Research on the UBI Car Insurance Rate Determination Model Based on the CNN-HVSVM Algorithm

CHUN YAN\textsuperscript{1}, XINDONG WANG\textsuperscript{1}, XINHONG LIU\textsuperscript{2}, WEI LIU\textsuperscript{3}, (Member, IEEE), AND JIAHUI LIU\textsuperscript{1}

\textsuperscript{1}College of Mathematics and System Sciences, Shandong University of Science and Technology, Qingdao 266590, China
\textsuperscript{2}Department of Mathematics and Physics, Beijing Institute of Petro-chemical Technology, Beijing 102617, China
\textsuperscript{3}College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao 266590, China

Corresponding authors: Xinhong Liu (10285653@qq.com), Xindong Wang (937215471@qq.com), and Wei Liu (liuwei_doctor@yeah.net)

This work was supported in part by the General project of science and technology plan of Beijing Municipal Commission of Education under Grant KM202010017001, in part by the China Statistical Science Research under Grant 2019LZ10, and in part by the National Natural Science Foundation of China under Grant 61502280.

ABSTRACT With the support of Internet of Vehicles technology, UBI (Usage Based Insurance) car insurance rate determination has certain guiding significance for achieving accurate pricing of car insurance rates and satisfying the personalized needs of users. Based on the CNN (Convolutional Neural Networks) algorithm and SVM (Support Vector Machine) algorithm, this paper establishes a rating model for UBI car insurance rates. The model first performs a series of operations such as convolutions, pooling and nonlinear activation function mapping using the CNN algorithm so that it can extract the features from the driving behavior data of UBI customers. Then, it introduces the Hull Vector to optimize the operating efficiency of the SVM algorithm. The HVSVM (Hull Vector Support Vector Machine) algorithm classifies customers according to their driving behavior, and thus obtains UBI customer car insurance rate grades. Therefore, this paper proposes a UBI car insurance rate grade determination model based on the CNN-HVSVM algorithm. The empirical results of the model show that the CNN-HVSVM algorithm has higher discrimination accuracy in the risk rating process of UBI customer driving behavior than the CNN algorithm, BP neural network algorithm and SVM algorithm; and when dealing with large training sets, it has a speed advantage over the CNN-SVM algorithm. Furthermore, it is easy to realize in the process of establishing the UBI car insurance rate determination model and it has good robustness, which can adapt to diverse data sets, thus achieving better results in the car insurance rate determination process. Therefore, the CNN-HVSVM model can predict the grade of UBI car insurance users more accurately and efficiently, and the prediction results are more consistent with the actual situation, which has strong applicability and flexibility. The UBI car insurance premium rate determination model based on the CNN-HVSVM algorithm can determine driver behavior more fairly and reasonably, and has certain practical significance for promoting car insurance rate market reform, which can better promote future UBI research work.

INDEX TERMS CNN algorithm, CNN-HVSVM algorithm, rate grade judgment, SVM algorithm, UBI car insurance

I. INTRODUCTION

With the progress of the automobile industry, China’s total car volume is expected to become the first in the world by 2020, which will form a large car insurance and related industry market [1]. With the advent of the era of big data, using big data knowledge to deal with problems is becoming increasingly more innovative and practical in this industry [2]. The maturity of new Internet technologies such as car networking, big data, the blockchain and so on will bring new opportunities and challenges to China’s traditional car insurance and related industries. Traditional Chinese car
insurance is car-based, which mainly calculates the basic premium price of the vehicle through attributes such as vehicle age, vehicle use, purchase price, etc., and then gives the customer a corresponding discount according to the number of vehicles insured. Under the background of big data, the current car insurance rate determination method has had difficulties meeting market requirements, and reforming the car insurance rate determination method has become the general trend. With the development of related technologies such as the Internet of Vehicles, the technical costs of capturing and collecting information have been gradually reduced, and therefore the differentiated pricing of car insurance can better meet the needs of the insurance market. The development of the UBI car insurance model is the trend of the current car insurance industry.

After an investigation, it was found that only 17.5% of the total car insurance policyholders have frequent accidents while the remaining 82.5% of the car insurance policyholders with relatively good driving behavior have to bear unreasonably high insurance premiums. The traditional car-based insurance premium calculation method has not been able to provide reasonable premium prices for car owners fairly. First, China’s car models are relatively monotonous, and the way of determining vehicle insurance premiums is relatively traditional, which leads to little difference between vehicle premiums. Second, traditional car insurance claims cannot deal with the complex and changing environment, and cannot provide an accurate and effective judgment basis for insurance companies in time. The premium price is determined by various factors such as the vehicle accident rate, the total mileage of the vehicle, the age of the vehicle, and the purchase price; but there will be many unavoidable insurance frauds, such as fraudulent accident scenes, maliciously adjusting odometers, delays in reporting insurance claims, etc., which will bring many problems for claims settlement. Charity Mkajuma Wamwea et al. (2015) believe that the current traditional car insurance pricing is very unreasonable. Policyholders with different risk levels and driving mileage need to pay the same premium, which is not rigorous from the actuarial point of view. They thought that users should pay premiums according to their specific circumstances; therefore, their research determined the most reasonable compensation model, the zero inflation negative binomial model, and proved the rationality of flexible car insurance pricing, that is, the feasibility of UBI car insurance rate pricing [3]. Hunsnjak Silvan Forenbacher and others (2015) believed that the current car insurance industry only considered traditional factors such as vehicle age, displacement, and vehicle use, which is not comprehensive; and the user’s driving behavior should also be included in the premium calculation [4]. Guo et al. (2017) used driving data to study the accident risk of short- and long-term drivers, mainly focusing on the microdriving behavior that causes driver distraction [5]. M. F. Carfora et al. (2018) used real case studies to evaluate the effectiveness of UBI car insurance, and proposed a method for identifying driver behavior using unsupervised machine learning algorithms [6]. Y Bian et al. (2018) proposed that UBI car insurance can use big data services to convert premiums from annual measurements to variable costs and real-time vehicle related parameters and driving conditions can also be uploaded to the network platform in time to allow insurance companies to be more accurate and timely in finding target customers [7]. Michel Denuit et al. (2018) believe that UBI car insurance provides actuaries with behavioral risk factors, which could be included in dynamic actuarial pricing through a reputation updating mechanism [8]. Due to the exponential growth of the data flow between insurance companies and customers, Subramanian Arumugan et al. (2019) introduced the MHYD (management-how-you-drive, MHYD) model as the basis of UBI, and proposed a method to predict aggressive driving behavior by considering drivers’ daily behaviors and habitual emotions [9].

Domestic scholars have been studying and discussing the issue of car insurance rate marketization. Jin Ying (2012) pointed out that the problems in China’s car insurance rate market should be dealt with in terms of supply and demand and supervision, and proposed related solutions [10]. Yan Chun (2017) believes that frequent vehicle insurance fraud has disturbed the operating order of China’s vehicle insurance industry to a certain extent and hindered the marketization process of China’s car insurance industry [11].

