Journal on Internet of Things
DOI:10.32604/jiot.2020.09868
Article

Gender Forecast Based on the Information about People Who Violated Traffic Principle

Rui Li[1], Guang Sun[1,*], Jingyi He[1], Ying Jiang[1], Rui Sun[1], Haixia Li[1], Peng Guo[1,2] and Jianjun Zhang[3]

[1]Hunan University of Finance and Economics, Changsha, China
[2]University Malaysia Sabah, Kota Kinabalu, Malaysia
[3]Hunan Normal University, Changsha, China
[*]Corresponding Author: Guang Sun. Email: imon5115@163.com
Received: 08 November 2019; Accepted: 07 June 2020

Abstract: User portrait has been a booming concept in big data industry in recent years which is a direct way to restore users' information. When it talks about user portrait, it will be connected with precise marketing and operating. However, there are more ways which can reflect the good use of user portrait. Commercial use is the most acceptable use but it also can be used in different industries widely. The goal of this paper is forecasting gender by user portrait and making it useful in transportation safety. It can extract the information from people who violated traffic principle to know the features of them then forecast the gender of these people. Finally, it will analyze the prediction based on characteristics correlation and forecasting results from models which can verify if gender can have an obvious influence on the traffic violation. Also we hope give some advice to drivers and traffic department by doing this research.

Keywords: User portrait; gender forecast; feature selection; correlation analysis; traffic violation

1 Introduction

User portrait has been used widely in different areas as an efficient way of describing goal users and knowing users' requirements. In this period of big data, internet develops quickly so users' information is dispersed and everywhere [1,2]. User portraits put each specific piece of information of users into a label then visualize users through these labels. It builds relationships between data that seems irregular and makes them become clear, vivid and easy to understand so it can be accepted and used by all walks of life. It has become a research hotspot in recent years [3].

The main research point of this paper belongs to the prediction of gender with user portraits. Gender is one of essential attributes of users [4]. Also, it is one of the most different characteristics among human beings. Gender will influence users' behaviors, hobbies and requirements. As the same way, we can predict users' gender based on extracting the characteristics from all kinds of users' information to provider specific service to users, and it can also help us know our users better [5].

Prediction of gender with user portrait has aroused researchers' interest all over the world because of the potential applications like accurate marketing, injecting advertisements and personalized recommendations. The describing ways based on the users' behaviors, moods and hobbies, topics, personalities are the important points and hotspots for researching in academic circles and economic circles [6]. There are various predicting ways of gender like Random Forest (RF), Decision Tree (DT), Logistic Regression (LR) and Support Vector Machine (SVM) that are common classification, predicting methods and models in computer study. But what should be pointed out is that there is just few research

results of user portrait and its relevant content that mostly depends on documents. However, most of the textual information is private, making it more difficult to obtain, also it is hard to know the textual characteristics [7]. In addition, most researches are based on user behaviors such as e-commerce shopping and social networks, while traffic safety and violations are rarely involved. Although major influencing factors of traffic accidents involving drivers of different genders have been studied, gender prediction using the information of drivers violating the rules is rarely mentioned.

This study is based on the information of traffic violators in Maryland, USA. On the one hand, the public has different impressions and ideas about male and female drivers, but it is generally believed that men have an innate advantage in driving [8]. Especially in recent years in all kinds of media reports, female drivers have become synonymous with the "road killer". One of the key issues in the debate is whether there is a correlation between gender and traffic accidents [9]. This paper intends to verify and summarize the results through prediction. On the other hand, different from the more common traffic violation rate projections, this paper starts from another Angle to study the correlation between various factors of traffic violation and gender, which can provide relevant theoretical basis and data support for traffic accident prevention, and is also of great significance for traffic safety management [10].

According to the chosen data characteristics, we decided to use Decision Tree Model. In the process of data pretreatment, we made data redundancy and noise-removing handle then extracted the new characteristics. We made participle and word frequency statistics for the text message and transferred all the transformable class variable to numerical categories. Finally, we calculated the correlation value between each characteristic and gender to get final graph of characteristics.

