Research on The Model of UBI Car Insurance Rates Rating Based on CNN-Softmax Algorithm

Meijie Li1, Yanqi Sun2, Xindong Wang2 and Yongkui Shi1, *

1College of Energy and Mining Engineering of Shandong University of Science and Technology, Qingdao, China
2College of Mathematics and System Science of Shandong University of Science and Technology, Qingdao, China

*Corresponding author: shiyongkui@sdu.edu.cn

Abstract. With the support of vehicular networking technology, the rating of UBI auto insurance rate has a certain guiding significance for accurate pricing of the rate and personalized demand. Based on CNN-Softmax algorithm, a rating model of UBI auto insurance rate is proposed in this paper. The model first performs a series of operations such as convolution, pooling, and non-linear activation function mapping on CNN algorithm to enable it to extract features based on UBI customers' driving behavior data, and then uses Softmax algorithm to classify customers according to their driving behaviors, thereby obtain UBI customer's auto insurance rate rating. The empirical results show that compared with, CNN algorithm, BP neural network algorithm and SVM algorithm, CNN-Softmax algorithm has a higher discrimination accuracy in the risk assessment of UBI customers' driving behavior. And it is easy to implement in establishing rating model of UBI auto insurance rate. What's more, it also has a good robustness, which can achieve better grade evaluations.

Keywords: UBI car insurance, Rating of rates, CNN-Softmax algorithm, CNN algorithm.

1. Introduction
With the development of the automobile industry, China's total vehicle volume is expected to become the world's largest by 2020, when a huge market for auto insurance and related industries will be formed [1]. The maturity of emerging Internet technologies, such as Internet of vehicles, big data and blockchain, will bring new opportunities and challenges to China's traditional auto insurance and related industries [2]. Using these technologies to deal with problems in the industry is also of innovative and practical value [3]. China's traditional auto insurance rate is determined according to the vehicle itself, mainly through the age of the vehicle, vehicle use, purchase price and other attributes to calculate the basic premium price of the vehicle, and then according to the number of vehicle accidents to give customers the corresponding discount. As the gradual maturity of related technologies and lower technical cost of information capture and collection, personalized pricing of auto insurance has become the new demand, thus giving birth to the development of UBI auto insurance model.
The traditional car-based insurance premium calculation method has not been able to provide car owners with a reasonable premium price. A survey found that applicants had frequent auto accidents accounted for only 17.5% of the total, while the remaining 82.5% with relatively good driving behavior had to bear unreasonably high premiums [4]. This is partly because China's car types are relatively single, which caused a little difference in premiums between vehicles. In addition, the traditional pricing of auto insurance premiums is determined by various factors such as vehicle accident rate, total mileage and the purchase price. Therefore, insurance fraud like forging accident scenes, maliciously changing the odometer and delaying insurance reporting cannot be avoided. Based on this situation, many scholars at home and abroad had proposed many improvements to the traditional auto insurance rate rating. Charity Mkajuma Wamwea et al. (2015) claimed that applicants with different risk levels and driving mileage had to pay the same premium was not rigorous from an actuarial perspective. Users should pay the premium according to their specific circumstances. His research believed that the zero-inflated negative binomial model was the most reasonable one, which proved the rationality and feasibility of UBI auto insurance pricing [5]. With the development of Internet technology, the concept of big data is gradually penetrating into various industries, and UBI auto insurance is a typical case of big data technology adapting to the insurance industry. Hunsnjak Slvan Forenbachera et al. (2015) believed that in addition to the existing factors, the driving behavior of users should also be included in the premium calculation [6]. Guo et al. (2017) used driving data to study the accident risk rates of drivers with different driving ages, mainly focusing on micro-driving behaviors that led to driver distraction [7].

The problem of auto insurance rates in the market is a hot topic that domestic scholars studying and discussing. Jin Ying (2012) pointed out that these issues should be addressed from the perspective of supply and demand, supervision, and proposed related solutions [8]. Yan Chun (2017) believed that frequent auto insurance frauds disrupted the operating order of the auto insurance industry in China to a certain extent and hindered the marketization process of the auto insurance industry [9].