In terms of the determining factors of the vehicle insurance rate, the current research in China mainly focuses on the factors of “vehicle”, “people”, “land”, etc. The selected risk factors mainly involve the driving environment, driving behavior, driving vehicles and other aspects, which are an important part of the determination of vehicle insurance rates. Xie Qinglian (2007) put forward relevant suggestions based on the data provided by the motor vehicle insurance rate factor of the people’s Insurance Company of China; the influencing factors such as vehicle, people, and land that are worth learning; and the analysis results [12]. Guofa Li (2020) examined the influence of traffic congestion on driver behavior on the post-congestion roads [13]. Zhu Shuang (2015) introduced the entropy weight analytic hierarchy process into the selection of risk factors, and extracted a number of high-weighted indicators that affect the driving behavior risk, which is of great significance for the determination of UBI car insurance rate grades [14]. Liu Ruigang (2017) established a driving behavior scoring model based on a generalized linear model and verified the feasibility of the model [15]. Liu Jian (2019) started from the construction of the car Internet UBI car insurance big data analysis model and the development of the car Internet UBI car insurance big data analysis platform, and developed a reasonable Commercial vehicle UBI car insurance rate model based on the four priorities, driving mileage, and the commercial car coefficient [16]. Guoafa Li (2017) presented a novel way to identify driving style not in terms of the durations or frequencies of individual maneuver states, but rather the transition patterns between them to see how they are interrelated [17].
Bai Yunfeng (2019), from the National Development Institute of Peking University, proposed that insurance companies should be able to focus on technological innovation and build a gold standard for the UBI behavior of high-quality users in the industry to positively encourage and guide excellent car owners to obtain better discounts [18]. Guofo Li (2019) aimed to investigate drivers’ visual scanning behavior at signalized and unsignalized intersections, which is of great significance for the determination of UBI car insurance rate grades [19].

In recent years, the CNN model has been widely used in speech recognition, face recognition, natural language processing and other fields, and has made great breakthroughs. It has become a more widely used and mature technology in these fields. Haifeng Wu (2019) proposed a CNN-SVM combined regression model to estimate the knee joint angle at different motion speeds [20]; Jian Xu (2019) built a novel improved hybrid CNN-SVM based model for bearing fault diagnosis through transforming the vibration signal into an image with two-dimensional representation [21]; Yongqiang Wang (2019) used CNN-SVM model to recognize the offline handwritten New Tai Lue characters, and the model has a good recognition rate [22]; The efficient classification of remote sensing images (RSIs) has become the key of remote sensing application. To tackle the high computational cost in the traditional classification method, Xiankun Sun (2019) proposed a new RSI classification method based on improved convolutional neural network (CNN) and support vector machine (SVM) (CNN-SVM), and obtained good experimental results [23]; Deepak B Desai (2019) verified the good performance of CNN-SVM in Face anti-spoofing Technique through experiments [24]; In many water distribution systems, a significant amount of water is lost because of leakage during transit from the water treatment plant to consumers. To solve this problem, Jiheon Kang (2017) propose a novel, fast, and accurate water leakage detection system with an adaptive design that fuses a one-dimensional CNN and SVM [25]; Rida Dong (2019) propose a novel method, CNN-SVM with Embedded Recurrent Structure, for social emotion prediction, Experimental results show that CNN-SVM outperforms the state-of-the-art methods in social emotion prediction by a significant margin [26]. Xiaolong Sun (2017) proposed a novel hybrid model that integrates the synergy of two superior classifiers for functional magnetic resonance imaging (fMRI) recognition, namely, convolutional neural networks (CNNs) and support vector machines (SVMs), both of which have proven results in the field of image recognition [27].

The SVM algorithm is widely used in the classification and prediction of different insurance actuarial fields. Many mixed models with SVM algorithms and other models have emerged to meet the needs of different insurance actuarial problems. The classification functions are mostly used in artificial intelligence, medical and Satellite remote sensing image analysis and other fields. Chun Yan (2019) established a UBI car insurance rate determination model based on the K-S algorithm by combining the KNN algorithm and the SVM algorithm. This model constructs the corresponding relationship between a driver’s behavior grade and vehicle insurance rate, and then obtains the corresponding adjustment coefficient of their insurance rate according to the driving behavior grade. This forms a more reasonable and fair measurement standard for a driver’s behavior, which has certain practical significance for promoting the reform of China’s vehicle insurance rate marketization [28]. Zhao Shangmei (2015) improved the SVM algorithm based on Hull vector and SMOTE, and compared the results with traditional SVM algorithm, BP neural network and GSA-BP neural network. The results show that the improved algorithm can obtain better prediction accuracy in the field of motor vehicle insurance fraud recognition [29]. Xue Zhiwen (2018) proposed the ARIMA-SVM combined prediction model by combining the advantages of the ARIMA model and the support vector machine (SVM) model. The empirical results showed that the improved forecasting model had a significantly improved fitting effect on the number of claims [30], Tang Shiyong (2019) analyzed the superiority of combining rough sets and the SVM algorithm, and proposed applying the RS-SVM to the performance evaluation model of a social pension insurance system [31]; To solve the problem that the classical support vector machine has difficulties carrying out incremental learning quickly and effectively, Wenbo (2013) proposed an incremental learning algorithm based on the KKT conditions and the hull vector. The experimental results show that the algorithm not only guarantees the accuracy of the learning machine and good improvement ability, but also has a faster learning speed than the classic SMO and can be used for incremental learning [32]. Mao Yanlei (2017) used the characteristics of the Hull Vector to retain nonsupport vectors that may contain hidden information in the training sample set, and only added incremental samples that violate the KKT conditions to the new training set to improve the computing efficiency [33]. Shuang Yu (2019) propose a new SVM with sample-based misclassification cost invariance with the aim of constructing a relatively reliable classifier [34]; Meng Lei (2018) propose NIR spectroscopy combined with Medium Gaussian SVM can be used as a good non-destructive method to predict origins of coal, with an accuracy rate of 98.8%, which strengthens the supervision of coal quality [35]; Yajuan Tian (2019) proposed an automatic classification method of sandstone based on support vector machine, in order to efficiently and flexibly classify the rock features extracted from thin plate images [36]; Devi KN et al. (2015) used Cuckoo search (CS) to adjust the parameters of the SVM model and form a hybrid CS- The SVM model. By comparing with the results obtained by the artificial neural network (ANN) and the SVM model, it proved that the hybrid CS-SVM technology could produce better results when predicting the price changes of financial products [37]; Kordonis J et al. (2016) used machine learning techniques such as Naïve Bayes Bernoulli classification and SVM model to analyze how tweets were related to market price behavior of financial products [38]; Plakandaras V (2017) used dynamic probability models and SVM models to analyze...
leading indicators and predict the extent of the overall US economic recession. The results show that, compared with the probability model, the SVM model has stronger recognition accuracy in a longer time range, and can better distinguish between depression and quiet periods [39]. Mukhoti S (2018) introduced leverage parameters into the SVM model to predict the returns of risky assets and their volatility over time [40]. Xiaohu Wu (2018) propose F-SVM can also be incorporated with deep convolutional networks to improve image classification performance. Experiments on the UCI, LFW, MNIST, CIFAR-10, CIFAR-100, and Caltech101 data sets demonstrate the effectiveness of F-SVM [41]. Yan Song (2018) investigate and overview mathematical optimization algorithms and parallel technologies of SVM, propose a Vote Parallel SVM to reduce the training time [42].

