the Element Analysis

2.3.1 Matter-Element Definition of Multiple classifier
As an important method for qualitative evaluation, the matter-element analysis has been successfully used in many fields[13][14]. In matter-element analysis, the class of traffic accident can be divided into several minimum units which are represented by $M[15]$. The matter-element about the traffic accident can be represented by the orderly ternary array $R$ as shown in (1).
\[ R = (M, c, v) = \begin{bmatrix} M & c_1 & v_1 \\ c_2 & v_2 \\ \vdots & \vdots \\ c_i & v_i \\ \vdots & \vdots \\ c_s & v_s \end{bmatrix} \] (1)

Where \( M \) is class of traffic accident, and \( c \) is the base classifier used for assessing \( M \), and \( v \) is the confidence coefficient where the traffic accident belongs to one certain class. If traffic accident is depicted by \( n \) base classifiers, \( R \) would be expressed as a \( n \)-dimension array. In this paper, \( c_1 \) is LR, and \( c_2 \) is DT, and \( c_3 \) is BPNN.

### 2.3.2 Classical Domain and Segmented Domain of Matter-Element

The traffic accident is divided into \( m \) classes with level of \( M_j (j=1,...,m) \). For each \( M_j \) subjected to class \( j \), the ranges of probability of all \( c_i \) are called the classical domain\[16\], which is expressed as \( R_{cj} \). In matter-element \( R_c \), the whole range of probability \( v_{ci} \) of base classifier \( c_i \) would be divided into \( m \) intervals as \( a_{ci}-b_{ci} \) (\( j=1,2,...,m \)). Traffic accident can be identified as the \( j \)th class when the confidence coefficient obtained by using the \( ith \) base classifier is close to \( v_{cj} \). The matrix of the classical domain \( R_{cj} \) can be achieved as (2).

\[
R_{cj} = (M_j, c, v_{cj}) = \begin{bmatrix} M_j & c_1 & v_{cj1} \\ c_2 & v_{cj2} \\ \vdots & \vdots \\ c_i & v_{cj} \\ \vdots & \vdots \\ c_s & v_{cjs} \end{bmatrix}
\] (2)

Segmented domain of matter-element indicates the whole probability range of \( c_i \). For \( c_i \) in equation (2), the segmented domain would be the interval \( a_{ci}-b_{ci} \), as expressed in (3).

\[
R_p = (M_p, c, v_p) = \begin{bmatrix} M_p & c_1 & v_{p1} \\ c_2 & v_{p2} \\ \vdots & \vdots \\ c_i & v_{pi} \\ \vdots & \vdots \\ c_s & v_{ps} \end{bmatrix}
\] (3)

Where \( R_p \) represents the matter-element of segmented domain, \( M_p \) and \( v_p \) indicate the same
definition as \( M \) and \( v \) in (2) but apply to the segmented domain.

### 2.3.3 Correlation Degree Calculation

Based on the classical domain and the segmented domain, correlation degree between each model \( c_i \) and each class \( j \) is calculated by using the (4)-(6). \( K_j(v_i) \) is the correlation degree between base classifier \( i \) and class \( j \), and \( \rho(v_i, v_{oj}) \) is the distance between the \( v_i \) and \( v_{oj} \), and \( \rho(v_{pi}, v_{pi}) \) is the distance between the \( v_i \) and \( v_{pj} \).

\[
K_j(v_i) = \begin{cases} 
\rho(v_i, v_{oj}) & v_j \notin v_{oj} \\
\rho(v_i, v_{oj}) - \rho(v_{pi}, v_{pj}) & v_j \in v_{oj} 
\end{cases}
\]  

(4)

\[
\rho(v_i, v_{oj}) = |v_j - 0.5(a_{oj} + b_{oj})| - 0.5(b_{oj} - a_{oj}) \tag{5}
\]

\[
\rho(v_{pi}, v_{pj}) = |v_j - 0.5(a_{pi} + b_{pi})| - 0.5(b_{pi} - a_{pi}) \tag{6}
\]

### 2.3.4 Weight Calculation and Class Evaluation

The integrated correlation degree between the traffic accident and the evaluation class is one of the most important components of multiple classifier. The integrated correlation degree of each matter-element \( M_{cj} \) can be calculated by following equation:

\[
K_j(M_{cj}) = \sum_{i=1}^{\infty} w_i K_j(v_{oi}) \tag{7}
\]

Where \( K_j(M_{cj}) \) indicates the integrated correlation degree between traffic accident and class \( j \). \( w_i \) is the weighted value of base model \( c_i \). If \( K_j = \max[K_j(M_{cj})](j = 1, \ldots, m) \), \( M_{cj} \) belongs to class \( l \).