When doing visualized analysis to the result of this research, bar chart is the best way to make statistics of the people who violated traffic principles. Make word cloud map to show the direct result of text segmentation and statistics. Then thermodynamic chart can reflect the correlation between the features. Finally, confusion matrix and classification report can evaluate the result of models which can tells if it is a good research.

2 Overview of Decision Tree (DT) Model

As it mentioned, there are lots of models that can be used to classify and predict users'' gender, all of them are different from each other and have own key points. We select the Decision Tree Model according to the features of the data to provide convenience and learning for the following research.

Decision Tree is a graphical and decision analyzing method which uses probability analyzing directly and based on knowing the probabilities of all kinds of situations. It decides the feasibility through estimating the risk of an item and knowing the probability that is equal or greater than 0 by making decision tree. It is called Decision Tree because the branch of this decision is painted like a branch of a tree. Entropy means how messy in this system and this meaning of "Entropy" is based on informatics theory, so we can get entropy by using algorithm ID3, C4.5 and C5.0 to generate decision tree [11]. Decision Tree is divided into two category-classification tree and regression tree. Regression tree aims at continuous variable and classification tree aims at discrete variable. Decision tree has a tree shape, every internal node represents a test of quality, every branch represents an output of test, and each leaf node represents a category.

There are three processes of generating a decision tree. Firstly, choose the characteristics: It means choosing characteristic from the different characteristics in training data as a splitting standard of present node. There are many different quantitative estimation standards of how to choose characteristics which derives many different decision tree algorithms. Secondly, generation of decision tree: Child nodes will generate from above down recursively according to the chosen characteristics' estimation standards. Recursive structure is the easiest one to understand in decision tree. And thirdly, pruning: It is easy for decision tree to become over-fitting so it needed to be pruned to downsize and alleviate. Pruning includes prepruning and backward pruning [12].

And there also are some requirements of using decision tree: Goals that decision makers want to get. There will be at least two feasible alternatives that decision makers can choose from. There will be at least two uncertain factors that decision maker cannot control. The loss and gain can be calculated in different schemes with different factors. The decision makers can estimate the probability of uncertain factors [13].

2.1 The advantages and Disadvantages of This Algorithm

It is easy to understand and accomplish. It can deal with both data-type and regular-type attributes, and can make feasible and effective results for large data sources in a relatively short time. Then it is easy to measure the reliability of the model by static test. Given an observed model, it can derive logical expressions based on the resulting decision tree easily.

But continuous fields are harder to predict. A lot of preprocessing is required for chronological data. When there are too many categories, errors may increase more quickly. The general algorithms classify only depend on one field.

2.1 The scope of Application

Decision tree can help analyze the risk and direction of operating for decision maker of an enterprise easily. The accuracy of a decision depends on scientific decision ways.

The enterprises and analyzers can easily use decision tree when they have accumulated enough data and resources of their customers if they want to classify user.

Decision tree always link with analyzing goal and background, for example: decision tree can estimate the risk of debt in the financial industry. It can also be used for promoting some kinds of insurance in the insurance industry and it can generate auxiliary diagnosis model in the medical industry [14].

3 Process and Methods

3.1 Global Mutual Interference Coefficient Based on Matrix

The data in this research are based on a data set The Information about People Who Had Violated Traffic Principles in Maryland from 2012 to 2018 which is from the open data in Kaggle. The origin data set (Traffic_violations.csv) is a 1048576*35 CSV form and it records the information about drivers and cars, the process of those accidents and the reasons why those accidents happened from 2012 to 2018. Some data is as Fig. 1 shown.