At present, deep learning algorithms have been increasingly more recognized for their efficiency and the accuracy of their results compared to traditional models. With the rise of big data and artificial intelligence, machine learning algorithms have been studied in various fields, and the SVM algorithm has been widely used in the field of insurance actuarial work. Therefore, this paper uses the CNN algorithm for feature extraction, then introduces the Hull Vector to optimize the operating efficiency of the SVM algorithm and finally combines the two algorithms to establish the CNN-HVSVM algorithm, which can be used to solve the problem of determining the driving behavior risk level of UBI customers. The model combing the CNN algorithm and HVSVM algorithm has not been applied in the field of UBI car insurance rate determination, but the combination of these two methods is feasible. The CNN model can effectively extract sample features through convolution and pooling operations, and transfer the feature vector to the HVSVM model for multi-classification operation. Then, customers can be classified according to their driving behavior through the HVSVM algorithm. The HVSVM algorithm will first select the Hull Vector that contains all the support vectors, use the KKT conditions to eliminate the useless samples in the new samples, reduce the number of samples involved in training, and then quickly train the support vector machine for incremental learning using the new training set. The combination of the CNN algorithm and the HVSVM algorithm can effectively solve the problems of overfitting and local optimization and improve the efficiency of the model’s operations; moreover, the model uses a 10-fold cross-validation method to process data, which reduces the sensitivity of the model to the results of the data set division and makes the model have better robustness to achieve a better judgment effect. In addition, the BP neural network algorithm, CNN algorithm, SVM algorithm and CNN-SVM algorithm are introduced in this paper and are compared with the CNN-HVSVM algorithm, highlighting the superiority of the CNN-HVSVM algorithm in UBI vehicle insurance rate determination, and making the determination result of the model more convincing.

II. ALGORITHM PRINCIPLE AND MODEL CONSTRUCTION

To realize the determination of UBI car insurance premium rates, the data need to be processed by the classifier. Common classifiers include the CNN algorithm, SVM algorithm, BP neural network and so on. The CNN algorithm has good feature extraction capabilities, but it is prone to overfitting. The multi-classification problem of the HVSVM model solves the convex quadratic optimization problem by introducing slack variables, effectively avoiding the overfitting problem, and the solution obtained is the global best. Therefore, the fully connected layer of the CNN algorithm is replaced with the HVSVM classifier, and the CNN-HVSVM algorithm is constructed to solve the problem of determining the UBI car insurance rate level. This paper mainly compares the CNN-HVSVM algorithm and CNN-SVM algorithm, forecasts the same sample with the two respective algorithms, obtains the two prediction results of the sample, and then compares the prediction accuracy and algorithm efficiency of the two algorithms. The following is an introduction to the principles of the related algorithms.

A. PRINCIPLE OF THE CNN ALGORITHM

In addition to the more common input layer, fully connected layer and output layer in the neural network, the CNN model also includes a relatively special convolution layer and pooling layer. The structure of the CNN model is shown in Fig. 1.

The prediction problem of the CNN model first conducts the input data forward, then outputs feature maps at various levels, and finally outputs the conditional probability distribution based on the input data at the fully connected layer. The special layer structures of the CNN model are the convolution layer and pooling layer. It is the convolution operation of the convolution layer and the pooling operation of the pooling layer that make the CNN model have a strong feature extraction ability.

The convolutional layer of the CNN model extracts features through a convolution operation. The input is usually the original image \( X \). In this paper, \( H_i \) is used to represent the feature map \( (H_0 = X) \) of layer \( H_i \) of the CNN model. Assuming that \( H_i \) is a convolutional layer, the generation process of \( H_i \) can be described as:

\[
H_i = f (H_{i-1} \otimes W_i + b_i). \tag{1}
\]

Here, the weight vector of the layer \( i \) convolution kernel is \( W_i \); “\( \otimes \)” represents the convolution operation of the convolution kernel and the layer \( i - 1 \) feature map. The output result of the convolution operation is added to the offset vector \( b_i \) of layer \( i \) and the sum of the two is input to the nonlinear excitation function \( f(x) \) to obtain the characteristic graph \( H_i \) of layer \( i \).

After the convolutional layer, the pooling layer performs the pooling operation. The pooling layer removes redundant information, compresses feature maps, and simplifies network calculations. The most practical pooling method is Max pooling, which selects the maximum value of each part in
the input matrix as the value of the corresponding part of the output matrix. Assuming that \( H_i \) is the down sampling layer, we get the following:

\[
H_i = \text{subsampling}(H_{i-1}). \tag{2}
\]

The Max pooling operation is shown in Fig. 2. The largest number in the upper left corner of the input matrix is 9. Then, the corresponding upper left corner of the output matrix is 9, and so on.

The convolutional layer and the downsampling layer are usually arranged alternately. A convolutional layer is followed by a pooling layer. There can be multiple sets of convolutional layers and pooling layers to perform multiple data transformations and dimensionality reduction. The CNN model is a mathematical model that allows the original image data transformations and dimensionality reduction. The CNN model is a mathematical model that allows the original image data to be transformed and dimensionality reduced and maps to a new feature expression \((Y_{0})\) where \(l_{i}\) represents label category \(i\) is calculated as shown in formula (3):

\[
Y(i) = P(L = l_{i}|H_0; (W, b)). \tag{3}
\]

B. HV SVM ALGORITHM PRINCIPLES

1) SVM ALGORITHM PRINCIPLES

The Support Vector Machine (SVM), based on the principle of minimizing the structural risk in statistical learning theory, seeks a weighted and minimized empirical risk and confidence range. Assume a set of training data, \((x_1, y_1), \ldots, (x_n, y_n)\), \(x_i \in \mathbb{R}^d, y_i \in \{+1, -1\}, i = 1, 2, \ldots, n\). If the kernel function is \(k(x, y)\), the SVM problem solves the minimum objective function \(W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j k(x_i, x_j)\), which also has to satisfy the constraint conditions \(0 \leq \alpha_i \leq C\) and \(\sum_{i=1}^{n} \alpha_i = 0\); and the obtained classifier function obtained can be represented as \(f(x) = \text{sgn} [\sum_{i=1}^{n} y_i \alpha_i^* k(x, x_i) + b^*]\). The support vector is the training sample corresponding to \(\alpha_i\) that is not equal to 0.

