Hindawi
Advances in Mathematical Physics
Volume 2022, Article ID 7622906, 10 pages
https://doi.org/10.1155/2022/7622906

Research Article

Financial Futures Prediction Using Fuzzy Rough Set and Synthetic Minority Oversampling Technique

Shangkun Deng, Yingke Zhu, Ruijie Liu, and Wanyu Xu

College of Economics and Management, China Three Gorges University, Yichang 443002, China

Correspondence should be addressed to Yingke Zhu; 202012530021115@ctgu.edu.cn

Received 30 June 2022; Revised 15 September 2022; Accepted 8 October 2022; Published 16 November 2022

Academic Editor: Khalid K. Ali

In this research, a novel approach called SMOTE-FRS is proposed for movement prediction and trading simulation of the Chinese Stock Index 300 (CSI300) futures, which is the most crucial financial futures in the Chinese A-share market. First, the SMOTE- (Synthetic Minority Oversampling Technique-) based method is employed to address the sample unbalance problem by oversampling the minority class and undersampling the majority class of the futures price change. Then, the FRS- (fuzzy rough set-) based method, as an efficient tool for analyzing complex and nonlinear information with high noise and uncertainty of financial time series, is adopted for the price change multiclassification of the CSI300 futures. Next, based on the multiclassification results of the futures price movement, a trading strategy is developed to execute a one-year simulated trading for an out-of-sample test of the trained model. From the experimental results, it is found that the proposed method averagely yielded an accumulated return of 6.36%, a F1-measure of 65.94%, and a hit ratio of 62.39% in the four testing periods, indicating that the proposed method is more accurate and more profitable than the benchmarks. Therefore, the proposed method could be applied by the market participants as an alternative prediction and trading system to forecast and trade in the Chinese financial futures market.

1. Introduction

As a crucial part of the world financial markets, the Chinese financial futures market could have a significant impact on the global economy [1, 2]. Stock index futures, which are efficient financial derivatives for hedging trading risk, have become more and more popular among market participants, and numerous scholars have conducted research on their price predictions [3–7]. With the fast development of communication technology, the ability of investors to capture opportunities in the shorter term gradually increases [8]. Subsequently, there is an increasing number of individual and institutional investors participating in High-Frequency Trading (HFT), and many researchers have focused on the studies of high-frequency price forecasting [9, 10]. However, some scholars found that the traditional methods are difficult to achieve a satisfactory performance due to the nonlinear and uncertain character of financial time series [11, 12].

In the last few decades, with the rapid development of artificial intelligence (AI) technologies, machine learning-based approaches have been widely applied to the analysis of massive and nonlinear data in various applications, which include the finance field [13–15]. Among them, the fuzzy set and rough set are efficient tools for analyzing complex and nonlinear information with high noise and uncertainty. Thus, some researchers combined fuzzy set- and rough set-based theories to solve relevant problems. For instance, Dubois and Prade designed the fuzzy rough set-based method by combining two theories [16, 17], and it has been widely applied by many researchers. The complex and nonlinear concept is approximated by the fuzzy rough lower and upper approximation, and it allows the elements to be recognizable from each other to some extent, rather than being either discernible or not.

With the rapid development of technology, an increasing number of investors prefer high-frequency trading [9, 10].

However, the performance of a trading decision support model will be affected by significant differences in the base price of various stocks [18]. Therefore, their trading decision support systems tend to forecast price movements as a trading signal for the trading strategies. Additionally, for solving the multiclassification problem of financial price movement prediction, the training samples of each class are usually unbalanced, which could lead to biased prediction results and unsatisfactory accuracy [19, 20]. Therefore, it is also necessary to balance the sample labels of price direction and magnitude for the CSI300 futures.

In this research, by integrating the SMOTE-based oversampling method and fuzzy rough set (FRS), we propose a high-frequency price trend multiclassification and simulation trading method for the CSI300 futures, which is the most crucial financial futures in the Chinese A-share market. The SMOTE-based method is adopted to balance the label ratios, and the FRS is employed as the base classifier for price movement prediction. Based on the multiclassification prediction results, we also design a trading strategy for simulation trading. The main contributions of this study could be summed up as follows: (1) by integration of SMOTE and FRS-based methods, a novel price movement multiclassification and simulation trading approach is developed for the CSI300 futures; (2) the SMOTE-based method is applied in this study to deal with the unbalanced samples, which effectively avoided biased prediction results and improved the prediction accuracy; and (3) a trading strategy based on the multiclassification results is designed for enhancing the trading performance of the proposed method.

The rest of this article is arranged as follows: Section 2 introduces the related works of this study. The background of relevant methods is described in Section 3. In Section 4, we provide an explanation of the proposed method in detail. The experimental results are reported and discussed in Section 5. In Section 6, we conclude this study and provide several research directions.

2. Related Work

In the last two decades, machine learning-based methods have been widely used as an efficient and remarkable classification and regression tool in the financial fields. For instance, Lin et al. constructed a novel ensemble machine learning method with six commonly used machine learning algorithms including SVM (Support Vector Machine), RF (Random Forest), and KNN (K-Nearest Neighbor) to predict the daily price movements of stocks in the Chinese stock market. The experimental results show that the accuracy and profitability of their proposed method outperformed the traditional methods [21]. Kamalov proposed a Neural Network- (NN-) based method for significant change prediction in stock price, and the experimental results show that the proposed method obtained the best accuracy [22]. Yu and Yan developed a stock price prediction model based on a deep learning- (DL-) based algorithm, and they concluded that their proposed method produced a larger prediction accuracy than traditional models [23]. However, those methods not only require a large amount of complete data but also need preprocessing prior to the model training.