Instead of deploying the traditional methods for calculating the weighted value, such as analytic hierarchy process (AHP) and fuzzy synthetic evaluation (FSE), this paper applies goodness of fit as the weight to reduce the subjectivity, which is illustrated carefully in (10).

## 3. Case studying

### 3.1. Data Description

In this paper, 200 traffic accident recordings in United Kingdom are used to clearly demonstrate the steps of modeling and test its validity. The severity of traffic accident has been divided into 5 class corresponding to \( M = \{M_{c1}, M_{c2}, M_{c3}, M_{c4}, M_{c5}\} \), and other fields (independent variables) are shown in the table I.

### 3.2. Modeling for Matter-Element Analysis

**Step 1: model fitting and classifying based on the base model**

| filed name            | description                              | symbol |
|-----------------------|------------------------------------------|--------|
| Road_Class            | the class of road where traffic accident happened | \( x_1 \) |
| Junction_Control     | the types of control for junction around which traffic accident happened | \( x_2 \) |
| Pedestrian_Crossing-Physical_Facilities | If there is facilities used for crossing road by pedestrian at the place where traffic accident happened | \( x_3 \) |
In this paper, 200 traffic accident recordings are divided into two parts: one is the training set, and
the other is test set. The training set is employed to fit all the base classifiers to obtain all the parameters as well as goodness of fitting.

Table II-IV show the information of fitting for LR, BPNN, and DT. According to goodness of fitting and accuracy of classification (shown in the Fig.1) in train set, the performance of DT is very poor. Although the LR is better than others, there is still a certain space for improvement. After that, the classes of traffic accidents with the confidence coefficients (shown in the Fig.2) in test set can be identified by means of these base classifiers.

![Figure 1. The difference between actual value and results of classification from different classifiers in train set](image1)

![Figure 2. The demonstration of confidence coefficient obtained in SPSS](image2)

**Step 2: Building matter-element of accident severity.**

Based on the above mentioned results, the matter-element model consisting of LR, BPNN, and DT is established to identify the class of accident for every recording in test set. Taking the first record in test set as an example, the detailed steps of modeling are shown as following. Table V shows gradation of accident severity based on the confidence coefficient. This means that we can evaluate the class of traffic accident by calculating the distance between the confidence coefficient from one certain base classifier and confidence interval in the Table V.

| Base model | $M_{c1}$ | $M_{c2}$ | $M_{c3}$ | $M_{c4}$ | $M_{c5}$ |
|------------|----------|----------|----------|----------|----------|
| LR         | [0.8,1]  | [0.8,1]  | [0.8,1]  | [0.8,1]  | [0.8,1]  |

![TABLE V. Gradation of accident severity](image3)
Step 3: Building the classical domain and segment domain

It is unnecessary to normalize the Table V, because the probability where the traffic accident belongs to one certain class is within 0-1. Based on the theory about matter-element analysis, classical domain for $M_{ci}$ is shown as following:

$$R_{ci} = \begin{bmatrix} M_{ci} & c_1 & [0.8,1] \\ c_2 & [0.8,1] \\ c_3 & [0.8,1] \end{bmatrix}$$

(8)

Where $c_1$ is $LR$, and $c_2$ is $NN$, and $c_3$ is $DT$. $R_{c2}$, $R_{c3}$, $R_{c4}$, $R_{c5}$ can be calculated similarly as $R_{c1}$. Because of the same scale (0-1) for confidence coefficient ($v_i$), segment domain is expressed as following equation:

$$R_{p} = \begin{bmatrix} M & c_1 & [0,1] \\ & [0,1] \\ & [0,1] \end{bmatrix}$$

(9)

Step 4: Weight evaluation based on identification accuracy

In the existing theoretical studies, AHP and FSE are widely employed to determine the weight of index. AHP combines the qualitative analysis with quantitative analysis into make a decision[17], but there is lots of subjectivity in steps of evaluation. This paper deploys the classification accuracies of the 3 base classifiers in the train set for weight calculation, the values of which embody the individual performance. The weight of each $c_i$ can be calculated by following:

$$\omega_i = \frac{|aop_i|}{\sum_i |aop_i|}$$

(10)

where the $aop_i$ is the accuracy of identification for $c_i$ in train set, and $\omega_i$ is weight of $c_i$ which is calculated as $\omega = (0.336, 0.336, 0.300)$.

