Journal of Database Management
Volume 34 • Issue 3

Incorporating I Ching Knowledge Into Prediction Task via Data Mining

Wenjie Liu, Nanjing University of Posts and Telecommunications, China
Sai Chen, Nanjing University of Posts and Telecommunications, China
Guoyao Huang, Nanjing University of Posts and Telecommunications, China
Lingfeng Lu, Nanjing University of Posts and Telecommunications, China
Huakang Li, Xi'an Jiaotong-Liverpool University, China
Guozi Sun, Nanjing University of Posts and Telecommunications, China*

ABSTRACT

Many real-world applications require prediction that takes the most advantage of data. Classic data mining mechanisms tend to feed a prediction model pivotal data to achieve a promising result, which needs to be adjusted in different application scenarios. Recent studies have shown the potential of I Ching mechanism to improve prediction capacity. However, the I Ching prediction mechanism has several issues, including underutilized I Ching knowledge and incomplete data conversion. To address these issues, the authors designed a model to leverage I Ching knowledge and transform traditional I Ching prediction processing into data mining. The authors' investigation revealed promising results in the stock market compared to popular machine learning and deep learning algorithms such as support vector machine (SVM), extreme gradient boosting (XGBoost), and long short-term memory (LSTM).

KEYWORDS

Artificial Intelligence, Data Mining, I Ching Knowledge, Machine Learning, Stock Prediction

INTRODUCTION

I Ching is an ancient Chinese document which contains a wide variety of rules. I Ching divination uses Ying and Yang and the Four Signs to explain the laws of the changes in the world and predict all kinds of affairs. Nowadays, researchers leverage the I Ching numerical hexagram model to stock market prediction (Guo & Lu, 2020), and apply hexagrams to business guidance (Chen, 2021). However, previous I Ching-based stock prediction mechanisms have several weak points: Small scale stock market data lead to unconvincing statistical results, oversimplified indicators lead to incomplete initial feature, and many abstractive I Ching concepts cannot transform into available modus.

DOI: 10.4018/JDM.322097 *Corresponding Author

To solve the above problems, the authors integrated I Ching prediction mechanism and machine learning, presenting a prediction model based on I Ching. Traditional I Ching predicting methods mainly utilize the explanations of Yao and hexagram to guide people make reasonable judgments.

Feature selection is an important step in building the I Ching prediction model. The selection algorithm can filter out the features with higher importance from the high-dimensional feature set. Currently, many feature selection methods are widely used in optics science (Huang & Liu, 2021) and medical science field (Liu et al., 2019). Classic integrated feature selection algorithms such as random forest (RF)(Breiman, 2001) integrate the results of feature selection methods by training multiple features. In real-world applications, the Three Vitals especially reflect the influence of nonhuman factors such as national policy and environmental conditions.

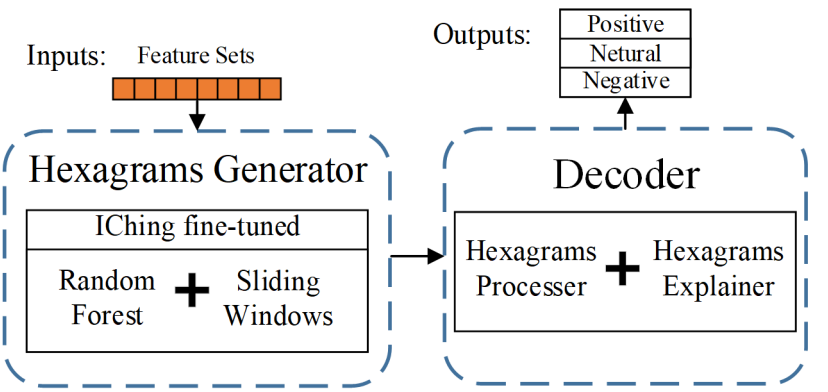
Note. Left: The generator receives unprocessed feature sets. The authors leverage Three Vitals and Six Yao to fine-tuned RF algorithm and sliding windows technique, which improves the prediction capacity and flexibility. Right: The decoder receives the hexagram as input, and extracts hexagrams processer and explainer from "ZhouYi."