In recent years, the CNN-Softmax model has been a widely used and mature technology in many fields such as speech recognition, face recognition, as well as natural language processing, and has made great breakthroughs. G. wimmer (2016) applied the CNN-Softmax algorithm to the automatic diagnosis of celiac disease, and achieved good classification results in the processing of duodenal endoscopy images of celiac disease [10]. J. wang (2018) proposed a CNN algorithm for feature extraction of human facial features, and then classify face images by Softmax algorithm, so as to realize face recognition. Experiments on data sets LFW and MegaFace show that this kind of algorithm has good performance [11]. M.Y.Jiang (2018) proposed a hybrid model of text classification algorithm based on DBN and Softmax algorithm, with DBN algorithm firstly used for feature extraction, and then Softmax regression to study text classification in the feature space. Experimental results on the corpus of reuters-21, 578 and 20-newsgroup show that the model can converge in the fine-tuning stage, and its performance is significantly better than classical algorithms such as SVM and KNN [12]. L.F.Chen (2018) proposed a deep sparse self-coding network (SRDSAN) based on Softmax regression for face emotion recognition in human-computer interaction. This method makes use of the characteristics of neural network and makes the overall depth of neural network more robust by fine-tuning the weights of neural network, thus improving the performance of facial emotion recognition [13].

With the rise of big data and artificial intelligence, machine learning algorithms have gained wide attention, among which Softmax regression model has been widely used in a variety of fields. At present, the deep learning algorithm is more and more recognized for its efficiency and accuracy compared with the traditional model. Therefore, this paper proposes to use CNN algorithm for feature extraction and then classify the data by Softmax algorithm. This model can effectively solve the problem of determining the risk level of driving behavior of UBI customers. CNN algorithm and Softmax algorithm combining model has not been applied in the field of UBI car insurance rates set, but the combination of these two kinds of machine learning methods is feasible: The CNN model can effectively extract the sample features of UBI auto insurance users through convolution and pooling,
and pass the feature vectors to the Softmax model for classification operation, so as to determine the premium rate of UBI auto insurance users. The combination of the two models can effectively solve the problems of overfitting and local optimization. At the same time, the 10-fold cross-validation method is used to process the data in the empirical study, which reduces the sensitivity of the model to the results of data set partitioning, and makes the model have better robustness, so as to achieve better judgment effect.

2. Introduction of relevant models

For realizing the rating of UBI auto insurance rate level, data need to be processed through a classifier. Some common classifiers include CNN algorithm, SVM algorithm, and BP neural network. This paper respectively uses CNN algorithm and the CNN-SVM algorithm to perform level prediction on the same test sample, and compares the two prediction results in precision. The following is an introduction to the principles of related algorithms.

2.1. CNN Algorithm

CNN algorithms include two basic types of processes: feedforward operations and feedback operations. Feedforward calculation is a process of prediction and reasoning. It stacks a series of operations such as convolution (corresponding to the convolution layer), pooling (corresponding to the pooling layer), and non-linear activation function mapping to form a feature extraction network, extracts high-level semantic information from the original data input layer step by step, and abstracts it into a feature map. Then the last layer of the convolutional neural network formalizes its target task (classification, regression, etc.) into an objective function, and finally outputs the classification and regression results.

The input of a convolutional neural network is usually the original image $X$. This paper uses $H_i$ to represent the $i$-th layer's feature map of convolutional neural network ($H_0 = X$). Assuming $H_i$ is a convolutional layer, the generation process of it can be described as:

$$H_i = f(H_{i-1} \otimes W_i + b_i)$$  \hspace{1cm} (1)

$W_i$ represents the weight vector of the $i$-th layer convolution kernel; the operation symbol “$\otimes$” represents convolution operation between the convolution kernel and the $i-1$th layer image or feature map, and the output of the convolution is added to the offset vector $b_i$ of the $i$-th layer, then the feature map $H_i$ of the $i$-th layer is obtained through the nonlinear excitation function $f(x)$.