The fuzzy set and rough set, as efficient tools in machine learning algorithms for analyzing complex and nonlinear information with high noise and uncertainty, have been widely applied in the financial fields. For instance, Sun et al. proposed a price prediction model for the stock index in the Chinese stock market by combining the traditional fuzzy time series model and rough set method [24]. Kumar et al. proposed a stock price forecasting method based on the fuzzy set, and they tested it in the Indian stock market. Experimental results showed that the proposed method outperformed the benchmark methods [25]. In addition to these applications of fuzzy sets and rough sets to build classifiers for forecasting stock prices, it is also widely used for reducing data dimensionality [26, 27]. Jensen et al. proposed a novel hybrid fuzzy rough rule induction approach, which combines the process of rule induction and attribute reduction. They improved the greedy hill-climbing strategy, which made it perform better than the benchmark methods [28, 29]. Thus, in this article, the CSI 300 index futures prediction is selected as the research object, and the approach proposed by Jensen et al. [28, 29] is employed to generate rules for its price change prediction.

Additionally, for solving the multiclassification problem of financial price movement prediction, the training samples of each class are usually unbalanced, which leads to biased classification results and low accuracy [30, 31]. The Synthetic Minority Oversampling Technique (SMOTE), which was proposed by Chawla et al. [32], is an efficient method for solving unbalanced samples by oversampling the minority [33], and it has been successfully and widely applied in many fields [33–36]. Therefore, following the research of Chawla et al. [32], the SMOTE-based approach is employed and integrated into the proposed method to balance the model training samples of different classes before the model training of fuzzy rough set (FRS).

3. Background

The fuzzy set approach can be used to handle fuzzy data, while rough sets can deal with incomplete information. By expanding equivalence relations in rough sets to fuzzy equivalence relations, it results in an integration of rough set and fuzzy set theories [37–39]. For variables x, y, z in U $(\forall x, y, z \in U)$, the fuzzy equivalence relation R should satisfy the following three properties: (1) reflexivity: $\mu_R(x, x) = 1$; (2) symmetry: $\mu_R(x, y) = \mu_R(y, x)$; and (3) transitivity: $\mu_R(x, z) \geq \mu_R(x, y) \wedge \mu_R(y, z)$. The partition of U, generated by the associated equivalence relation R_P of nonempty finite set P of attributes, $U/P = \{F_1, \cdots, F_m\}$, which can be calculated by using the conjunction of constituent fuzzy equivalence classes F_i $(1 \leq i \leq m)$. For any fuzzy concept X in the universe of discourse to be approximated $(\forall X \in U)$, the fuzzy lower and upper approximations are redefined as

$$\mu_{\underline{P}X}(x) = \sup_{F \in U/P} \min \left(\mu_F(x), \inf_{y \in U} \max \{1 - \mu_F(y), \mu_X(y)\} \right),$$
$$\mu_{\bar{P}X}(x) = \sup_{F \in U/P} \min \left(\mu_F(x), \sup_{y \in U} \min \{\mu_F(y), \mu_X(y)\} \right), \tag{1}$$

where the tuple $\langle \bar{P}X, \underline{P}X \rangle$ that generated from the fuzzy lower and upper approximations is the fuzzy rough set. The fuzzy positive region can be defined as

$$\mu_{\mathrm{POS}_P(Q)}(x) = \sup_{X \in U/Q} \mu_{\underline{P}X}(x). \tag{2}$$

In addition, the fuzzy rough dependency function could be defined as follows:

$$\gamma'_P(Q) = \frac{\sum_{x \in U} \mu_{\mathrm{pos}_P(Q)}(x)}{|U|}. \tag{3}$$

The dependency of Q on P is equal to the proportion of identifiable objects in the entire dataset, which corresponds to determining the fuzzy cardinality of positive region $\mu_{\mathrm{POS}_P(Q)}(x)$ divided by the total number of objects in the universe U. R is the approximation of the set C for all conditional properties when $\gamma'_{R-|a|}(D) \neq \gamma'_C(D)(\forall a \in R)$ and $\gamma'_R(D) = \gamma'_C(D)$.

For the fuzzy rough rule induction and feature selection approach proposed by Jensen et al., it merges the processes of rule induction and feature selection, and it improves the hill-climbing strategy of the original algorithm, which can generate a rule on the fly that completely covers the training samples [28, 29]. Equation (4) is used to assess the quality of approximation of all conditional attributes. The core features are identified through the dependency change of the full set of the conditional features when the individual attributes are removed:

$$\mathrm{Core}(C) = \left\{ a \in C | \gamma'_{C-\{a\}}(Q) < \gamma'_C(Q) \right\}. \tag{4}$$

A subset of the attribute set that maintains invariance with the fuzzy rough positive region is then defined as the relative reduction, and each rule generated from the fuzzy rough set will contain a more compact subset [29, 37].

4. Proposed Method

In this study, a novel approach SMOTE-FRS is proposed for the price movement multiclassification of the CSI300 futures. The main structure of the proposed method is presented in Figure 1. There are mainly four parts of the proposed method: (1) Data preprocessing part. In this part, the 1 min frequency trading data of the CSI300 futures are collected and transformed into the 1-hour frequency data and features. Then, the datasets containing the normalized data of features are divided into several training and testing datasets. (2) Training sample reconstruction part. The SMOTE-based approach is employed for minority class oversampling and majority class undersampling in the training dataset to generate a balanced group of training samples. (3) Signal generation part. The training datasets are used for model training to generate trading signals based on the fuzzy rough rule (see more details in Section 5.2). (4) Simulated trading and result evaluation part. In this part, a predesigned trading strategy is applied, and simulated trading is carried out for one year of out-of-sample testing. Finally, three evaluation indicators are employed to judge the prediction performance and profit-making ability of the proposed method.