	Accident	Belts	Personal	Property I	Fatal	Commercial	HAZMAT	Commercial	Alcohol	Work Zone	State	VehicleType	Year	Make	Model	Color
2	No	No	No	No	No	No	No	No	No	No	MD	02 - Automobile	2008	FORD	4S	BLACK
3	No	No	No	No	No	No	No	No	No	No	VA	02 - Automobile	2001	TOYOTA	COROLLA	GREEN
4	No	No	No	Yes	No	No	No	No	No	No	MD	02 - Automobile	2001	HONDA	ACCORD	SILVER
5	No	No	No	Yes	No	No	No	No	No	No	MD	02 - Automobile	1998	DODG	DAKOTA	WHITE
6	No	No	No	No	No	No	No	No	No	No	MD	02 - Automobile	2015	MINI COOPER	2S	WHITE
7	No	No	No	No	No	No	No	No	No	No	MD	02 - Automobile	2013	HYUNDAI	ELANTRA	GRAY
8	No	No	No	No	No	No	No	No	No	No	MD	02 - Automobile	1993	FORD	PICKUP	BLACK
9	No	No	No	No	No	No	No	No	No	No	VA	02 - Automobile	2003	DODGE	SPRINTER	WHITE
10	No	No	No	No	No	No	No	No	No	No	MD	02 - Automobile	2015	MINI COOPER	2S	WHITE
11	No	No	No	No	No	No	No	No	No	No	MD	02 - Automobile	2015	MINI COOPER	2S	WHITE
12	No	No	No	No	No	No	No	No	No	No	MD	02 - Automobile	2005	CADI	STS	BLACK
13	No	No	No	No	No	No	No	No	No	No	VA	02 - Automobile	1996	HONDA	CIVIC	SILVER
14	No	No	No	No	No	No	No	No	No	No	MD	02 - Automobile	2004	CHEVROLET	IMPALA	SILVER
15	No	No	No	No	No	No	No	No	No	No	MD	02 - Automobile	2005	AUDI	4S	GRAY
16	No	No	No	No	No	No	No	No	No	No	MD	02 - Automobile	2002	TOYT	4S	RED
17	No	Yes	No	No	No	No	No	No	No	No	MD	02 - Automobile	2009	DODGE	CHARGER	BLACK
18	No	No	No	No	No	No	No	No	No	No	MD	02 - Automobile	2000	SATURN	LS	SILVER
19	No	No	No	No	No	No	No	No	No	No	MD	02 - Automobile	2003	HONDA	ACCORD	GOLD
20	No	No	No	No	No	No	No	No	No	No	MD	02 - Automobile	2015	MINI COOPER	2S	WHITE
21	No	No	No	No	No	No	No	No	No	No	MD	02 - Automobile	2005	TOYOTA	CAMRY	BLACK
22	No	No	No	No	No	No	No	No	No	No	MD	02 - Automobile	2015	MINI COOPER	2S	WHITE
23	No	No	No	No	No	No	No	No	No	No	VA	28 - Other	2002	FORD	ECONOLINE	WHITE

Figure 1: Part of the origin data

3.2 Data Pretreatment

We can extract some useful characteristics to make a new form after data reading, because the data in that form is irregular which has text forms and Boolean forms. Beside of it, during data cleaning, text data and noise data that cannot be converted into numerical type are deleted, then it will reduce the steps and difficulty of subsequent machine processing [15].

The next step is to check null value and then delete all rows which include null value and reset index. Some of characteristic from origin data need to be transformed to gain more useful information. Firstly, choose the right date from the above years to create a “date” attribute. Then deduct the date when the car had already been made well to know how long this car had been kept. In the end, add this result to the form.

To know the specific situations well about the violation, we extracted the information "the specific description violation" solely to make text segmentation in English and statistic the frequency of words. The participle can be separated by spacing directly because all of the English words are capital letters. But when deleting the stop words, all of the letters need to be transferred to minuscule in case of analyzing. The following image is part of description after transformation and deletion [16].

	Description
0	driving vehicle highway suspended registration
1	driver failure obey properly traffic control...
2	failure yield hwy
3	failure yield
4	failure dr lane change avail lane immed ...
5	negligent driving vehicle careless imprudent...
6	driving vehicle highway suspended registration
7	driver fail flashing red traffic signal
8	failure individual driving highway display ...
9	driving vehicle highway expired license
10	failure drive hand roadway divided hwy
11	driver hands handheld telephone whilemotor ...
12	occupant restrained seatbelt

Figure 2: Part description of violation

When choosing the characteristics, it can be more convenient to know the relativity between characteristics and gender attribute through correlation coefficient method if using numerical value to replace the categorical variables of each characteristic [17]. Then we can choose characteristics in series. Pearson coefficient correlation can reflect the level of correlation between two variables, but the result reflects that the coefficient correlation is not very high between each characteristic and gender attribute also there are positive and negative numbers at the same time. So we use Kendall which is used to reflect the index of categorical variables’ correlation coefficient and Spearman which points at nonlinear data to calculate and then we find the calculation result is similar with Pearson’s. Therefore, for the sake of the future prediction model, only the characteristics of positive correlation are left.