Downsampling layer usually follows convolutional layer. The feature map is down-sampled according to a certain down-sampling rule. This step mainly has the following two functions:

1) Dimension reduction of feature maps;

2) Maintain the scale-invariant characteristics of the features to a certain extent.

Suppose $G_i$ is the downsampling layer:

$$G_i = subsampling(G_{i-1})$$  \hspace{1cm} (2)

After the alternate transfer of multiple convolutional layers and down-sampling layers, the convolutional neural network relies on a fully connected network to classify the extracted features and obtain an input-based probability distribution $\hat{Y}$ ($l_i$ represents the $i$-th label category). As shown in the formula, the convolutional neural network is essentially a mathematical model that transforms the original matrix ($G_0$) through multiple levels of data transformation or dimensionality reduction to a new feature expression ($\hat{Y}$).
The training goal of a convolutional neural network is to minimize the loss function \( L(W,b) \) of the network. The difference between the input \( G_0 \) and the expected value calculated by the loss function after forward conduction is called "residual error". Common loss functions include Mean Squared Error (MSE) functions, Negative Log Likelihood (NLL) functions, and so on.

To alleviate overfitting, the final loss function usually controls the overfitting of the weights by increasing \( L_2 \) norm, and the strength of the overfitting effect is controlled by the parameter \( \lambda \) (weight decay):

\[
E(W,b) = L(W,b) + \frac{\lambda}{2} W^T W
\]

During training, gradient descent is the commonly used optimization method for convolutional neural networks. Residuals are back-propagated through gradient descent, and trainable parameters \((W,b)\) of each layer in convolutional neural network are updated layer by layer. Learning rate parameter \((\eta)\) is used to control the strength of residual backpropagation:

\[
W_{i+1} = W_i - \eta \frac{\partial E(W,b)}{\partial W_i}
\]

\[
b_{i+1} = b_i - \eta \frac{\partial E(W,b)}{\partial b_i}
\]

2.2. Softmax algorithm principle

Softmax logistic regression model is a generalization of logistic regression model on multi-classification problems, in which the class label \( y \) can take more than two values. Softmax regression is the general form of Logistic regression, while Logistic regression is the special form of Softmax regression when \( k=2 \).

There are \( k \) categories of input data \( \{(x_1,y_1),(x_2,y_2),\ldots,(x_m,y_m)\} \), namely \( y_i \in \{1,2,\ldots,k\} \), so Softmax regression mainly estimates the probability that input data \( x_i \) belongs to each category, that is:

\[
h_{\theta}(x_i) = \left[ \begin{array}{c} p(y_i = 1|x_i; \theta) \\ p(y_i = 2|x_i; \theta) \\ \vdots \\ p(y_i = k|x_i; \theta) \end{array} \right] = \left[ \frac{1}{\sum_{j=1}^{k} e^{\theta_j^T x_i}} e^{\theta_1^T x_i} e^{\theta_2^T x_i} \ldots e^{\theta_k^T x_i} \right]
\]

Where \( \theta_1, \theta_2, \ldots, \theta_k \in \Theta \) is the parameter of the model, and \( \frac{1}{\sum_{j=1}^{k} e^{\theta_j^T x_i}} \) is multiplied to make the probability between [0,1] and the sum of probabilities 1. The probability that Softmax regression attributes the input data \( x_i \) to category \( j \) is:
\[ p(y_i = j | x_i; \theta) = \frac{e^{\theta_j^T x_i}}{\sum_{l=1}^k e^{\theta_l^T x_i}} \]  

The parameter matrix \( \theta \) of Softmax regression can be denoted as:

\[
\theta = \begin{bmatrix}
\theta_1^T \\
\theta_2^T \\
\vdots \\
\theta_k^T 
\end{bmatrix}
\]

The cost function of Softmax regression is defined as:

\[
L(\theta) = -\frac{1}{m} \left[ \sum_{i=1}^m \sum_{j=1}^k 1\{y_i = j\} \log \frac{e^{\theta_j^T x_i}}{\sum_{l=1}^k e^{\theta_l^T x_i}} \right]
\]