5. Experimental Design

5.1. Data Preprocessing. In the data preprocessing part, first, the 1 min frequency trading data of the CSI300 futures that range from January 2020 and December 2021 is derived from the Choice Database (the formal website of the Choice Database is http://choice.eastmoney.com/). The trading data for experiments consists of the open and close prices, trading volume, and open interest in the 1 min timeframe. The original data are used to calculate the hourly return (Return), volume change rate (VCR), and the open interest change rate (OICR). The calculation ways of those indicators are shown as Equation (5). The indicators within the ten hours prior to the prediction points are then standardized to provide the prediction features for the initial input datasets, as listed in Table 1, in which the Return, VCR, and OICR are denoted by R, V, and H, respectively. For instance, R_4 represents the Return four hours before the forecasting point. Then, the entire dataset is separated into training and testing datasets with a ratio of about 4 : 1. Next, the SMOTE-based approach is used to address the sample unbalanced problem by oversampling the minority class and undersampling the majority class of samples. Details about the unbalanced sample processing are reported in Table 2, in which label = 1, 2, 3, 4 are multiclassification classes that represent the small rise, large rise, small fall, and large fall in price, respectively. Label = 0 represents the minor changes in price that do not meet the transaction conditions. Additionally, the experiment dataset window will be slid forward one period (three months) by the sliding window technique after one round of model training and testing, and the entire testing period lasts for one year in total. Details of the experiment data design are provided in Table 3:

$$\mathrm{Return}_t = \frac{\mathrm{Close}_t - \mathrm{Open}_{t-9}}{\mathrm{Open}_{t-9}},$$
$$\mathrm{VCR}_t = \frac{\mathrm{Volume}_t - \mathrm{Volume}_{t-9}}{\mathrm{Volume}_{t-9}}, \tag{5}$$
$$\mathrm{OICR}_t = \frac{\mathrm{OI}_t - \mathrm{OI}_{t-9}}{\mathrm{OI}_{t-9}},$$

where Close_t, Open_t, Volume_t, and OI_t, respectively, represent the closing price, opening price, trading volume, and open interest at the t hour.

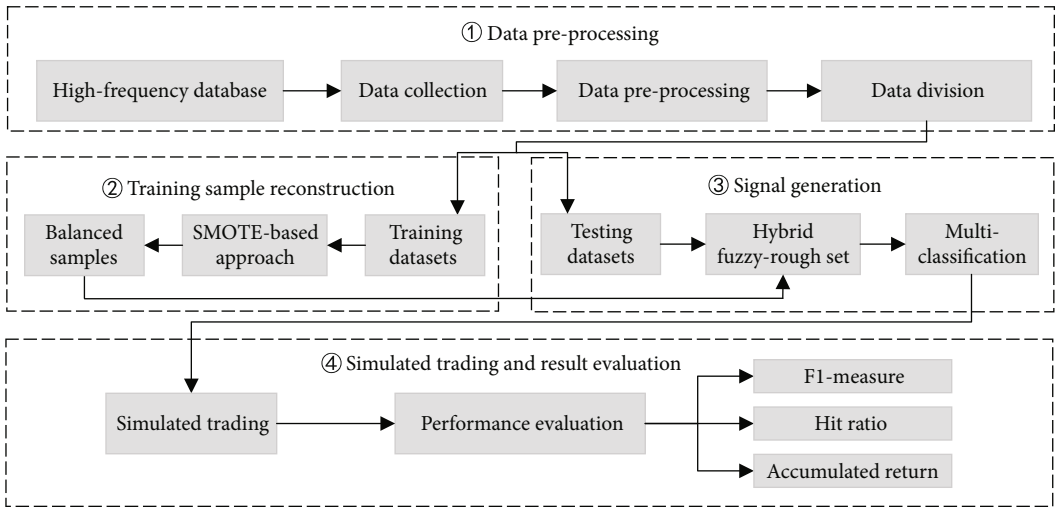

Figure 1: The main structure and working procedures of the proposed method SMOTE-FRS.

Table 1: The list of input features for multiclassification.

Indicator	Input features
Hourly return (Return)	$R_{10}, R_9, R_8, R_7, R_6, R_5, R_4, R_3, R_2, R_1$
Volume change rate (VCR)	$V_{10}, V_9, V_8, V_7, V_6, V_5, V_4, V_3, V_2, V_1$
Open interest change rate (OICR)	$H_{10}, H_9, H_8, H_7, H_6, H_5, H_4, H_3, H_2, H_1$

Table 2: The data preprocessing results of the SMOTE-based approach. Note that the class imbalance ratio = majority class/minority class.

Multiclassification	Label = 0	Label = 1	Label = 2	Label = 3	Label = 4
Class	Majority class	Minority class	Minority class	Minority class	Minority class
The unbalanced samples before processing					
Training period 1	717	99	18	92	21
Class imbalance ratio	\	7.24	39.83	7.79	34.14
Training period 2	770	93	12	80	17
Class imbalance ratio	\	8.28	64.17	9.63	45.29
Training period 3	756	104	11	87	18
Class imbalance ratio	\	7.27	68.73	8.69	42.00
Training period 4	772	90	8	87	11
Class imbalance ratio	\	8.58	96.50	8.87	70.18
The sample numbers after processing by SMOTE					
Training period 1	396	297	54	276	63
Class imbalance ratio	\	1.33	7.33	1.43	6.29
Training period 2	372	279	36	240	51
Class imbalance ratio	\	1.33	10.33	1.55	7.29
Training period 3	416	312	33	261	54
Class imbalance ratio	\	1.33	12.61	1.59	7.70
Training period 4	360	270	24	261	33
Class imbalance ratio	\	1.33	15.00	1.38	10.91

5.2. Trading Strategy Design. The training datasets are used to generate rules based on a fuzzy rough set for multiclassification of the CSI300 futures direction change, which results in labels representing the price changes (expressed as the FR, Forecasting Return) one hour after the prediction. Additionally, a predesigned trading strategy is employed to validate

Table 3: The four subdatasets for model training and model testing.