In the classification problem of nonlinear support vector machines, the SVM model converts the inner product \((x_i \cdot x_j)\) of the transformation space into a function \(K(x_i \cdot x_j) = (\phi(x_i) \cdot \phi(x_j))\) in the original space by introducing a kernel function \(K(x_i \cdot x_j)\), maps the sample \(x\) to a high-dimensional space \(E\) and divides the original problem linearly. After replacing the inner product with a kernel function, the original quadratic programming problem is still a convex problem, which also has a global optimal solution. We construct the following dual optimization problem:

\[
\max_{\alpha} Q(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j)
\]

\[\text{s.t. } \sum_{i=1}^{n} \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C, \quad i = 1, 2, \ldots, n. \tag{4}\]

The corresponding optimal decision function at this time is as follows:

\[
f(x) = \text{sgn} [\sum_{i=1}^{n} y_i \alpha_i^* K(x, x_i) + b^*]. \tag{5}\]

2) BASED ON THE BASIC IDEA OF HV SVM FAST ALGORITHM

The operation speed of the support vector machine is mainly determined by the number of training samples. The more training samples there are, the slower the speed. In fact, the resulting support vector is only a part of the training sample and is usually a very small part. In other words, most of the training samples cannot become support vectors, but they increase the calculation time; and the algorithm can construct the correct optimal classification surface only by getting the support vector. If the sample most likely to be a support vector can be selected from the training sample set in advance, other samples can be discarded, and then training the new sample set will greatly improve the operating speed. In addition, since all the reserved samples are likely to be support vectors, the accuracy of the algorithm will not be affected.

It can be seen from Fig. 3 that only through the classification surface of the sample located at the vertex of the convex hull is it possible to correctly classify the two types of samples, and the support vector is generated in it, rather than using the sample point located inside the convex hull.

Therefore, we introduced the Hull Vector, which includes all the samples located at the edge of the training set, that is, the samples located on the convex hull of the training set. In fact, the Hull Vector is the convex vertex of the...
sample set. The Hull Vector includes all support vectors, but not necessarily all nonsupport vectors; however, the support vectors must be in the Hull Vector. In this way, the set of Hull Vectors is used as the training sample set and then handed over to the SVM for incremental training, and the support vector can be further obtained. Fig. 3 shows the case where the sample set is linearly separable. For the nonlinear sample set, Freund’s method of mapping the original training space to a high-dimensional space is used to make the sample set linearly separable in this space [43].

3) SVM INCREMENTAL LEARNING ALGORITHM

The dual problem of the SVM learning problem can be expressed as follows:

$$\min_{0 \leq \alpha_i \leq C} Q = \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j K_{ij} - \sum_i \alpha_i + b \sum_i y_i \alpha_i.$$  

Here $\alpha_i$ is the corresponding coefficient, $b$ is the offset, $K_{ij} = y_i y_j k(x_i, x_j)$ is the symmetric positive definite kernel matrix, and $C$ is the penalty coefficient.

The KKT conditions for the dual problem are as follows:

$$g_i = \frac{\partial Q}{\partial \alpha_i} = \sum_j K_{ij} \alpha_j + y_i b - \frac{\partial Q}{\partial b} = \sum_j y_j \alpha_j = 0. \quad (7)$$

The KKT conditions divide the training set $D$ into three parts: the support vector set $S(0 < \alpha_i < C, \ g_i < 0)$, the error vector set $E(\alpha_i = C, \ g_i < 0)$, and the correct classification set $R(\alpha_i = 0, \ g_i > 0)$.

If a new sample $c$ is added to the training set, the change of $g_i$ is as follows:

$$\Delta g_i = K_{ic} \Delta \alpha_c + \sum_j K_{ij} \Delta \alpha_j + y_i \Delta b \quad 0 = y_i \Delta \alpha_c. \quad (8)$$

Therefore: $\Delta b = \beta \Delta \alpha_c, \ \Delta \alpha_i = \beta \Delta \alpha_c$.

If the newly added training sample is a support vector, then update $\alpha$ and $b$ as follows; otherwise, no update is required.

$$\Delta \alpha_c = \frac{\sum_{j \in S} K_{cj} \alpha_j + y_c b - 1}{K_{cc} + \sum_{j \in S} K_{cj} \beta_j + y_c \beta} \quad (9)$$

$$\alpha = \left[ \begin{array}{c} \alpha + \Delta \alpha_c \\ \alpha_c \end{array} \right] b = b + \Delta b \quad (10)$$

4) BASIC STEPS OF THE SVM FAST INCREMENTAL LEARNING ALGORITHM BASED ON THE HULL VECTOR

Let the sample set be: $(x_i, y_i), i = 1, 2, \cdots, n, x \in R_d, y \in \{+1, -1\}$ is the category symbol.

1) If the sample set is linearly separable, then proceed directly to step 2; otherwise, the method in sections 1 and 2 is used to map the original training space to a high-dimensional space so that the sample set is linearly separable in this space.

2) Find the vertices of the convex hull of the two sample sets, $HV_i = (x_{HV_i}, y_i), 0 \leq i \leq r, r < n$. The resulting convex vectors are convex vectors. The following algorithms are for the new training sample set composed of Hull Vectors.

3) Initialization. The support vector set $S = \{Two recent heterogeneous samples\}, \alpha_c = 0$;

4) Start iteration. Starting from the $r$ Hull Vector samples, add one sample at a time for training.

5) If $g_i > 0$, the added sample is not a support vector or an error vector. Then, the process ends and returns to step (4).

6) If $g_i \leq 0$, incremental learning is applied so that one of the following conditions occurs. 1) If $g_i = 0$, add the added sample $c$ to the support vector set $S$, and update $A$. Then, end the process. 2) If $\alpha_c = C$, add the added sample $c$ to the error set $E$;

7) Bookkeeping. Check whether the following conditions are true, and update the three sample sets accordingly: the support vector set $S$, the error vector $E$, and the samples in the correct classification set $R$:

$$g_c \leq 0; \alpha_c \leq C; 0 \leq \alpha_j \leq C, \ \forall j \in S;$$

$$g_i \leq 0; g_i > 0, \ \forall i \in R;$$

8) After Bookkeeping, return to steps (4)-(8). When all Hull Vector samples are added to the training, the algorithm ends.

In this way, in the process of seeking the Hull Vector in the first step, the quadratic optimization problem is avoided, and the model can be solved by linear programming, which greatly reduces the operating complexity and operating time. At the same time, the number of new training sets of the hull vector is only a part of the original training set, and so it greatly reduces the complexity and training time of the subsequent training process.