Where, \( 1\{\cdot\} \) is the indicator function, that is, \( 1\{\text{value is true expression}\} = 1 \), \( 1\{\text{value is false expression}\} = 0 \). Using the gradient descent method to minimize the cost function, let's solve for the gradient of \( \theta \). The gradient of \( L(\theta) \) with respect to \( \theta_j \) is solved as follows:

\[
\frac{\partial L(\theta)}{\partial \theta_j} = -\frac{1}{m} \frac{\partial}{\partial \theta_j} \left[ \sum_{i=1}^m \sum_{j=1}^k 1\{y_i = j\} \log \frac{e^{\theta_j^T x_i}}{\sum_{l=1}^k e^{\theta_l^T x_i}} \right]
\]

\[
= -\frac{1}{m} \frac{\partial}{\partial \theta_j} \left[ \sum_{i=1}^m 1\{y_i = j\} \left( x_i - \sum_{l=1}^k \frac{e^{\theta_l^T x_i}}{\sum_{l=1}^k e^{\theta_l^T x_i}} \right) \right]
\]

\[
= -\frac{1}{m} \sum_{i=1}^m x_i 1\{y_i = j\} \left( 1 - \sum_{l=1}^k \frac{e^{\theta_l^T x_i}}{\sum_{l=1}^k e^{\theta_l^T x_i}} \right)
\]

2.3. CNN-Softmax model construction

According to its main functions, the CNN model can be divided into feature extraction part and classification part. The feature extraction part includes convolution layer and pooling layer, and the
The classification part includes full connection layer. The input-based probability distribution can be obtained at the full connection layer, but it is prone to overfitting and local optimization. However, Softmax model can effectively avoid the overfitting problem by introducing a regular term (12) after the loss function, and obtained the globally optimal solution.

Regular term:

$$L(\theta) = -\frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{k} 1(y_i = j) \log \left( \frac{e^{\theta_j x_i}}{\sum_{j=1}^{k} e^{\theta_j x_i}} \right) + \lambda \sum_{i=1}^{m} \sum_{j=1}^{n} \theta_{ij}^2$$

The gradient of the new loss function:

$$\frac{\partial L(\theta)}{\partial \theta_j} = -\frac{1}{m} \sum_{i=1}^{m} x_i (1(y_i = j) - p(y_i = j|x_i; \theta)) + \lambda \theta_j$$

Therefore, the full connection layer of the CNN model was replaced by the Softmax classifier, and the CNN-Softmax model was constructed to solve the problem of determining the UBI auto insurance rate grade. The CNN model was trained to make the accuracy of the training set stable. Then the high dimensional feature representation of the test set obtained by the sampling layer was extracted and used as the input of the Softmax model. In the model, the CNN model is used to extract features, and the Softmax model is used as a multi-classifier for prediction classification.

The specific operation steps of CNN-Softmax model are as follows:

1. Pretreatment of sample data;
2. Set parameters and layers of CNN model and Softmax model, and initialize the model;
3. Convert the panel data of the training set into a matrix and input it into the CNN model as an input vector; input the feature vector output from the CNN model into Softmax model as the input vector; train model parameters with training data;
4. Convert the panel data of the test set into a matrix, and take it into the well-trained CNN-Softmax model to output multi-classification results; predict the user's level according to the output of each user's probability vector, and finally compare them with the actual level of the customer, observe the effect of classification.

3. Index selection and data preprocessing of UBI car insurance rate rating mode

3.1. Index selection

According to the entropy weight analytic hierarchy method proposed by Zhu Shuang (2015), this paper obtains the weights of multiple risk factors in the cause of the accident. Among them, the time factor with high weight risk factor higher than 120km/h, the number of rapid accelerations, the mileage per month, and the ratio of driving time at night were selected as the experimental sample indicators. The specific weights of the risk factors are shown in Table 1.

| Risk factor                    | Weights derived from entropy weight analytic hierarchy process |
|-------------------------------|-------------------------------------------------------------|
| Mileage per month             | 0.1483                                                      |
| Time ratio higher than 120km/h | 0.5042                                                      |
| Acceleration times            | 0.01452                                                     |
| Night driving time ratio      | 0.2023                                                      |
Note: Data source: Zhu Shuang. UBI-based automobile insurance rate determination model and method research in the context of connected cars [D]. Beijing Jiaotong University, 2015.