Subdataset	Model training period	Model testing period
Dataset 1	2020/Jan.–2020/Dec. (1 year)	2021/Jan.–2021/Mar. (3 months)
Dataset 2	2020/Apr.–2021/Mar. (1 year)	2021/Apr.–2021/June (3 months)
Dataset 3	2020/Jul.–2021/June (1 year)	2021/Jul.–2021/Sept. (3 months)
Dataset 4	2020/Oct.–2021/Sept. (1 year)	2021/Oct.–2021/Dec. (3 months)

the prediction accuracy and profitability of the proposed method in trading simulation based on the classification results. An example of multiclassification and trading simulation of the proposed method is plotted in Figure 2. As shown in Figure 2, the hourly return (Return), volume change rate (VCR), and open interest change rate (OICR) within the ten hours prior to the prediction points are employed as the input features, and the FRS is used as the base classifier to forecast the price changes one hour after the forecast points with the output of the price change label (label). If the Forecasting Return (FR) is greater than T_4, the classification label is 2; if FR is larger than T_3 and less than or equal to T_4, the classification label is 1; when FR is greater than or equal to T_2 and less than or equal to T_3, the classification label is 0; if FR is larger than or equal to T_1 and less than T_2, the classification label is 3; if FR is smaller than T_1, the classification label is 4. As reported in Table 4, the multiclassification results are then also used as trading signals to design a trading strategy, which is set out as follows: if the classification label is 2, a long transaction with a leverage of 2 is applied; if the classification label is 1, a long transaction with small leverage of 1 will be used; when the classification label is 0, no transaction will be executed; if the classification label is 3, a short-selling transaction with small leverage of 1 is executed; if the classification label is 4, the proposed method will execute a short-selling transaction with large leverage of 2. Note that the abovementioned T_1, T_2, T_3, and T_4 are the level thresholds, in which T_1 is set to -0.02, T_2 is set to -0.01, T_3 is set to 0.01, and T_4 is set to 0.02. Additionally, the value of small leverage is set to 1, and the large leverage value is set to 2. The trading commission is set to 0.1% per transaction. Finally, the position holding period length for each transaction is set to five hours.

5.3. Benchmark Design. For judging the performance of the proposed method SMOTE-FRS, several popular machine learning methods are adopted to design the benchmarks. In the benchmark methods, the SVM, ANN, RF, XGBoost, and the deep learning method multilayer perceptron (MLP) are adopted as the basic classifier for multiclassification of the CSI300 futures movement. Note that for each benchmark method, the SMOTE-based approach is also used by them to produce balanced samples for model training. In addition, the FRS-based method without using SMOTE (FRS-no-SMOTE) is designed as one of the benchmarks, and it is used for testing the functions of the SMOTE method in the proposed method. Furthermore, two classic passive trading strategies, Buy-and-Hold (BAH) and Short-and-Hold (SAH), are employed as benchmark methods to evaluate the performance of the proposed method.

5.4. Performance Evaluation Measures

5.4.1. F1-Measure. In order to evaluate the performance of the proposed model in price change prediction of the CSI300 futures, the F1-measure (see Equation (6)) is employed as the accuracy evaluator based on the results of the confusion matrix (see Table 5):

$$\text{F1-measure} = \frac{2^{*}\text{TPR}^{*}\text{PPV}}{\text{TPR} + \text{PPV}}, \tag{6}$$

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{7}$$

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{8}$$

In Table 5, TP represents the correct times of positive predictions (including small and large rises for the price change, label = 1 or label = 2); TN represents the correct times of negative predictions (both small and large declines for the price change, label = 3 or label = 4); FN indicates the times of positive price changes that are incorrectly predicted as negative changes, and FP denotes the times of negative changes that are incorrectly predicted as positive changes. TPR and PPV stands for true positive rate and positive predictive value, respectively.

5.4.2. Hit Ratio (HR). The HR is a measure of the price direction forecasting accuracy, which can be calculated from

$$\text{HR} = \frac{\text{PF} + \text{NF}}{N}, \tag{9}$$

where PF denotes the times of correct positive forecasting, NF is the times of correct negative forecasting, and N means the total times of direction forecasting.

5.4.3. Accumulated Return (AR). Accumulated return (AR) is an indicator that measures the profitability of the trading system with the formulas shown in

$$\text{AR} = \sum_{n=1}^{N} (P_n * l_n - C), \tag{10}$$

$$P_n = \frac{\text{Close}_{n+1} - \text{Open}_n}{\text{Open}_n}, \tag{11}$$

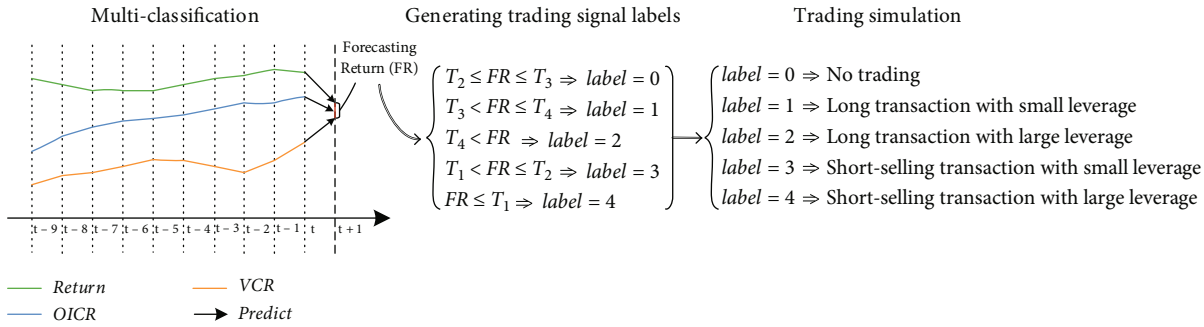

Figure 2: The design of multiclassification and trading simulation of the proposed method.

Table 4: The design of label, threshold, trading signal, and leverage of the trading strategy for the proposed method.

Label	Level threshold	Trading signal	Leverage
0	$[T_2, T_3]$	No trading	N/A
1	$(T_3, T_4]$	Long	1
2	$>T_4$	Long	2
3	$[T_1, T_2)$	Short-selling	1
4	$<T_1$	Short-selling	2

Table 5: The confusion matrix for price movement prediction of the CSI300 futures.