The multiclassification method selected in this paper conducts one-to-many classification (OVR SVM), which constructs a multiclassifier by combining multiple classifiers. During training, the OVR SVM model first classifies the samples of category $i(i = 1, 2, 3, 4, 5)$ into one category and the other remaining samples into another category, thus constructing 5 two-class SVMs. This article divide the new Hull Vector training set into 5 training sets, uses the 5 training sets to train separately, and then brings them into the test set for testing. Each test has a result set $\{f_1(x), f_2(x), f_3(x), f_4(x), f_5(x)\}$, and the largest of the 5 values
and the category corresponding to the value are used as the classification result.

C. CNN-HVSVM MODEL CONSTRUCTION
The CNN model can be divided into the feature extraction part and classification part according to its main function. The feature extraction part includes a convolutional layer and a pooling layer, and the classification part includes a fully connected layer. The input-based probability distribution is obtained at the fully connected layer, but this is prone to overfitting and local optimization problems. The multiclassification problem of the HVSVM model solves the convex quadratic optimization problem by introducing slack variables, which effectively avoids the overfitting problem, and obtains the optimal solution. Therefore, the fully connected layer of the CNN model is replaced by the HVSVM classifier and used to construct the CNN-HVSVM model to determine the UBI car insurance rate grade. When the CNN model is trained to stabilize the training set’s accuracy rate, the high-dimensional feature representation of the test set obtained by the sampling layer is extracted and used as the input of the HVSVM model. In the model, the CNN model is used to extract features, and the HVSVM model is used as a multiclassifier for predictive classification.

The structure of the CNN-HVSVM model is shown in Fig. 4 in which the fully connected layer of the CNN model is replaced by the HVSVM classifier. The CNN model is used to extract the features of the input UBI car insurance user driving data, and the extracted features are used as the input of the HVSVM model. The HVSVM model is used to complete the multiclassification process, and so its output is the classification prediction value.

The specific operation steps of the CNN-HVSVM model are:
1) Pre-process the sample data;
2) Set the parameters and layers of CNN model and HVSVM model to initialize the model;
3) Convert the panel data of the training set into a matrix, enter the CNN model as the input vector, import the feature vector output from the CNN model into the HVSVM model as the input vector, and use the training data to train the model parameters;
4) Convert the panel data of the test set into a matrix, and bring it into the trained CNN-HVSVM model to obtain multi-classification results. Compare with the actual level of the customer and observe the effect of classification and the speed of the model.

III. USING THE CNN-HVSVM MODEL FOR UBI CAR INSURANCE RATE DETERMINATION

A. INDEX SELECTION
According to the weights of the multiple risk factors that cause accidents derived from the entropy weight hierarchy analysis method proposed by Zhu Shuang (2015), the high weighted risk factors in the calculation results of this paper are selected as the sample indicators and include the ratio of time driving more than 120 km/h, the number of times of rapid acceleration, the monthly driving mileage and the night driving time ratio. The specific weights of the risk factors are shown in Table. 1.

| Risk factor                  | Weights obtained by entropy weight analytic hierarchy process |
|-----------------------------|---------------------------------------------------------------|
| Monthly mileage             | 0.0765                                                        |
| Higher than 120km/h time ratio | 0.2601                                      |
| Rapid acceleration          | 0.0749                                                        |
| Night driving time ratio    | 0.1044                                                        |

Note: Data source: Zhu Shuang. Research on UBI-based car insurance rate determination model and method based on UBI [D]. Beijing Jiaotong University, 2015.

B. DATA PREPROCESSING
The rating of one’s driving behavior is the basis for the effective determination of UBI car insurance rates. The CNN-HVSVM algorithm can match groups with similar driving characteristics to a large number of historical data groups, and then evaluate the user rating. The driving behavior data of the driver in this article comes from the “Internet car insurance UBI product design” proposed by Jiang Lei (2017), which collects the drivers’ monthly mileages, night driving time ratios, ratios of time driving more than 120 km/h, and rapid acceleration behavior as the 4 indicator variables. The data are composed of two parts, namely, the panel data of the indicator system and the customers’ driving behavior level. Examples of some users’ driving behavior data are shown in Table. 2. Screen the above risk factor data in the original database and establish a historical data state vector database. The established data state vector library is divided into 90% training set and 10% test set. Due to space limitations, this article only shows part of the test set data as shown in Table. 2.
TABLE 2. Some examples of user driving behavior data.

| User | Monthly mileage (KM) | Proportion of time higher than 120km/h (%) | Rapid acceleration | Night driving time (h) | Grade |
|------|----------------------|------------------------------------------|-------------------|-----------------------|-------|
| 1    | 84                   | 0                                        | 0                 | 0                     | A     |
| 2    | 362                  | 0                                        | 0.9               | A                     |
| 3    | 1337                 | 0.66                                     | 1                 | B                     |
| 4    | 870                  | 0.48                                     | 4                 | B                     |
| 5    | 754                  | 1.02                                     | 2                 | C                     |
| 6    | 1180                 | 0.63                                     | 3                 | C                     |
| 7    | 1240                 | 1.24                                     | 13                | D                     |
| 8    | 574                  | 3.04                                     | 6                 | D                     |
| 9    | 1710                 | 0.95                                     | 34                | E                     |
| 10   | 1185                 | 4.86                                     | 4                 | E                     |

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

The customers’ driving behavior levels are divided into categories A, B, C, D, and E, which respectively correspond to five driving behavior risk levels: extremely low, low, medium, high, and extremely high. The representation method of the matrix \( A = (0, 0, 0, 0, 0, 0, 0) \) in the model of class i’s rank replaces the element at the corresponding position of the matrix with 1. For example, class B is \( A_2 = (0, 1, 0, 0, 0, 0, 0, 0, 0, 0) \).

There is a dimensional difference between the indicators of the index system, and so normalization needs to be performed. The specific normalization operation is implemented by the mapminmax function, and the implemented formula is shown in (11):

\[
y = \frac{(y_{\max} - y_{\min}) \times (x - x_{\min})}{x_{\max} - x_{\min}} + y_{\min} \quad (11)
\]

Here, in the formula, \( x \) is the feature vector that needs to be normalized, \( x_{\max} \) and \( x_{\min} \) represent the largest and smallest vectors in \( D \), and \( y_{\max} \) and \( y_{\min} \) correspond to the largest and smallest values in the interval range after the data are normalized. At this time, all the values in the original stock characteristic data are converted into the interval \([0, 1]\). Such processing can avoid the different dimensions between different characteristic attributes without affecting the internal relationship of the same attribute value.

Since the training and test sets are randomly divided, in order to reduce the error caused by this randomness, the input data set that was common in the past is changed to multiple input sets. It can be seen from the theorem of large numbers that the greater the number of trials is, the greater the probability of this occurring. To make the final prediction accuracy converge and the model to not be too complicated, the process of randomly dividing the training set and the test set is repeated 10 times to obtain 10 different data sets, namely, 10-fold cross-validation is conducted. Through the 10-fold cross-validation testing, the prediction results of these independent and different data subsets can be obtained, and finally, the average of the results of the 10 groups of subexperiments. The sensitivity of 10-fold cross-validation to the division of the data set is relatively low; therefore, each sample has the opportunity to be divided into a training set or a test set. While obtaining the experimental results, we can also observe the variance of the model prediction results and this can reflect the robustness of the model.