3.2. Data preprocessing
Hierarchical scoring of driving behavior is the basis for effective rating of UBI auto insurance rates. The CNN-SVM algorithm can be used to match similar driving behaviors among a large number of historical data groups, thereby realizing the evaluation of user rates. The driving behavior data in this article comes from "Internet Auto Insurance UBI Product Design" proposed by Jiang Lei (2017), which collected four indicator data, the monthly driving mileage, night driving time ratio, time ratio higher than 120km/h, and rapid acceleration behavior. The data consists of two parts, the panel data of indicator system and the customer's driving behavior level. Some examples of driving behavior data are shown in Table 2.

### Tab. 2 Examples of part driving behavior data

| user | Mileage per month (KM) | Proportion of time above 120km/h (%) | Rapid acceleration times (times) | Night driving time (h) grade |
|------|------------------------|-------------------------------------|---------------------------------|-----------------------------|
| 1    | 84                     | 0                                   | 0                               | 0 A                         |
| 2    | 362                    | 0                                   | 1                               | 0.9 A                       |
| 3    | 1137                   | 0.66                                | 1                               | 0 B                         |
| 4    | 870                    | 0.48                                | 4                               | 0 B                         |
| 5    | 754                    | 1.02                                | 2                               | 1.1 C                       |
| 6    | 1180                   | 0.63                                | 3                               | 4.3 C                       |
| 7    | 1240                   | 1.24                                | 13                              | 6 D                         |
| 8    | 574                    | 3.04                                | 6                               | 1.6 D                       |
| 9    | 1710                   | 0.95                                | 34                              | 7.3 E                       |
| 10   | 1185                   | 4.86                                | 4                               | 6.1 E                       |

Note: Data source: Jiang Lei. UBI product design for Internet auto insurance [D]. Zhejiang University, 2017.

The customer's driving behavior level is represented by A, B, C, D, and E, corresponding to five driving behavior risk levels: extremely low, low, medium, high, and extremely high. The matrix representation method of the i-th level in model is to replace the element at the corresponding position of matrix $A = (0,0,0,0,0,0,0,0,0,0,0,0,0,0)$ with 1, for example, for class B, $A = (0,1,0,0,0,0,0,0,0,0,0,0,0,0)$.

Dimensional differences exist among the indicators of the indicator system, so normalization processing is needed. The specific normalization operation is realized by the $\text{mapminmax}$ function, and the realized formula is shown in formula (14):

$$y = \frac{(y_{\text{max}} - y_{\text{min}}) \times (x - x_{\text{min}})}{x_{\text{max}} - x_{\text{min}}} + y_{\text{min}}$$

Among them, $x$ in the formula is the feature vector to be normalized, $x_{\text{max}}$ and $x_{\text{min}}$ respectively represent the largest and smallest vectors in $x$, and $y_{\text{max}}$ and $y_{\text{min}}$ correspond to the maximum and minimum values in the interval range after the data is normalized. At this time, all the values in the original individual stock characteristic data are transformed into the interval $[0,1]$. Such processing avoids dimensional differences between different feature attributes without affecting the internal relationship of the same attribute value.
Due to the randomness of the training and test sets, this paper changed from inputting one set to multiple sets for reducing error. Known from the large number theorem, the more trials, the more stable the probability of the event. In order to make the final prediction accuracy converge and the model easy, this paper repeats the process of randomly dividing the training set and test set for 10 times to get 10 different data sets, namely a 10-fold cross-check. From this, the prediction results of these independent and different data subsets are obtained, and the average results of the 10 sub-tests is obtained.