	Positive change	Negative change
Positive prediction	True positive (TP)	False positive (FP)
Negative prediction	False negative (FN)	True negative (TN)

where P_n denotes the return yielded by the nth transaction, which can be calculated from Equation (11), and l_n indicates the leverage chosen for the nth transaction; C denotes the trading cost for each transaction. Note that the trading cost C is zero for the current trading if the current trading signal is the same as the former one, because there is no need to close the position if the current trading signal is identical to the former one. Otherwise, the value of the trading cost C is 0.1% per round trip. N means the total transaction times.

6. Experimental Results

6.1. Multiclassification Results. In this study, the FRS is used as the rule-based classifier for price change multiclassification of CSI300 futures, resulting in the price change labels. The decision rules extracted based on the FRS are in the form of IF-THEN, and some examples of the rules are shown as follows.

Rule 1. IF R6 is around 0.0089 and H3 is around 0.1015 and R1 is around 0.0038 and R5 is around 0.0064 and R8 is around 0.0036 and V8 is around 0.1866 THEN label is 4.

Rule 2. IF R4 is around −0.0113 and H3 is around 0.0918 and R1 is around −0.0012 and R5 is around −0.0082 and R8 is around −0.0005 and V8 is around 0.1072 THEN label is 0.

Rule 3. IF R7 is around −0.0018 and H3 is around 0.1235 and R1 is around 0.0025 and R5 is around 0.0004 and R8 is around 0.0018 and V8 is around 0.0648 THEN label is 3.

Rule 4. IF R10 is around 0.0152 and H3 is around 0.1098 and R1 is around −0.0034 and R5 is around −0.0035 and R8 is around 0.0012 and V8 is around 0.0726 THEN label is 1.

Rule 5. IF R9 is around −0.0010 and H3 is around 0.1130 and R1 is around −0.0003 and R5 is around 0.0002 and R8 is around 0.0025 and V8 is around 0.1536 THEN label is 1.

Based on the decision rules extracted from the training datasets with the FRS, a predesigned trading strategy is applied for transaction simulation with the multiclassification results out-of-sample. The confusion matrix results of the proposed method over the four testing periods are presented in Figure 3, where the horizontal blocks in each subplot indicate the predicted classes and actual classes on the vertical blocks. The darker the color of the blocks, the greater the number of classes.

Based on the confusion matrix results, the F1-measure results of the proposed method and benchmark methods for the testing periods are reported in Table 6. First, as shown in Table 6, the average result of the F1-measure over the four testing periods for the proposed method (SMOTE-FRS) is 65.94%, which is larger than the results of SMOTE-SVM (60.63%), SMOTE-ANN (60.66%), SMOTE-RF (61.59%), and SMOTE-XGBoost (62.02%). Moreover, the results of all benchmark methods experienced at least one F1-measure lower than 60% within the testing periods. It indicates that compared to these traditional machine learning algorithms, the proposed method produced a more accurate and robust performance in price change multiclassification of the CSI300 futures. Although the SMOTE-MLP-based method produced the excellent F1-measure result in the fourth quarter (72.41%), the results within the second and third quarters are less than 65%, while the proposed method consistently yielded F1-measure results

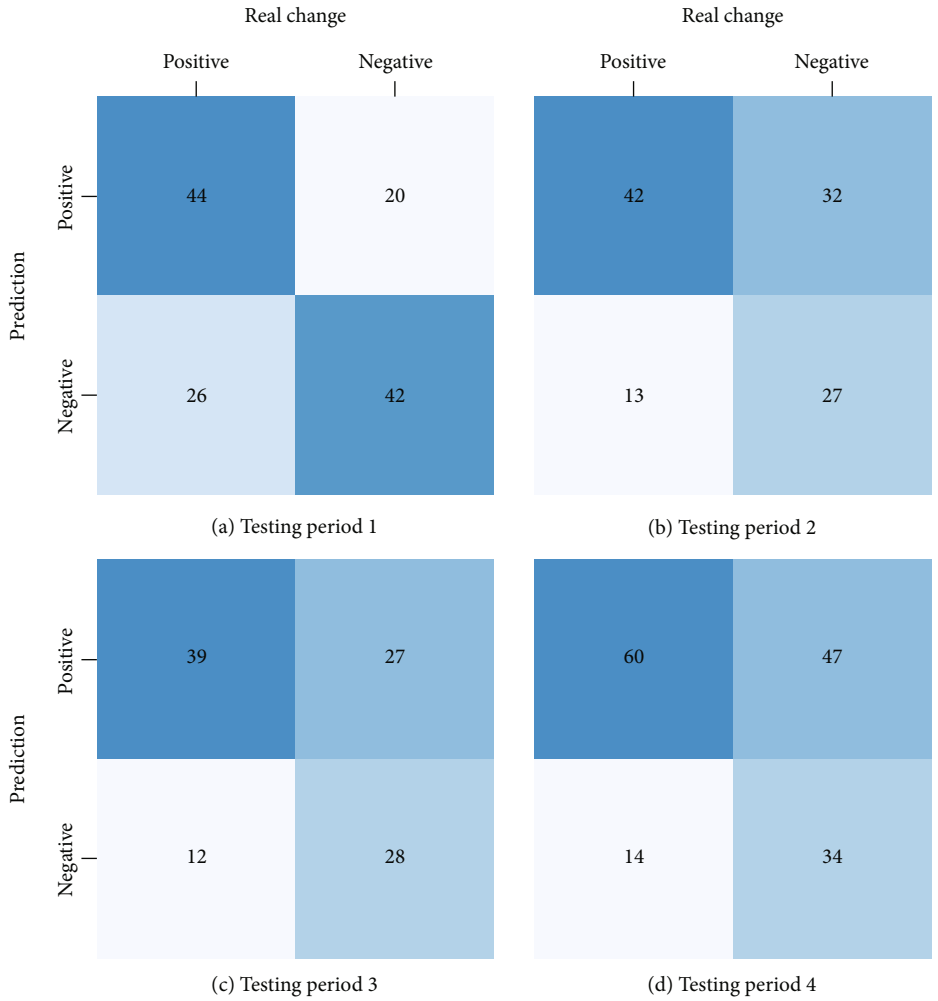

Figure 3: The confusion matrix results of direction prediction for the proposed method in the four testing periods.