After the sifting we get final form(feature.csv) and its shape is (1034594, 8), the details of its characteristics are as the following Tab. 1 shown.

Table 1: Features table of violators information

Name of attribute	Illustration of attribute
Gender	Female (F = 1), Male (M = 0)
Belts	Whether the violation involves the use of seat belt (Yes = 1, No = 0)
Fatal	Whether there were any deaths in the violation（Yes = 1, No = 0）
Year	When the car had been made
Color	The color of the violation vehicle (Mark from 1–26)
Violation Type	which type of the violation (Citation = 1, Warning = 2
Arrest Type	What was the reason for the arrest (Mark from 1–6)
Date	Violation year

After getting the final characteristics form we begin to create models and predict. We use Scikit-learn, a third-party library from Python. Scikit-learn were based on Numpy library and Matplotlib library which can used for classifying, regression, dimension reduction and cluster analysis. The efficiency of machine learning can be improved quickly if using there 4 modules' advantages correctly [18]. Because the work of this paper is to predict the gender of violators, which is a binary classification problem. In addition, according to the characteristics of the data, the decision tree algorithm is selected to model the training set, conduct attribute comparison from top to bottom, and predict the gender of the test set.

Firstly, divide data into the training set accounted for 80% and the test set for 20% by using train_test_split. Assume "gender" is the dependent parameter y and other characteristics are the independent parameter "x". Secondly, train the training set by using decision tree model and then factor test set into model to predict which can get the prediction the gender (X_text). Finally, make comparison between x and the real value (Y_text) then get a score and have the result from models. At last the data should be operated and disposed visually.

3.3 Analyze the Visual Experiments Results

According to the research about the relativity between kinds of factors, characteristics and attributes of gender, the gender can be predicted finally, so gender is indispensable in this study. Now we make statistics about the number of females and males in the data set and make the following bar chart (Fig. 3).

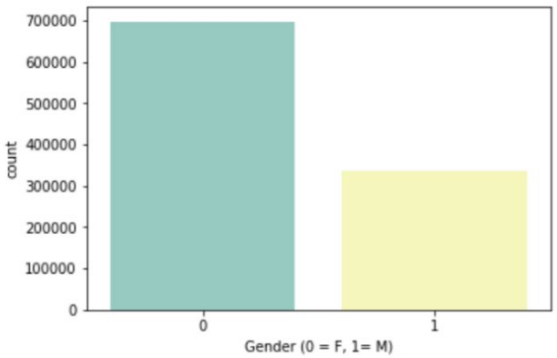

Figure 3: Statistical bar chart of the number of gender

From the above image we can know the probabilities cannot get a balance in the people who violated traffic principles from 2012 to 2018. Female is about 700000 which means there are about twice as many females as males and that indicates it is more possible for women to violate the traffic principles.

After making text segmentation of "Description" we made statistics about the frequency of words and made the following word cloud map.

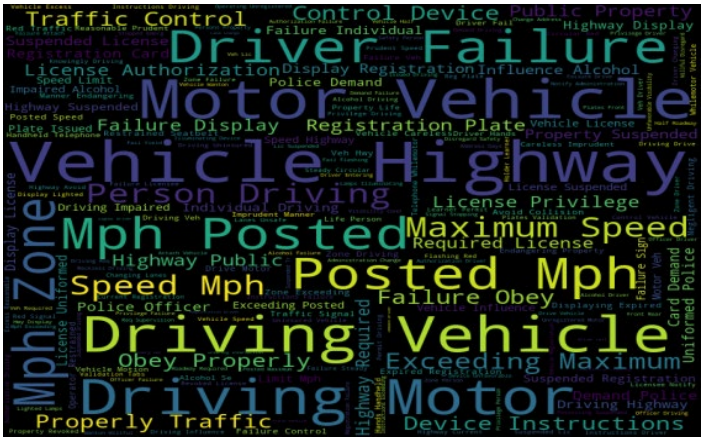

Figure 4: Word cloud map of violation description

As the Fig. 4 shown, apart from the words "Driving", "Vehicle", "Motor", "Highway" and other words that describe the Vehicle and road condition, we find that the words "Posted", "Mph", "Maximum Speed", "License", "obey" and other words are bigger. It can be speculated that the reasons of the violation may also be related to speeding, driving License and failure to obey the correct traffic rules.