The thought of Ten, Chi, and Jin in the "ZhouYi" is also called the Three Vitals, which represent heaven, earth, and human. The Three Vitals assert that human beings are a separate entity isolated from heaven and earth, but also recognize that human behavior can affect the operation of the entire world. Specifically, the authors propose a feature selection method based on machine learning to rank feature importance in the first place, then leverage the Three Vitals concept to filter out the most important feature in current scene. Secondly, the authors utilize sliding windows algorithm to transform features into hexagram; then, they determine the original hexagram and change the hexagram according to the I Ching hexagram generating method. Furthermore, the authors utilize the I Ching explaining method to determine stock trend and finally get the prediction result by comparing predict value and tag value. Combining I Ching and machine learning gives the model more flexibility. The experiment result shows promising outcomes in the stock market prediction task, compared to popular machine learning and deep learning algorithms such as support vector machine (SVM), eXterme Gradient Boosting (XGBoost), attention long short-term memory (LSTM), and gate recurrent unit (GRU).

BACKGROUND

The concept of Yin and Yang originated from the Chinese I Ching culture system and it is the foundation of ancient Chinese philosophical thought (Li, 2014). Yin and Yang theory shows two different states and teaches people to see the world's problems in two ways. In real-world applications,

Figure 1.
Prediction model framework



Yin and Yang can correspond to both good and bad aspects of one thing. Any complex thing can be restricted and divided by the Yin and Yang.

The Four Signs theory expanded from the concept of Yin and Yang (Li, 2014). They are Old Yin, Old Yang, Young Yin and Young Yang, and used to show more states of things. In I Ching knowledge, things will develop in the opposite direction when they become extreme. According to this, the Old Yin and Old Yang will change to Young Yang and Young Yin when they are at extremes. In real-world situation, things do not stay in a certain state all the time. For example, a stock in the securities market will bottom out when it falls to a certain level, and the price will start to rise.

I Ching specifically applies the way of nature to the Three Vitals of heaven, earth, and people. People live in nature, with the earth under their feet and the sky overhead. Every Yaos in the hexagrams corresponds to different locations in the Three Vitals.

The original hexagram presents information that predicts the initial stage of things through a certain method, which focuses on the past and present situation, while the changed hexagram focuses on the future. The hexagrams do not exist independently and need to be combined with word and Yao's word of hexagrams to explain information. The hexagrams and Yao's word are a text sequence used to explain the meaning of the whole hexagrams; there are 64 hexagrams word and 384 Yao's word.

Specifically, there are some hexagrams words showing a dangerous phenomenon and prompting people to stop action or take measures to reduce or avoid losses, while other hexagrams showing positive phenomenon and prompting people to act boldly and gain profit.

RELATED WORK

In recent years, using data mining technology to make macropredictions on data in specific scenarios has gradually become a research hotspot. Modeling the filtered data and predicting future trends is essentially a multivariate time series problem. The most widely used multivariate time series models mainly include autoregressive integrated moving average (ARIMA) and LSTM. The predicted results of ARIMA are close to the historical average and more suitable when there is more noise and real value fluctuations are relatively stable. LSTM achieves good results in multivariate time series; however, as a deep learning model, it requires a large number of training samples which lead to strict restrictions on the real-world applications. ARIMA can achieve better robustness with online learning algorithms (Liu et al., 2016), and LSTM can achieve better long sequence prediction performance with tensorization. When the attention mechanism is added on top of LSTM, it can make predicting time series more accurate. The ARIMA predicted results are close to the historical average; it is suitable when there is more noise and the real value fluctuation is relatively stable.

However, the current macroprediction model is not suitable for every scenario. The accuracy of classic prediction mechanisms in stock prediction task is barely more than 50% (Kannan et al., 2010); the F1 score of RF and SVM in the task of predicting the demand for shared bicycles is up to 62% and 51% separately (Hu, 2013).

In the process of I Ching model prediction, first the authors select six features as the basis of six Yaos. In order to make the selection more scientific and effective, the authors filter out 30 important features by combining data mining algorithm. There are many types of feature selection algorithms including filtering feature selection which can quickly remove noise features, with high computational efficiency and strong versatility. Experiments show significant performance advantages using feature sets identified with precise mutual information in classification tasks (Brunato, 2016).

The wrapping feature selection algorithm is also a typical feature selection method. A wrapping feature selection method leverages missing data to improve the performance of classifier (Cao et al., 2016). However, the wrapping method requires multiple training, the computational overhead is large, compared to the filtering method, and it is not suitable for high-dimensional datasets.