Table 6: The F1-measure results of the benchmarks and the proposed method SMOTE-FRS.

Method	SMOTE-SVM	SMOTE-ANN	SMOTE-RF	SMOTE-XGBoost	SMOTE-MLP	FRS-no-SMOTE	SMOTE-FRS
Period 1	68.92%	65.77%	60.53%	61.39%	65.79%	38.10%	65.67%
Period 2	61.31%	51.69%	66.67%	56.10%	62.30%	60.00%	65.12%
Period 3	59.20%	64.58%	61.54%	63.92%	64.46%	60.87%	66.67%
Period 4	53.10%	60.61%	57.63%	66.67%	72.41%	50.00%	66.30%
Average	60.63%	60.66%	61.59%	62.02%	66.24%	52.24%	65.94%

greater than 65% in all four quarters. It can be concluded that although a deep learning-based algorithm may produce a wonderful performance than the traditional machine learning models, while in the case of price trend multiclassification for CSI300 futures, as evidenced by the confusion matrix results in Figure 3, the method proposed in this research successfully produced a more robust performance. Moreover, compared with the F1-measure results of FRS without SMOTE (FRS-no-SMOTE), the proposed method produced a superior prediction performance after adopting the SMOTE-based method to solve the sample imbalance problem.

6.2. Hit Ratio Results. To further evaluate the performance of the proposed method in price change prediction, the hit ratio results produced by the benchmarks and the proposed method are reported in Table 7. First, it could be observed that the average hit ratio of the proposed method in four subtesting periods is 62.39%, which outperforms that of the benchmark methods, including the SMOTE-SVM (59.94%), SMOTE-ANN (59.16%), SMOTE-RF (59.97%), SMOTE-XGBoost (59.57%), and SMOTE-MLP (61.99%). Additionally, the proposed method yielded the best direction prediction accuracy in all of the four subtesting periods,

Table 7: The hit ratio results of the benchmarks and the proposed method SMOTE-FRS.

Method	SMOTE-SVM	SMOTE-ANN	SMOTE-RF	SMOTE-XGBoost	SMOTE-MLP	FRS-no-SMOTE	SMOTE-FRS
Period 1	64.05%	58.96%	61.54%	61.25%	62.60%	35.00%	65.15%
Period 2	58.59%	58.25%	59.32%	57.65%	58.18%	55.56%	60.53%
Period 3	58.54%	60.47%	59.02%	62.37%	61.95%	50.00%	63.21%
Period 4	58.59%	58.95%	60.00%	57.00%	65.22%	33.33%	60.65%
Average	59.94%	59.16%	59.97%	59.57%	61.99%	43.47%	62.39%

Table 8: Friedman test on the hit ratio results for the proposed method SMOTE-FRS against the benchmark methods.

Compared models	Significant level $\alpha = 0.05$
SMOTE-FRS versus SMOTE-SVM	$H_0 : n1 = n2 = n3 = n4 = n5 = n6 = n7$
SMOTE-FRS versus SMOTE-ANN	
SMOTE-FRS versus SMOTE-RF	$F = 15.96$
SMOTE-FRS versus SMOTE-XGBoost	
SMOTE-FRS versus SMOTE-MLP	$p = 0.014$ (reject H_0)
SMOTE-FRS versus FRS-no-SMOTE	

Table 9: The accumulated return results of the benchmarks and the proposed method SMOTE-FRS.

Method	SMOTE-SVM	SMOTE-ANN	SMOTE-RF	SMOTE-XGBoost	SMOTE-MLP	FRS-no-SMOTE	BAH	SAH	SMOTE-FRS
Period 1	3.28%	- 0.35%	- 2.08%	- 11.23%	0.15%	- 1.05%	- 2.62%	2.42%	6.49%
Period 2	- 2.55%	- 3.40%	- 6.38%	- 1.28%	- 14.46%	0.18%	7.98%	- 8.18%	5.90%
Period 3	- 11.63%	4.25%	- 13.16%	0.39%	2.31%	- 4.18%	3.00%	- 3.20%	6.25%
Period 4	- 1.79%	- 7.44%	- 9.67%	- 4.64%	- 3.91%	- 0.53%	4.55%	- 4.75%	6.80%
Average	- 3.17%	- 1.74%	- 7.82%	- 4.19%	- 3.98%	- 1.40%	3.23%	- 3.43%	6.36%

which indicates that compared to the most popular machine learning methods, the proposed method performed better when applied to the price direction prediction of the CSI300 futures. The results of the proposed method are better compared to the FRS without SMOTE (FRS-no-SMOTE), which indicates that the performance of the proposed method can be enhanced after applying the SMOTE-based method to deal with the sample unbalanced problem. Furthermore, the Friedman test [40] is employed to evaluate whether the proposed method performed better than the benchmarks significantly. The Friedman test results of the hit ratio are reported in Table 8, from which we can find that the significance is at the 0.05 level for the one-tailed test, demonstrating that the direction prediction accuracy of the proposed method is significantly better than that of the benchmarks.