10 groups of data sets are input into the CNN-Softmax model using a circular algorithm. Each set is subjected to convolution operations and pooling operations so as to extract data features and then pass them to Softmax model. Softmax model performs multi-classification processing and then obtains the classification results. At the same time, the classification accuracy of each group is also obtained. Repeating this process for ten times, then the average value is obtained and the accuracy rate of the UBI auto insurance premium rate rating based on the CNN-Softmax algorithm is obtained either. The algorithm is implemented in Python 3.6 (64-bit).

4. Empirical analysis

4.1. Effect analysis of the UBI auto insurance premium rating model based on CNN algorithm

CNN model selects piecewise linear activation function reLU and nonlinear activation function sigmoid, which are commonly used in multi-classification problems, to observe the effects. The accuracy of the predicted value is calculated by a convolutional neural network and is shown in Table 3. As can be seen from the table, the accuracy rate of the CNN model using the reLU activation function is 82.9% and the Sigmoid activation function is 79.1%. Obviously, the prediction effect using the reLU activation function is better.

| Data set | Accuracy under reLU activation function | Accuracy under reLU activation function |
|----------|---------------------------------------|---------------------------------------|
| 1        | 0.834                                 | 0.784                                 |
| 2        | 0.828                                 | 0.780                                 |
| 3        | 0.831                                 | 0.781                                 |
| 4        | 0.838                                 | 0.809                                 |
| 5        | 0.831                                 | 0.792                                 |
| 6        | 0.831                                 | 0.797                                 |
| 7        | 0.829                                 | 0.780                                 |
| 8        | 0.804                                 | 0.798                                 |
| 9        | 0.836                                 | 0.787                                 |
| 10       | 0.827                                 | 0.797                                 |
| Average  | 0.829                                 | 0.791                                 |

Observe the effect of iterations on the prediction accuracy of the convolutional neural network. Table 4 lists the prediction accuracy of the convolutional neural network under different iterations when the reLU activation function is used.
It can be seen from Table 4 that as the number of iterations increases from 10 to 1000, the prediction accuracy of the convolutional neural network continues to increase from 0.491 to 0.829, while the operating efficiency is greatly reducing. In order to take into account both accuracy and operating efficiency, this paper selects 1,000 iterations, where the accuracy is 82.9%.

### 4.2. Effect analysis of UBI auto insurance premium rate rating model based on CNN-Softmax algorithm

As can be seen from the previous section, when iterations of the convolutional neural network are 1,000, the model has the best accuracy, feature extraction effect and the highest efficiency, therefore the iteration is still 1000. Softmax uses a one-to-many classification method. When the reLU function is used for convolutional neural network, the accuracy obtained from the Softmax using linear activation function and rbf function is shown in Table 5. It can be seen that the model using rbf activation function has a higher prediction classification accuracy value of 86.4%.

### Tab. 5 CNN-Softmax model prediction accuracy when CNN activation function is reLU

| Data set | Accuracy under Linear activation function | Accuracy under rbf activation function |
|----------|-------------------------------------------|---------------------------------------|
| 1        | 0.856                                     | 0.872                                 |
| 2        | 0.854                                     | 0.853                                 |
| 3        | 0.860                                     | 0.873                                 |
| 4        | 0.851                                     | 0.865                                 |
| 5        | 0.848                                     | 0.861                                 |
| 6        | 0.849                                     | 0.866                                 |
| 7        | 0.855                                     | 0.865                                 |
| 8        | 0.850                                     | 0.864                                 |
| 9        | 0.853                                     | 0.868                                 |
| 10       | 0.851                                     | 0.853                                 |
| Average accuracy | 0.853                                 | 0.864                                 |