The 10 data sets are input into the CNN-HVSVM model using a cyclic algorithm, each group extracts data features according to the convolution operation and the pooling operation and then passes them to the HVSVM model. The HVSVM model performs multiclassification to obtain the classification results, and at the same time it obtains the classification accuracy of the results. After 10 cycles, according to the average of the 10 groups of experimental results, the accuracy rate of the UBI vehicle insurance rate grade judgment based on the CNN-HVSVM algorithm is obtained. The algorithm is implemented with Python 3.6 (64 bit).

IV. EMPIRICAL ANALYSIS

Using 10-fold cross-validation, the training set is divided into 10 subsets without duplication, 9 of which are used for model training, and the remaining one is used for testing. After repeating this 10 times, 10 models and their performance evaluations can be obtained.

A. ANALYSIS OF THE EFFECT OF THE UBI CAR INSURANCE RATE DETERMINATION MODEL BASED ON THE CNN ALGORITHM

The activation function in the CNN model uses the segmented linear activation ReLU function and the nonlinear activation sigmoid function that are commonly used in multiclassification problems to observe the respective effects. We use the convolutional neural network to calculate the accuracy of the predicted value, and the results are shown in Table 3. From the table, when the CNN model uses the ReLU activation function, the accuracy obtained is 82.9%. When it uses the Sigmoid activation function, the rate is 79.1%. Obviously, the prediction effect from using the ReLU activation function is better.

Next, we observe the influence of the number of iterations on the prediction accuracy of the convolutional neural network. Table. 4 lists the prediction accuracies of the convolutional neural network under different numbers of iterations when using the ReLU activation function.
It can be seen from Table 4 that as the number of iterations increases from 10 to 5000, the prediction accuracy of the convolutional neural network continuously improves from 0.491 to 0.830. However, as the number of iterations increases, the operating efficiency will be greatly reduced. To balance the accuracy and operating efficiency, this paper sets the number of iterations as 1000, and the accuracy rate is 82.9%.

### Table 4. Prediction accuracy of convolutional neural networks under different iterations.

| Data set | 10 | 50 | 100 | 500 | 1000 | 2000 | 5000 |
|----------|----|----|-----|-----|------|------|------|
| 1        | 0.494 | 0.736 | 0.795 | 0.806 | 0.811 | 0.834 | 0.832 | 0.834 |
| 2        | 0.499 | 0.736 | 0.777 | 0.807 | 0.816 | 0.828 | 0.831 | 0.830 |
| 3        | 0.495 | 0.769 | 0.789 | 0.795 | 0.821 | 0.831 | 0.831 | 0.832 |
| 4        | 0.500 | 0.741 | 0.791 | 0.797 | 0.811 | 0.838 | 0.824 | 0.831 |
| 5        | 0.484 | 0.758 | 0.777 | 0.799 | 0.823 | 0.836 | 0.826 | 0.823 |
| 6        | 0.494 | 0.763 | 0.786 | 0.811 | 0.812 | 0.831 | 0.835 | 0.826 |
| 7        | 0.490 | 0.763 | 0.777 | 0.810 | 0.819 | 0.829 | 0.832 | 0.833 |
| 8        | 0.481 | 0.757 | 0.786 | 0.809 | 0.816 | 0.804 | 0.834 | 0.823 |
| 9        | 0.486 | 0.746 | 0.789 | 0.800 | 0.813 | 0.836 | 0.826 | 0.822 |
| 10       | 0.490 | 0.770 | 0.784 | 0.799 | 0.818 | 0.827 | 0.821 | 0.836 |
| Average  | 0.491 | 0.754 | 0.785 | 0.803 | 0.816 | 0.829 | 0.830 | 0.829 |

### Table 5. CNN-HVSVM model prediction accuracy rate when CNN activation function is ReLU.

| Data set | Accuracy under linear activation function | Accuracy under RBF activation function |
|----------|------------------------------------------|----------------------------------------|
| 1        | 0.836 | 0.870 |
| 2        | 0.847 | 0.860 |
| 3        | 0.841 | 0.866 |
| 4        | 0.837 | 0.870 |
| 5        | 0.841 | 0.866 |
| 6        | 0.834 | 0.865 |
| 7        | 0.832 | 0.851 |
| 8        | 0.847 | 0.852 |
| 9        | 0.849 | 0.854 |
| 10       | 0.845 | 0.854 |
| Average  | 0.841 | 0.861 |

### Table 6. CNN-HVSVM model prediction accuracy rate when CNN activation function is sigmoid.

| Data set | Accuracy under linear activation function | Accuracy under RBF activation function |
|----------|------------------------------------------|----------------------------------------|
| 1        | 0.783 | 0.813 |
| 2        | 0.793 | 0.818 |
| 3        | 0.799 | 0.829 |
| 4        | 0.800 | 0.825 |
| 5        | 0.793 | 0.830 |
| 6        | 0.781 | 0.836 |
| 7        | 0.800 | 0.840 |
| 8        | 0.782 | 0.839 |
| 9        | 0.790 | 0.820 |
| 10       | 0.794 | 0.829 |
| Average  | 0.792 | 0.825 |

B. ANALYSIS OF THE EFFECT OF THE UBI CAR INSURANCE RATE DETERMINATION MODEL BASED ON THE CNN-HVSVM ALGORITHM

It can be seen from the previous section that when the number of iterations of the convolutional neural network is 1000, the accuracy and feature extraction of the model are the best and the most efficient; therefore, the number of iterations of the convolutional neural network is still set to 1000 and the one-to-many classification method is also used as the support vector machine classification method. We observe the effect when the convolutional neural network uses the ReLU function and the support vector machine uses the linear activation function or the RBF function, respectively. The obtained accuracy rates are shown in Table 5. It can be seen that the model prediction classification accuracy of the RBF activation function is 86.1%.

Table 6 shows the accuracy results when the CNN algorithm uses the sigmoid activation function and the HVSVM algorithm also uses the linear activation function or the ReLU function, respectively. It can be seen that the prediction effects under the two activation functions is worse than that using the ReLU activation function alone, and they are 79.2% and 82.8%, respectively.

Based on the above experiments, when the CNN uses the ReLU activation function and the HVSVM uses the RBF activation function, the prediction accuracy is the highest at 86.1%. Under this model, some UBI users’ driving behavior data are predicted in Table 2, and the prediction results are shown in Table 7.

Except that the prediction result of User 5 is inconsistent with the actual result given in Table 2, all other prediction results are correct. All user levels are predicted, and the prediction accuracy rate is 85%, which is close to the accuracy rate of 86.1% obtained by the above experiment; therefore, the experimental error is within a reasonable range and there were no under-fitting and over-fitting problems during the experiment.