6.3. Accumulated Return Result. For market participants, an excellent trading decision support system should not only provide accurate signals of price direction change but also own excellent profit-making ability. Table 9 provides the accumulated return results of the proposed method SMOTE-FRS and all benchmarks. The average return of the proposed method over the four subtesting periods is 6.36%, which is superior to the results of the benchmarks, including SMOTE-SVM (- 3.17%), SMOTE-ANN (- 1.74%), SMOTE-RF (- 7.82%), SMOTE-XGBoost (- 4.19%), FRS-no-SMOTE (- 1.40%), and SMOTE-MLP (- 3.98%). In addition, the return generated by the proposed method in subtesting periods 1-4 is 6.49%, 5.90%, 6.25%, and 6.80%, all of which are positive returns. In contrast, the benchmark methods almost produced negative accumulated return results over the four subtesting periods. Although the classic passive trading strategy BAH produced an outstanding return in the second quarter, the proposed approach was capable of producing a more robust return over four quarters. Therefore, it is evident that the proposed method outperforms benchmark methods in terms of profit-making ability. Furthermore, the Friedman test results for accumulated return are

Table 10: The results of the Friedman test on accumulated return for SMOTE-FRS against the benchmark methods.

Compared models	Significant level $\alpha = 0.1$
SMOTE-FRS versus SMOTE-SVM	$H_0 : n1 = n2 = n3 = n4 = n5 = n6 = n7 = n8 = n9$
SMOTE-FRS versus SMOTE-ANN	
SMOTE-FRS versus SMOTE-RF	
SMOTE-FRS versus SMOTE-XGBoost	$F = 14.8$
SMOTE-FRS versus SMOTE-MLP	
SMOTE-FRS versus BAH	
SMOTE-FRS versus SAH	$p = 0.063$ (reject H_0)
SMOTE-FRS versus FRS-no-SMOTE	

displayed in Table 10. It is observed that the profitability of the proposed method is significantly better than that of the benchmarks at the 0.1 level, demonstrating that the method proposed in this research could be applied as an alternative trading support system for the market participants in the CSI300 futures market.

7. Conclusion

In this paper, we propose a novel approach SMOTE-FRS for high-frequency price prediction and trading simulation of the CSI300 futures. The SMOTE-based method is applied to solve the sample imbalanced problem, while the fuzzy rough set-based approach is employed to generate the movement prediction and simulation trading signal. Moreover, for the purpose of improving the profitability of the proposed method, a predesigned trading strategy was proposed, and one-year simulated trading was carried out for the out-of-sample test. For the proposed method, its average F1-measure was 65.94%, the average hit ratio was 62.39%, and the average accumulated return was 6.36%. In summary, compared to benchmark methods, the proposed method SMOTE-FRS produced the best prediction accuracy and trading profit results. The outstanding performance of the proposed method indicates that the proposed method could be applied as an efficient prediction and trading support system for the market participants. Additionally, employing the SMOTE-based method for solving sample unbalanced problems can effectively improve the performance of the proposed method. In future works, researchers could design a more sophisticated trading strategy to enhance the profitability of the method proposed in this research.

Data Availability

Publicly available datasets were analyzed in this study. This data can be found here: http://choice.eastmoney.com/ accessed on 1st June 2022.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgments

This research was sponsored by the Philosophy and Social Science Research Project of Hubei Provincial Department of Education (grant number 21Q035).

References

[1] D. Su and D. Kong, "Research on the regulation of futures market manipulation based on the evolutionary game theory," *Journal of Residuals Science & Technology*, vol. 13, 2016.

[2] J. Ma, Y. Pan, and Y. Zhang, "Selection of short-term investment strategy-judgment based on average adhesion state," *International Journal of Business and Management*, vol. 12, no. 6, p. 165, 2017.

[3] Y. Zhang and Y. Liu, "Risk aversion of stock index futures," *Journal of Beijing Institute of Technology*, vol. 10, no. 3, pp. 69–72, 2008.

[4] X. Wang, Q. Ye, F. Zhao, and Y. Kou, "Investor sentiment and the Chinese index futures market: evidence from the internet search," *Journal of Futures Markets*, vol. 38, no. 4, pp. 468–477, 2018.

[5] N. V. Voinov, M. K. Voroshilov, S. A. Molodyakov, P. D. Drobintsev, O. V. Prokofiev, and I. V. Zajtsev, "Predicting RTS index futures using machine learning," in *2021 XXIV International Conference on Soft Computing and Measurements (SCM)*, pp. 193–196, St. Petersburg, Russia, 2021.

[6] R. Jiang and C. Wen, "A comparison between parametric and nonparametric volatility forecasting of stock index futures in China," *Emerging Markets Finance and Trade*, vol. 58, no. 9, pp. 2522–2537, 2022.

[7] S. Cong, J. Liu, J. Liu, and X. Zhao, "Research on the price of stock index futures with ARIMA model," 2016, DEStech Transactions on Economics, Business and Management, iceme-ebm.

[8] J. C. Reboredo, J. M. Matías, and R. Garcia-Rubio, "Nonlinearity in forecasting of high-frequency stock returns," *Computational Economics*, vol. 40, no. 3, pp. 245–264, 2012.

[9] R. Savani, "High-frequency trading: the faster, the better?," *IEEE Intelligent Systems*, vol. 27, no. 4, pp. 70–73, 2012.

[10] A. Stenfors and M. Susai, "Liquidity withdrawal in the FX spot market: a cross-country study using high- frequency data," *Journal of International Financial Markets Institutions and Money*, vol. 59, pp. 36–57, 2019.

[11] W. Lu, C. Geng, and D. Yu, "A new method for futures price trends forecasting based on BPNN and structuring data,"

IEICE Transactions on Information and Systems, vol. E102.D, no. 9, pp. 1882–1886, 2019.

[12] S. Deng, Y. Zhu, X. Huang, S. Duan, and Z. Fu, "High-frequency direction forecasting of the futures market using a machine-learning-based method," *Future Internet*, vol. 14, no. 6, p. 180, 2022.

[13] S. Deng, C. Wang, Z. Fu, and M. Wang, "An intelligent system for insider trading identification in Chinese security market," *Computational Economics*, vol. 57, no. 2, pp. 593–616, 2021.