When sigmoid activation function is used, the Softmax also respectively uses linear function and the reLU function. The accuracy results obtained are shown in Table 6. It can be seen that the accuracy of the two activation functions and that of convolutional neural network using only the reLU activation function are 80.9% and 83.2%, proving that the former is less effective.
Tab. 6 CNN-Softmax model prediction accuracy when CNN activation function is sigmoid

| Data set | Accuracy under Linear activation function | Accuracy under rbf activation function |
|----------|------------------------------------------|----------------------------------------|
| 1        | 0.801                                    | 0.83                                   |
| 2        | 0.813                                    | 0.838                                  |
| 3        | 0.808                                    | 0.839                                  |
| 4        | 0.818                                    | 0.82                                   |
| 5        | 0.805                                    | 0.837                                  |

| Data set | Accuracy under Linear activation function | Accuracy under rbf activation function |
|----------|------------------------------------------|----------------------------------------|
| 6        | 0.81                                     | 0.835                                  |
| 7        | 0.813                                    | 0.838                                  |
| 8        | 0.801                                    | 0.828                                  |
| 9        | 0.819                                    | 0.824                                  |
| 10       | 0.801                                    | 0.834                                  |
| Average accuracy | 0.809                                   | 0.832                                  |

Based on the above experiments, when CNN using ReLU activation function and Softmax using the rbf activation function, it has the highest prediction accuracy rate of 86.4%. In this model, some driving behavior data of UBI users (seen in Table 2) are predicted, and the results are shown in Table 7.

Tab. 7 Example of driving behavior prediction results for some UBI users based on CNN-Softmax algorithm

| User | Monthly mileage (KM) | Proportion of time above 120km / h (%) | Rapid acceleration times (times) | Night driving time (h) | Grade |
|------|----------------------|----------------------------------------|---------------------------------|------------------------|-------|
| 1    | 84                   | 0                                      | 0                               | 0                      | A     |
| 2    | 362                  | 0                                      | 1                               | 0.9                    | A     |
| 3    | 1137                 | 0.66                                   | 1                               | 0                      | B     |
| 4    | 870                  | 0.48                                   | 4                               | 0                      | B     |
| 5    | 754                  | 1.02                                   | 2                               | 1.1                    | C     |
| 6    | 1180                 | 0.63                                   | 3                               | 4.3                    | C     |
| 7    | 1240                 | 1.24                                   | 13                              | 6                      | D     |
| 8    | 574                  | 3.04                                   | 6                               | 1.6                    | E     |
| 9    | 1710                 | 0.95                                   | 34                              | 7.3                    | E     |
| 10   | 1185                 | 4.86                                   | 4                               | 6.1                    | E     |

Except for the prediction result of user 5, the other prediction results are consistent with the true results given in Table 2. The accuracy rate of all user level predictions is 85.3%, which is close to the accuracy rate of 86.4% obtained by the above experiments, and the experimental error is within a reasonable range.

4.3. Results comparison of CNN-Softmax algorithm and CNN algorithm

The most commonly used criterion for evaluating the quality of a classifier is accuracy. The higher the accuracy, the better the classification effect. This paper mainly compares the CNN algorithm with the CNN-Softmax algorithm. In order to make the experimental results more convincing, BP neural
network algorithm, SVM algorithm and CNN-SVM algorithm are also introduced for auxiliary comparison. After a 10-fold cross-check, the accuracy of each algorithm's decision on user's driving behavior is shown in Table 8.

| Data set | CNN  | BP   | CNN-SVM | SVM  | CNN-Softmax |
|----------|------|------|---------|------|-------------|
| 1        | 0.834| 0.786| 0.87    | 0.796| 0.872       |
| 2        | 0.828| 0.773| 0.86    | 0.785| 0.853       |
| 3        | 0.831| 0.763| 0.866   | 0.795| 0.873       |
| 4        | 0.838| 0.777| 0.87    | 0.801| 0.865       |
| 5        | 0.831| 0.766| 0.866   | 0.787| 0.861       |
| 6        | 0.831| 0.770| 0.865   | 0.806| 0.866       |
| 7        | 0.829| 0.789| 0.851   | 0.800| 0.865       |
| 8        | 0.804| 0.750| 0.852   | 0.796| 0.864       |
| 9        | 0.836| 0.771| 0.854   | 0.788| 0.868       |
| 10       | 0.830| 0.772| 0.854   | 0.797| 0.853       |
| Average accuracy | 0.861 | 0.772 | 0.861   | 0.795 | 0.864       |

Experiments show that the average classification accuracy of CNN algorithm and CNN-Softmax algorithm is 82.9% and 86.4% respectively, which means CNN-Softmax is more suitable for UBI auto insurance rate rating for its higher accuracy rate. Besides, the 10-fold cross-test used in this paper has a smaller variance, therefore the model is more robust.