The most important thing about checking the correlation coefficient among different characteristics is checking the correlation coefficient between the gender in first column and other characteristics. Then make a thermodynamic diagram (see Fig. 5).

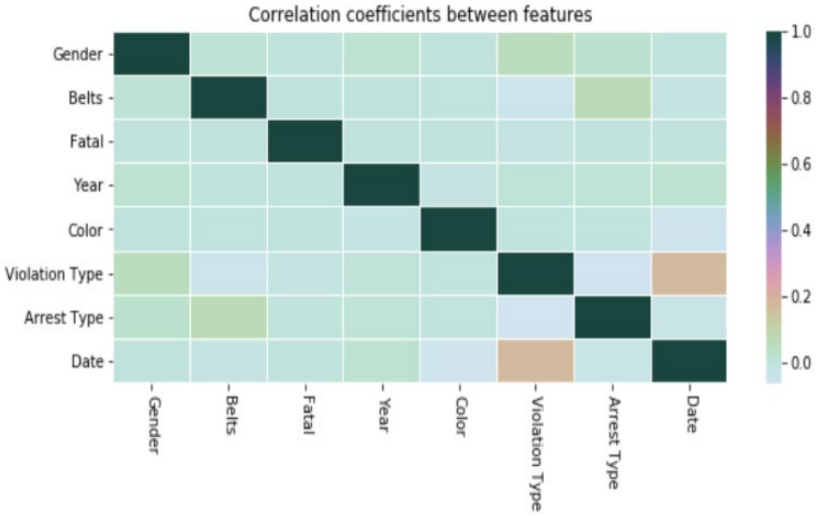

Figure 5: Thermodynamic diagram

We can know the correlation values of all features were very low, close to 0 or even negative. Therefore, it could be found that there was basically no linear correlation between all features through Pearson correlation coefficient method, while both Kendall and Spearman principle could verify that there was no obvious correlation between each feature and gender.

We get final result after making model and training by using decision tree. Comparing the real value with predictive value from gender and get the accuracy of the model test (see Fig. 6).

```
true=np.sum(clf.predict(X_test)==Y_test)
print('True: ',true)
print('False: ',clf.predict(X_test).shape[0]-true)
print('Accuracy on Test: ',true/clf.predict(X_test).shape[0])
```

```
True:  139997
False:  66922
Accuracy on Test:  0.6765787578714376
```

Figure 6: Diagram of experimental results

We can know 20% of experiment is centralized from Fig. 6, there are 139997 right results and 66922 wrong results which means the accuracy rate by using model prediction is 67.66%. It is not ideal.

Confusion matrix is a condition analysis graph which includes the predictions' conclusion and classification in machine study. Confusion matrix records the data together in matrix and concludes the two standards which are real category and classification model prediction. Fig. 7 is the confusion matrix which includes the real value and predicted value, the sum of (0,0) and (1,0) is the data of positive samples, and sum of (0,1) and (1,1) is the data of negative samples.