[14] J. Ayala, M. García-Torres, J. L. V. Noguera, F. Gómez-Vela, and F. Divina, "Technical analysis strategy optimization using a machine learning approach in stock market indices," *Knowledge-Based Systems*, vol. 225, p. 107119, 2021.

[15] L. I. Yan and Y. A. Jianhui, "Prediction of the price of stock index futures based on SVM and triangular fuzzy information granulation concerning investors sentiment," *Management Science and Engineering*, vol. 10, no. 3, pp. 28–34, 2016.

[16] S. Vluymans, Y. Saeys, L. D'Eer, and C. Cornelis, "Applications of fuzzy rough set theory in machine learning: a survey," *Fundamenta Informaticae*, vol. 142, no. 1-4, pp. 53–86, 2015.

[17] D. Dubois and H. Prade, "Rough fuzzy sets and fuzzy rough sets," *International Journal of General Systems*, vol. 17, no. 2-3, pp. 191–209, 1990.

[18] E. Akyildirim, O. Cepni, S. Corbet, and G. S. Uddin, "Forecasting mid-price movement of bitcoin futures using machine learning," *Annals of Operations Research*, vol. 1-32, pp. 1–32, 2021.

[19] R. Blagus and L. Lusa, "SMOTE for high-dimensional class-imbalanced data," *BMC Bioinformatics*, vol. 14, no. 1, p. 106, 2013.

[20] G. Douzas, F. Bacao, and F. Last, "Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE," *Information Sciences*, vol. 465, pp. 1–20, 2018.

[21] Y. Lin, S. Liu, H. Yang, and H. Wu, "Stock trend prediction using candlestick charting and ensemble machine learning techniques with a novelty feature engineering scheme," *IEEE Access*, vol. 9, pp. 101433–101446, 2021.

[22] F. Kamalov, "Forecasting significant stock price changes using neural networks," *Neural Computing and Applications*, vol. 32, no. 23, pp. 17655–17667, 2020.

[23] P. Yu and X. Yan, "Stock price prediction based on deep neural networks," *Neural Computing and Applications*, vol. 32, no. 6, pp. 1609–1628, 2020.

[24] B. Sun, H. Guo, H. R. Karimi, Y. Ge, and S. Xiong, "Prediction of stock index futures prices based on fuzzy sets and multivariate fuzzy time series," *Neurocomputing*, vol. 151, pp. 1528–1536, 2015.

[25] S. Kumar, K. Bisht, and K. K. Gupta, "Intuitionistic fuzzy time series forecasting based on dual hesitant fuzzy set for stock market," in *Exploring Critical Approaches of Evolutionary Computation*, IGI Global, 2019.

[26] D. Chen, W. Zhang, D. S. Yeung, and E. C. Tsang, "Rough approximations on a complete completely distributive lattice with applications to generalized rough sets," *Information Sciences*, vol. 176, no. 13, pp. 1829–1848, 2006.

[27] C. Cornelis, M. De Cock, and A. M. Radzikowska, "Fuzzy rough sets: from theory into practice," *Handbook of Granular Computing*, W. Pedrycz, A. Skowron, and V. Kreinovich, Eds., Wiley, 2008.

[28] R. Jensen, C. Cornelis, and Q. Shen, "Hybrid fuzzy-rough rule induction and feature selection," in *2009 IEEE International Conference on Fuzzy Systems*, pp. 1151–1156, Jeju, Korea (South), 2009.

[29] R. Jensen and Q. Shen, "New approaches to fuzzy-rough feature selection," *IEEE Transactions on Fuzzy Systems*, vol. 17, no. 4, pp. 824–838, 2009.

[30] B. Acharjya and S. Natarajan, "A fuzzy rough feature selection framework for investors behavior towards gold exchange-traded fund," *International Journal of Business Analytics*, vol. 6, no. 2, pp. 46–73, 2019.

[31] A. S. Roy and N. Chatterjee, *Forecasting of Indian stock market using rough set and fuzzy-rough set based models*, IETE Technical Review (Institution of Electronics and Telecommunication Engineers, India), 2021.

[32] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[33] D. Elreedy and A. F. Atiya, "A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance," *Information Sciences*, vol. 505, pp. 32–64, 2019.

[34] H. Guan, Y. Zhang, M. Xian, H. D. Cheng, and X. Tang, "SMOTE-WENN: solving class imbalance and small sample problems by oversampling and distance scaling," *Applied Intelligence*, vol. 51, no. 3, pp. 1394–1409, 2021.

[35] N. Mqadi, N. Naicker, and T. Adeliyi, "A SMOTe based oversampling data-point approach to solving the credit card data imbalance problem in financial fraud detection," *International Journal of Computing and Digital Systems*, vol. 10, no. 1, pp. 277–286, 2021.

[36] X. Huang, C. Zhang, and J. Yuan, "Predicting extreme financial risks on imbalanced dataset: a combined kernel FCM and kernel SMOTE based SVM classifier," *Computational Economics*, vol. 56, no. 1, pp. 187–216, 2020.

[37] R. Diao and Q. Shen, "A harmony search based approach to hybrid fuzzy-rough rule induction," in *2012 IEEE International Conference on Fuzzy Systems*, pp. 1–8, Brisbane, QLD, Australia, 2012.

[38] D. Dubois and H. Prade, "Putting rough sets and fuzzy sets together," in *Intelligent Decision Support*, R. Słowiński, Ed., vol. 11 of Theory and Decision Library, Springer, Dordrecht, 1992.

[39] H. Thiele, *Fuzzy Rough Sets versus Rough Fuzzy Sets - An Interpretation and a Comparative Study Using Concepts of Modal Logics*, Technical Representative-30/98 University, Dortmund, Germany, 1998.

[40] J. Derrac, S. García, D. Molina, and F. Herrera, "A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms," *Swarm and Evolutionary Computation*, vol. 1, no. 1, pp. 3–18, 2011.