Step 5: Evaluating the traffic accident class of 1th record

The confidence coefficients with which the accident severity of the first record in the test set belongs to one certain class ($M_{c1}, M_{c2}, M_{c3}, M_{c4}, M_{c5}$) according to base classifiers ($LR$, $NN$, $DT$) are substituted into (1) to build the matter-element of the traffic accident, which is expressed as following:

$$R_{e} = [M_{c,v}] = [M_{c1,v}, v]$$

(11)

By using (5)-(6), the correlation degrees between each index and class are calculated, which is presented in Table 6.

| TABLE VI. Correlation degree |
|-----------------------------|
| $LR$ | $M_{c1}$ | $M_{c2}$ | $M_{c3}$ | $M_{c4}$ | $M_{c5}$ |
| $NN$ | 0.089 | -0.878 | -0.822 | -0.878 | -0.878 |
| $DT$ | 0.006 | -0.886 | -0.888 | -0.889 | -0.887 |
|     | -0.022 | -0.822 | -0.889 | -0.889 | -0.722 |

The integrated correlation degrees are calculated as $K(M) = \omega * M_{e} = (0.028, -0.887, -0.887, -0.908, -0.857)$. Because the max integrated degree is $K(M) = 0.028$, the class of accident severity of the first record in the test set is $M_{c1}$, which is as same as the actual value. However, the class of it is classified as $M_{c2}$, $M_{c1}$, $M_{c3}$ according to $LR$, BPNN and DT respectively.

Step 6: Evaluation of all the recordings in the test set and testing the validity

The multiple classifier is employed to identify the classes of traffic accidents for 100 data
recordings in the test set, with which we could test the validity of it.

![Figure 3](image3.png)

**Figure 3.** The difference between actual value and results of classification from different classifiers in test set

![Figure 4](image4.png)

**Figure 4.** The accuracy of different classifiers

According to Fig.3 and Fig.4, with accuracy of only 38%, the performance of decision tree is worst, whereas logistic regression and neural network has better ability to identify the class of traffic accident, the accuracy of which is 84% and 87% respectively. However, by combining the different base classifiers into one based on matter-element analysis, there is obvious improvement in the multiple classifier because of its identification accuracy reaching up to 100%. In conclusion, the multiple classifier put forward in this paper is useful and valid, which provides a new approach to construct a reliable classifier in other fields.

4. CONCLUSION

This paper explores a new multiple classifier based on the hard decision model and the parallel structure, which realizes the classification of traffic accident with the better accuracy and stability than single model.

Matter-Element theory is selected to establish a multiple classifier consisting of different base classifiers, including LR, BPNN, and DT, which has potential to improve the accuracy of classification and effectively guarantee the application of the classification system for traffic accident. Furthermore,
the paper improves the matter-element evaluation method to increase the objectivity and dynamic in the criterion setting of the multiple classifiers, which is based on the mining of the confidence coefficient where the traffic accident belongs to one certain class. The data of traffic accidents in UK, including class of traffic accident, characteristic of driver, road condition for driving and crossing condition for pedestrian, are used to demonstrate the steps of modeling and test the validity and ability to classify. 200 recordings of traffic accident are randomly divided into two parts: train set and test set.

Furthermore, the corresponding data of 100 accident recordings in train set are used to fit the base classifiers and the multiple classifier, while other 100 recordings in the test set are employed to test the accuracy and stability. The results show that compared with the general base classifiers, the capacity to identify class of traffic accident can be strengthened obviously by using multiple classifier based on confidence coefficient and matter-element analysis because the classification accuracy is higher than that of LR, DT and BPNN.

ACKNOWLEDGMENTS

The authors wish to acknowledge the financial support of National Natural Science Foundation of China (51578197, 51578198).