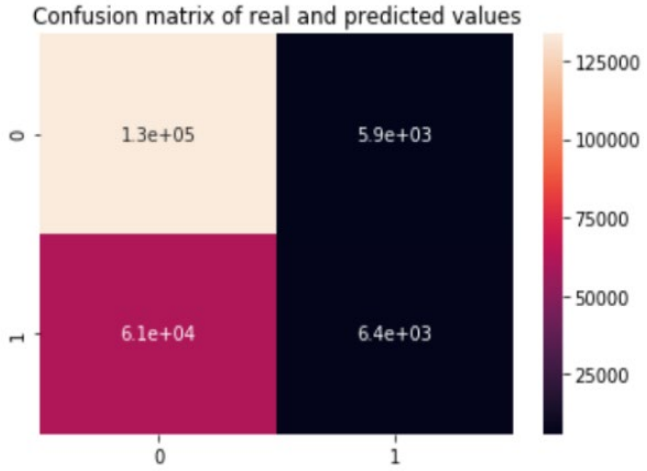

Figure 7: Confuse matrix diagram of real and predicted values

Recall means the proportion of right prediction positive samples in the whole positive samples. F1 is the evaluating index of precision index and recall index which is used to reflect the whole index synthetically [19]. The three indexes are calculated by confusion matrix (see Fig. 8).

```
from sklearn.metrics import classification_report
print('Report Of DT:\n',
      classification_report(clf.predict(X_test),Y_test))
```

```
Report Of DT:
              precision    recall  f1-score   support

           0       0.96      0.69      0.80    194573
           1       0.10      0.52      0.16     12346

avg / total       0.91      0.68      0.76    206919
```

Figure 8: The classification of decision tree

Combine the two pictures above we can know the prediction of female has higher right accuracy rate (96%) than male (10%). So this model does not fit male prediction well. From the recall rate, the proportion of positive samples correctly predicted by men and women is not high. Therefore, from theperspective of F1 score, the overall index is low and the model effect is not ideal.

4 Conclusions and Future Work

We find that accuracy rate of model prediction won't increase when the test set decrease and it will not decrease when the test set increase if not in accordance with the 20% of the test set divided. It is irregular in accuracy rate and has no correlation with the amount of training set and test set，but it will not be more than 69%. Then we can deduce that there is no direct correlation between gender and other factors in violation. So we cannot decide the gender directly only by gender which means gender difference has no specific form in traffic violation.

Generally, the non-ideal results can be caused by many reasons. Firstly, when pretreating data, there's no mature process of choosing and extracting so we haven't got more valuable characteristic attribute. Secondly, the chosen data set did not fit to make classification prediction. Because we just mentioned above that there is no obvious correlation between gender and each characteristic. There is no good standard for internal node to compare attributes for decision tree which will have difficulty in classifying. It also proves our research conclusion from another side.

From the experiment on the other hand also can prove that, Although the number and probability of traffic violation are higher for women than for men, but that is likely to be affected by the local population, the traffic environment or other factors, cannot be attributed to gender differences, also it can't think that gender factor is the cause of illegal violations occur, it may have influence, but is not directly factor. Therefore, in daily traffic travel, drivers, both men and women, should be more serious and strictly abide by traffic laws and regulations, which is responsible for themselves and others. Nor should accidents be blamed on gender weakness, since the law does not discriminate on the basis of sex.

This research is still relatively simple, there are many shortcomings. Since only the decision tree model is applied in this paper, we will make more efforts to understand the prediction models of various categories and optimize them based on this study to compare whether different models have different prediction results. In addition, in view of the shortcomings in feature selection, we will continue to analyze and understand the features in the original data, and further study on feature engineering, so as to make the research more serious and rigorous.

We gain lot from the gender prediction about user portrait [20]. The gender prediction asks not only gender attribute but also the correlation between it and other characteristics to check if it can be used to predict. Also the gender prediction is just a small part of user portrait so we should research more to learn it better when we can use it well [21]. We believe it can bring much convenience and information resources to people in the well-developed future.

Acknowledgement: This research work is implemented at the 2011 Collaborative Innovation Center for Development and Utilization of Finance and Economics Big Data Property, Universities of Hunan Province; Hunan Provincial Key Laboratory of Big Data Science and Technology, Finance and Economics; Key Laboratory of Information Technology and Security, Hunan Provincial Higher Education. This research is funded by the Open Foundation for the University Innovation Platform in the Hunan Province (Grant No. 18K103); Open Project (Grant Nos. 20181901CRP03, 20181901CRP04, 20181901CRP05); Hunan Provincial Education Science 13th Five Year Plan (Grant No. XJK016BXX001), Social Science Foundation of Hunan Province (Grant No. 17YBA049).

Funding Statement: This paper partly supported by the project 18K103.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] G. S. Gao, "Research review of user portrait construction methods," *Data Analysis and Knowledge Discovery*, vol. 3, no. 3, pp. 25-35, 2019.

[2] X. G. Shi, "Research and application of user portrait in user value enhancement," *Mobile Communications*, vol. 43, no. 4, pp. 70-74, 2019.