C. PERFORMANCE TEST OF CNN-HVSVM MODEL

The most commonly used evaluation criteria for evaluating the quality of a classifier are the accuracy and operating speed. The higher the accuracy is, the better the classification effect. As mentioned in the principles in 2.2, the Hull Vector includes all support vectors, but not necessarily all nonsupport vectors; however, the support vectors must be in the Hull Vector, and so there is no difference in the accuracy of the HVSVM algorithm and that of the SVM algorithm. To make the experimental results more convincing, this paper uses the BP neural network algorithm, CNN algorithm, SVM algorithm and CNN-HVSVM algorithm for the accuracy test.
The experimental results show that the CNN-HVSVM algorithm has an average classification accuracy of 86.1%. Because the CNN-HVSVM algorithm has high accuracy, it is more suitable for UBI vehicle insurance rate determination. Furthermore, the experiment in this paper uses the 10-fold cross-validation method, and the results obtained have less variance; therefore, the model is more robust.

Next, the operating speeds of the CNN-HVSVM algorithm and CNN-SVM algorithm are tested. The driver’s driving behavior data come from the “Internet car insurance UBI product design” proposed by Jiang Lei (2017), which collects a total of 2000 samples, including drivers’ monthly mileage, night driving time ratios, ratios of time driving more than 120 km/h, and rapid acceleration behavior as the 4 data indexes. The samples are randomly divided into 10 subsets to obtain 10 different data sets and 10-fold cross validation testing is performed. Each SVM parameter in the HVSVM and SVM takes the same value, and a polynomial kernel function \((q = 3)\) is used. The experimental results are shown in Table. 9.

As seen from Table. 9, the classification results of the CNN-HVSVM algorithm and the CNN-SVM algorithm are equivalent, but the speed of the CNN-HVSVM is increased by an average of 1.563 times. Similar results have been obtained when the same experiment was carried out for other kernel functions, that is, when processing a training set with a large number of samples, the CNN-HVSVM algorithm has an obvious speed advantage.

**V. CONCLUSION**

Based on the CNN-HVSVM algorithm, this paper establishes a rating model for UBI car insurance rates. The model combines the CNN’s superior feature extraction capabilities, the SVM’s stable classification capabilities, and the introduction of the Hull Vector to optimize the operating efficiency of the traditional SVM algorithm, thus significantly improving its operating speed. The CNN-HVSVM algorithm can prevent overfitting and local optimization problems to improve the calculation efficiency and effectively solve the UBI car insurance rate determination problem. In this paper, a 10-fold cross-validation method is used to process the data set, which reduces the sensitivity of the model to the results of the data set division and ensures that each sample has the possibility of being divided into the training set or the test set. The results verify the good robustness of the model. In addition, this paper also verifies the rationality of the CNN-HVSVM model in dealing with the feature extraction of the original data of the insurance index system and the multi classification of insurance samples. Compared with the BP neural network algorithm, the CNN algorithm, the SVM algorithm and the CNN-SVM algorithm, the CNN-HVSVM algorithm can accurately and efficiently predict the grades of UBI car insurance users, the prediction results are in line with the actual situation, and the results have strong applicability and flexibility. As Internet of vehicles technology matures, more risk factors can be collected to describe driving behavior more accurately. Model optimization and data enrichment can better promote future UBI research work.

In this study, the CNN-HVSVM algorithm is applied to the field of UBI car insurance rate determination to form a correspondence between driver behavioral levels and car insurance rates, which provides a reference for insurance companies to adjust customers’ rate coefficients. The results
show that the UBI car insurance rate determination model based on the CNN-HVSVM algorithm can determine drivers’ behavior more fairly and reasonably, which is also of practical significance to promoting the reform of China’s car insurance rate marketization.

REFERENCES

[1] D. L. Estrin, “Participatory sensing: Applications and architecture,” in Proc. 5th Int. Conf. Mobile Syst., Appl., Services (MobiSys), 2010, pp. 3–4, doi: 10.1145/1814433.1814435.

[2] Y. Su, “Analysis on the realization method of big data knowledge,” J. Shandong Univ. Sci. Technol., vol. 21, no. 1, pp. 20–26, Jan. 2019, doi: 10.16452/j.cnki.sjxsk.20190129.005.

[3] C. M. Wanweea, B. K. Mueema, and J. K. Mung’uattu, “Modelling a pay-as-you-drive insurance pricing structure using a generalized linear model: Case study of a company in Kaambu,” Amer. J. Theor. Appl. Statist., vol. 4, no. 6, pp. 527–533, Nov. 2015, doi: 10.11648/j.ajtas.20150406.23.

[4] S. Hussnjak, D. Perakovic, I. Forenbacher, and M. Mumdziev, “Telematics system in usage based motor insurance,” Procedia Eng., vol. 100, pp. 816–825, 2015, doi: 10.1016/j.proeng.2015.01.436.

[5] F. Guo, I. Kim, and S. G. Klauer, “Semiparametric Bayesian models for evaluating time-variant driving risk factors using naturalistic driving data and case-crossover approach,” Statist. Med., vol. 38, no. 2, pp. 160–174, Jan. 2019, doi: 10.1002/sim.7574.

[6] M. F. Carfora, F. Martinelli, F. Mercaldo, V. Nardone, A. Orlando, A. Santone, and G. Vaglini, “A ‘pay-as-you-drive’ car insurance approach through cluster analysis,” Soft Comput., vol. 23, no. 9, pp. 2863–2875, May 2019, doi: 10.1007/s00500-018-2374-y.

[7] Y. Bian, C. Yang, J. L. Zhao, and L. Liang, “Good drivers pay less: A study of usage-based vehicle insurance models,” Transp. Res. Part A, Policy Pract., vol. 107, pp. 20–34, Jan. 2018, doi: 10.1016/j.tra.2017.10.018.

[8] M. Denuit, M. Guillo, and J. Trufin, “Multivariate credibility modelling for usage-based motor insurance pricing with behavioural data,” Ann. Actuarial Sci., vol. 13, no. 2, pp. 378–399, 2019, doi: 10.1017/S1748499518000349.

[9] S. Arumugam and R. Bhargavi, “A survey on driving behavior analysis in usage based motor insurance based on ant colony algorithm to optimize random forest model,” IEEE Trans. Int. Transp. Electron., vol. 6, no. 5, pp. 4279–4289, May 2018, doi: 10.1011/TIE.2017.2764861.

[10] R. Dong, O. Peng, X. Li, and X. Guan, “CNN-SVM with embedded recurrent structure for social emotion prediction,” in Proc. Chin. Autom. Cong. (CAC), Xi’an, China, Nov. 2018, pp. 3024–3029, doi: 10.1086/6823318.