5. Conclusions
Based on CNN-Softmax algorithm, a rating model of UBI auto insurance rate is proposed in this paper. The model has both CNN's superior feature extraction ability and Softmax's stable classification ability, which could prevent over-fitting and local optimization, and therefore effectively achieve the rating of UBI auto insurance rate. It uses a 10-fold cross-validation method to process the data set for reducing the sensitivity of the model to the results of the data set partition, allows each sample having the possibility to become a training set or a test set, and makes the model more robust. At the same time, it also verifies the rationality of CNN-Softmax model in processing the feature extraction of the raw data in insurance index system and the multi-classification of insurance samples. During rating the UBI auto insurance rate, CNN-Softmax algorithm has the following characteristics: the model accuracy rate is significantly improved compared with CNN-SVM algorithm, CNN algorithm and SVM algorithm; the model prediction efficiency is higher known from the number of iterations. Therefore, CNN-Softmax model has a more accurate rating prediction of UBI auto insurance users, and the results are in line with the actual situation, which means a strong applicability and flexibility. With the maturity of connected vehicle technology, more collected data and established risk factors could be used to a better description of driving behavior. A better research environment could be promoted by optimized models and enrich data in the future.

6. References
[1] [Estrin D. Participatory sensing: applications and architecture [J]. Internet Computing, 2000, 14(1): 12-42.
[2] Jiangsu Nanyi Dina Digital Technology Co., Ltd. White paper about vehicular insurance in IOV [DB / OL].2015-03-06.
[3] Su Yuyuan. Analysis on the realization method of big data knowledge [J]. Journal of Shandong University of Science and Technology, 2019, 21 (1): 20-26.
[4] Jiangsu Nanyi Dina Digital Technology Co., Ltd. 4S Group's white paper on car networking solution [DB / OL].
[5] Charity Mkajuma Wamwea, Benjamin Kyalo Muema, Joseph Kyalo Mung'atu, et al. Modelling a Pay-As-You-Drive Insurance Pricing Structure Using a Generalized Linear Model Case study of a company in Kiambu [J]. American Journal of Theoretical and Applied Statistic, 2015, 4(6):527-533.

[6] Husnjak S, Perakovic D, Forenbacher I, et al. Telematics System in Usage Based Motor Insurance [J]. Procedia Engineering, 2015, 100:816-825.

[7] Guo F, Kim I, Klauer S G. Semiparametric Bayesian models for evaluating time (ariant driving risk factors using naturalistic driving data and case rossover approach [J]. Statistics in Medicine 2017(1).

[8] Jin Ying, Tian Xiaoli. Analysis of marketization of auto insurance rates from the perspective of supply and demand [J]. Cooperative Economy and Science and Technology, 2012 (1): 76-77.

[9] Yan Chun, Li Yaqi, Sun Haitang. Research on Auto Insurance Fraud Recognition Based on Ant Colony Algorithm to Optimize Random Forest Model [J]. Insurance Research, 2017 (6): 114-127.

[10] G. Wmmer. CNN transfer learning for the automated diagnosis of celiac disease [J]. IEEE, 10.1109/IPTA.2016.7821020.

[11] Wang, J. Cheng and W. Y. Liu. Additive Margin Softmax for Face Verification [J]. IEEE Signal Processing Letters, 2018, 25(7):926-930.

[12] M. Y. Jiang, Y. C. Liang. Text classification based on deep belief network and Softmax regression [J]. Neural Computing and Applications, 2018, 29:61–70.

[13] L. F. Chen, M. T. Zhou and W. J. Su. Softmax regression based deep sparse autoencoder network for facial emotion recognition in human-robot interaction [J]. Information Sciences, 2018, 428: 49–61.