REFERENCES

[1] X. Zou, Y. U. E. W. Long, and C. Li, “Road Traffic Accident Prediction using an SCGM (1,1)-c-Markov Model,” Asian Transp. Stud., vol. 5, 2018.

[2] H. H. Joni, A. M. Al-Dahawi, and O. J. Al-Tamimi, “Analysis of traffic accident severity in Baghdad city using Binary Logistic Regression Model,” IOP Conf. Ser. Mater. ence Eng., vol. 737, p. 12140, 2020.

[3] A. Dantas, K. Yamamoto, M. V Lamar, and Y. B. T.-I.-I.-. I. J. C. on N. N. Yamashita Ijcn, “Neural Network for Travel Demand Forecast Using GIS and Remote Sensing,” 2000.

[4] S. Moral-Garcia, J. G. Castellano, C. J. Mantas, A. Montella, and J. Abellán, “Decision Tree Ensemble Method for Analyzing Traffic Accidents of Novice Drivers in Urban Areas,” Entropy, vol. 21, no. 4, p. 360, 2019.

[5] E. O. Wiley, K. Menyset, T. Peterson, R. Robins, and A. Stewart, “Niche Modeling Perspective on Geographic Range Predictions in the Marine Environment Using a Machine-learning Algorithm,” Oceanography, vol. 16, no. 3, pp. 120–127, 2003.

[6] S. Petridis, V. Rajgarhia, and M. B. T.-J. C. on F. A. Pantic, “Comparison of Single-model and Multiple-model Prediction-based Audiovisual Fusion,” 2015.

[7] H. M. Khalid and C. H. Peng, “Immunity Toward Data-Injection Attacks Using Multisensor Track Fusion-Based Model Prediction,” IEEE Trans. Smart Grid, vol. 8, no. 2, pp. 697–707, 2017.

[8] J. Xie et al., “Chinese text classification based on attention mechanism and feature-enhanced fusion neural network,” Computing, vol. 102, no. 3, pp. 683–700, 2020.

[9] Q. Li, J. Dong, J. Zhong, Q. Li, and C. Wang, “A neural model for type classification of entities for text,” Knowledge-Based Syst., vol. 176, no. JUL. 15, pp. 122–132, 2019.

[10] X. Jin, P. Cheng, W. L. Chen, and H. Li, “Prediction model of velocity field around circular cylinder over various Reynolds numbers by fusion convolutional neural networks based on pressure on the cylinder,” Phys. Fluids, vol. 30, no. 4, p. 47105, 2018.

[11] D. Ruano-Ordás, L. Burggraaff, R. Liu, C. van der Horst, and G. J. P. van Westen, “A multiple classifier system identifies novel cannabinoid CB2 receptor ligands,” J. Cheminform., vol. 11, no. 1, 2019.

[12] Jian-Yu et al., “Predicting combinative drug pairs via multiple classifier system with positive samples only.,” Comput. Methods Programs Biomed., 2018.

[13] D. Yu and W. B. Chen, “Land ecological safety in Poyang Lake Eco-economic Zone: An evaluation based on matter-element model,” Ying Yong Sheng Tai Xue Bao, vol. 22, no. 10, pp. 2681–2685, 2011.
[14] M. Asghari, R. K. Saray, and E. Neshat, “Misfire detection of HCCI engines using matter-element extension theory and thermodynamic multi zone model,” *Environ. Prog. Sustain. Energy*, 2020.

[15] P. Gu, W. Xi, W. Ye, J. Shi, and J. Zhao, “Extenics matter-element analysis on dilemma problem in HMI design of nuclear power plant,” *Nucl. Eng. Des.*, vol. 350, no. AUG., pp. 176–181, 2019.

[16] H. B. Zhang, L. M. Liu, J. L. Zhang, and Z. Q. Zhu, “Matter-element model and its application to land resource eco-security assessment,” *J. Zhejiang Univ.*, 2007.

[17] A. E. Wolnowska and W. Konicki, “Multi-criterial analysis of oversize cargo transport through the city, using the AHP method,” *Transp. Res. Procedia*, vol. 39, no. 2018, pp. 614–623, 2019, doi: 10.1016/j.trpro.2019.06.063.