[11] X. Sun, J. Park, K. Kang, and J. Hur, “Novel hybrid CNN-SVM model for recognition of functional magnetic resonance images,” in Proc. IEEE Int. Conf. Syst. Man, Cybern. (SMC), Banff, AB, Canada, Oct. 2017, pp. 1001–1006, doi: 10.1109/SMC.2017.8122741.

[12] C. Yan and X. Yang, “UBI car insurance rate grade determination based on K-S algorithm,” J. Nonlinear Anal., vol. 20, no. 5, pp. 957–965, 2019.

[13] S. Zhao, T. Zhao, and J. Hou, “Introduction of SVM to Motor vehicle insurance fraud identification,” China Insurance, vol. 8, pp. 15–19, Aug. 2015.

[14] Z. Xue, “Prediction of car insurance claim frequency based on ARIMA-SVM,” M.S. thesis, Southwest Univ. Finance Econ., Chengdu, China, 2018.

[15] S. Tang, “Performance evaluation of social endowment insurance system based on RS-SVM,” M.S. thesis, Chongqing Normal Univ., Chongqing, China, 2019.

[16] B. Wen, G. Shan, and X. Duan, “Incremental learning algorithm based on KKT condition and shell vector,” Comput. Sci., vol. 40, no. 3, pp. 255–258, 2013.

[17] Y. Mao, Y. Wei, and J. Jia, “A SVM incremental learning algorithm based on KKT condition and shell vector,” Electron. Sci. Technol., vol. 29, no. 2, pp. 38–40 and 44, 2016.

[18] S. Yu, X. Li, X. Zhang, and H. Wang, “The OCS-SVM: An objective-cost-sensitive SVM with sample-based misclassification cost invariance,” IEEE Access, vol. 7, pp. 118931–118942, 2019, doi: 10.1109/ACCESS.2019.2934347.

[19] M. Lei, L. Zhang, M. Li, H. Chen, and X. Zhang, “Near-infrared spectrum for coal origin identification based on SVM algorithm,” in Proc. 35th Chin. Control Conf. (CCC), Wuhan, China, Jul. 2016, pp. 9016–9020, doi: 10.23919/ChiCC.2016.0843742.

[20] C. Yan, and X. Guo, L. V. F. Li, C. Gao, and Y. Liu, “Multi-color space rock shin-section image classification with SVM,” in Proc. IEEE 8th Joint Int. Inf. Technol. Artif. Intell. Conf. (ITAIAC), Chongqing, China, May 2019, pp. 571–574, doi: 10.1109/ITAIAC.2019.8785477.

[21] K. N. Devi, V. M. Bhaskaran, and G. P. Kumar, “Cuckoo optimized SVM for stock market prediction,” in Proc. Int. Conf. Innov. Inf., Embedded Commun. Syst. (ICIECS), Mar. 2015, pp. 1–5, doi: 10.1109/ICIECS.2015.7192906.

[22] J. Kordonis, S. Symeonidis, and A. Arampatzis, “Stock price forecasting via sentiment analysis on Twitter,” in Proc. 20th Pan-Hellenic Conf. Informat. (PCI), 2016, pp. 1–6, doi: 10.33406/pcii.2016.030378.

[23] V. Plakandaras, J. Cunado, R. Gupta, and M. E. Wohar, “Do leading indicators forecast U.S. recessions? A nonlinear re-evaluation using historical data,” Int. Finance, vol. 20, no. 3, pp. 289–316, Dec. 2017, doi: 10.1111/1650-2829.12111.
[40] S. Mukhoti and P. Ranjan, “A new class of discrete-time stochastic volatility model with correlated errors,” Appl. Econ., vol. 51, no. 3, pp. 259–277, Jan. 2019, doi: 10.1080/00036846.2018.1494804.

[41] X. Wu, W. Zuo, L. Lin, W. Jia, and D. Zhang, “F-SVM: Combination of feature transformation and SVM learning via convex relaxation,” IEEE Trans. Neural Netw. Learn. Syst., vol. 29, no. 11, pp. 5185–5199, Nov. 2018, doi: 10.1109/TNNLS.2018.2791507.

[42] Y. Song, Q. Jin, K. Yan, H. Lu, and J. Pan, “Vote parallel SVM: An extension of parallel support vector machine,” in Proc. IEEE SmartWorld, Ubiquitous Intell. Comput., Adv. Trusted Comput., Scalable Comput. Commun., Cloud Big Data Comput., Internet People Smart City Innov. (SmartWorld/SCALCOM/UI/ATC/CBDCom/IOP/SCI), Guangzhou, China, Oct. 2018, pp. 1942–1947, doi: 10.1109/SmartWorld.2018.00325.

[43] Y. Freund and R. E. Schapire, “Large margin classification using the perception algorithm,” Mach. Learn., vol. 37, no. 3, pp. 296–299, 1999.

CHUN YAN received the B.S., M.S., and Ph.D. degrees from the Shandong University of Science and Technology, Qingdao, China, in 2000, 2003, and 2011, respectively. She is currently a Professor with the Shandong University of Science and Technology, Qingdao, China. Her research interests include applied mathematics, statistics, machine learning, intelligent systems and economic management. She has about 30 technical articles published in journals and conference proceedings in her research areas.

XINDONG WANG received the B.S. degree in financial mathematics and the M.S. degree in statistics from the Shandong University of Science and Technology, Qingdao, China, in 2018 and in 2020, respectively. His research interests are in the areas of big data, blockchain, actuarial and computer algorithm improvement, and financial risk control.

XINHONG LIU received the bachelor’s and master’s degrees from the Shandong University of Science and Technology, in 2000 and 2003, respectively, and the Ph.D. degree from the Renmin University of China, in 2015. She is currently an Associate Professor with the Beijing Institute of Petro-chemical Technology, Beijing, China. Her research areas include: nonlife actuarial and statistical models, risk management, and catastrophe data analysis. She has about 20 technical articles published in journals and conference proceedings in her research areas.

WEI LIU (Member, IEEE) received the B.S., M.S. and Ph.D. degrees from the Shandong University of Science and Technology, Qingdao, China, in 1999, 2005, and 2012, respectively. He is currently a Professor with the Shandong University of Science and Technology. His research interests are in formal engineering methods and machine learning. He has led or participated in many projects supported by the National Natural Science Foundation, National Key Basic Research Development Program, and other important and key projects at provincial levels. He has published more than 30 articles in journals and conference proceedings, such as the IEEE TRANSACTIONS ON SYSTEMS MAN AND CYBERNETICS: SYSTEMS AND ENTERPRISE INFORMATION SYSTEMS. He was a recipient of Science and Technology Advancement Awards at the province level and the Excellent Doctoral Dissertation Award from Shandong province, China.

JIAHUI LIU received the B.S. and M.S. degrees in statistics from the Shandong University of Science and Technology, in 2019. Her research interests are in the areas of big data, statistics, actuarial science, and computer algorithm improvement.

* * *