[3] S. M. Park and Y. G. Kim, "User profile system based on sentiment analysis for mobile edge computing," *Computers, Materials & Continua*, vol. 62, no. 2, pp. 569-590, 2020.

[4] P. J. Zhu, "Research on gender prediction based on user behavior characteristics," *Computer Knowledge and Technology*, vol. 14, no. 2, pp. 158-160, 2018.

[5] C. P. Hu, Q. G. Shao and F. Y. Zhang, "User preference and Behavior analysis in personalized information service," *Science and Technology Information*, vol. 31, no. 1, pp. 4-6, 2008.

[6] Liu, Sun, Yi Su et al. "Literature review of persona at home and abroad," *Intelligence Theory and Practice*, vol. 41, no. 11, pp. 155-160, 2018.

[7] P. S. Zhu, T. Y. Qian and M. Q. Wu, "Identifying users' gender via social representations," *Computer Science*, vol. 44, no. 11A, pp. 160-165, 2017.

[8] D. P. Zheng, Z. H. Jiang and Q. Zhang, "Analysis of drivers' risky driving behavior and its influencing factors," *Chinese Journal of Ergonomics*, vol. 20, no. 1, pp. 20-25, 2014.

[9] Q. He, X. J. Zhu and M. N. Wan, "Association of Accident Risk with Demographic Factors and Environment Variables for Male and Female Drivers," *Chinese Journal of Ergonomics*, vol. 25, no. 1, pp. 31-35+51, 2019.

[10] Zhang, Deng and Lin, "Analysis of influencing factors of car accidents based on drivers' gender," *Safety and Environmental Engineering*, vol. 26, no. 3, pp. 166-170, 2019.

[11] X. B. Yang, J. Zhang and M. Li, "Decision tree algorithm and its core technology," *Computer Technology and Development*, vol. 17, no. 1, pp. 43-45, 2007.

[12] S. L. Han, H. Zhang and H. P. Zhou, "Decision tree classification algorithm based on correlation degree function," *Journal of Computer Applications*, vol. 25, no. 11, pp. 2655-2657, 2005.

[13] A. S. Koyuncugil1 and N. Ozgulbas, "Risk modeling by CHAID decision tree algorithm," *The International Conference on Computational & Experimental Engineering and Sciences*, vol. 11, no. 2, pp. 39-46, 2009.

[14] M. H. Shao, "Research review of decision tree typical algorithm," *Computer Knowledge and Technology*, vol. 14, no. 8, pp. 175-177, 2018.

[15] A. Maha, "A comparative study of machine learning methods for genre identification of classical arabic text," *Computers, Materials & Continua*, vol. 60, no. 2, pp. 421-433, 2019.

[16] B. Liu and J. C. Liu, "Modeling practice of forecasting electricity fees sensitive customers based on user profile analysis," *Power Systems and Big Data*, vol. 20, no. 8, pp. 20-24+19, 2017.

[17] B. Huang and G. Yu, "Research and application of public opinion retrieval based on user behavior modeling," *Neurocomputing*, vol. 167.no. 1, pp. 1596-603, 2015.

[18] M. Li, D. Mo and S. Lyu, "Using machine learning methods in the simulation of heat transfer and fluid flow: a brief review," in *The Int. Conf. on Computational & Experimental Engineering and Sciences*, vol. 22, no. 3, pp. 165-165, 2019.

[19] N. Yang, Y. Qian, S. Hany, R. Zhang and A. Wang, "Rapid detection of rice disease using microscopy image identification based on the synergistic judgment of texture and shape features and decision tree-confusion matrix method," *Journal of the Science of Food and Agriculture*, vol. 99, no. 14, pp. 6589-6600, 2019.

[20] D. Zhu, Y. Wang, C. You, J. Qiu and N. Cao, "MMLUP: Multi-Source & Multi-Task learning for user profiles in social network," *Computers, Materials & Continua*, vol. 61, no. 3, pp. 1105-1115, 2019.

[21] M. Q. Song, Y. Chen and R. Zhang, "User portrait research review," *Information Science*, vol. 37, no. 4, pp. 171-177, 2019.