Another feature selection method is embedded feature selection, which can handle high-dimensional datasets, but is prone to overfitting. An embedded feature selection algorithm based

on SVM probabilistic output sensitivity to evaluate the importance of specific features. As a typical integrated feature selection algorithm, RF (Breiman, 2001) has been widely used in data mining and other fields. RF has high prediction accuracy and a very strong tolerance for outliers and noisy data. It has the characteristics of being able to give multiple importance scores while analyzing high-dimensional data. These advantages make RF very suitable for the high-dimensional data research, and it has high application value in the field of data mining.

Figure 2 shows that the Vital Features include two features in each Vital with the order of Ten→Jin→Chi. The Yao information in sequence corresponds to the position of the previous feature. Then, it is necessary to progress eight times the permutation of the interior sequence and compare each one to ordering the Yao sequence. The Hexagram Processor simply follows the rules of the Four Signs theory to transform the YinYao and YangYao.

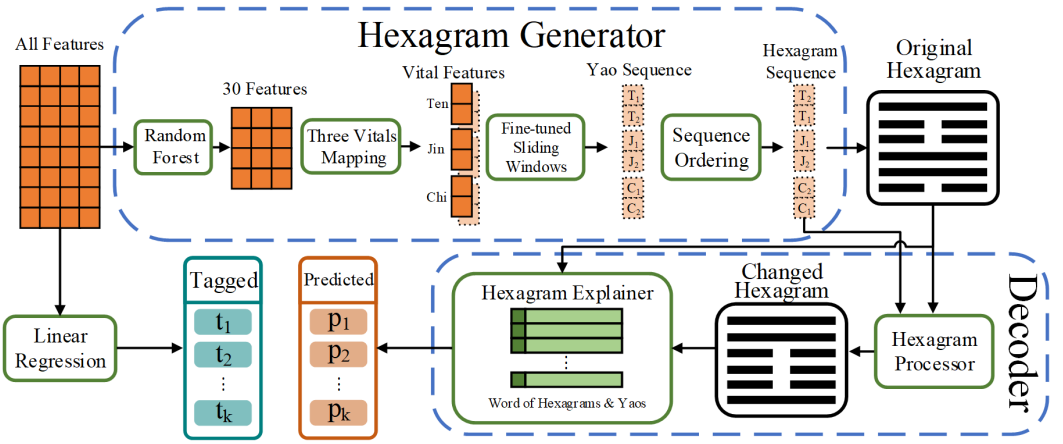
To extend the generality of the model and improve the performance, the authors present a macropredicting model which combines data mining technology and I Ching prediction mechanism. The I Ching predict the future development trending by deciphering the hexagrams. The mechanism of the Three Vitals in the I Ching combined with data mining technology can filter six important features from a large number of features. In addition, through the Four Signs in the I Ching, the Six Yao are converted into original hexagrams, which are 6-bit sequences. After that, the hexagrams explainer generates changed hexagrams according to the original one. Finally, the system analyzes the two sequence and provides the prediction result through hexagrams classifier.

METHODOLOGY

The factors included in the I Ching can be summarized into environmental factor, geographical factor, and human factor; they are also called the Three Vitals, which are Ten, Chi, and Jin. In a macroperspective, Ten indicates the overall environmental status of things which will not be influenced by people, Chi indicates physical properties and objective laws, and Jin characterize the thing itself. These macroscopic features need to be combined with the actual scene when applied in the real-world application.

When applied to the stocks market, Ten corresponds to the national policy, the world economic situation, and other overall environmental factors that are not affected by securities companies, Chi corresponds to company size and fund flow data, And Jin corresponds to company fundamental information (e.g., profit margin and debt ratio).

Figure 2.
Structure of the I Ching prediction model



Another important property of the classical I Ching prediction system is the Four Signs, which embodies the dynamic nature of the I Ching. In this paper, the authors adopt the method of sliding windows to restore the prediction method of the I Ching. Specifically, since the window is sliding, the Four Signs value formed by a data mapping at different time points is not fixed. The classical I Ching predicting method is also considered in the hexagram decoder. Researchers in traditional I Ching use a variety of methods to infer the change in hexagrams, but essentially they use a binomial distribution in a random process which performed three times. Therefore, the initial size of the sliding windows in the I Ching prediction model is set to eight. In this way, the authors make modern algorithmic techniques correspond to traditional I Ching predicting methods and also improve the ability of the model to fit nonlinear relationships.

LABEL CONSTRUCTION

There are more noise data in stock market scenarios, and Huber regression has good robustness, which can solve the noise points in the data very well. The authors mainly use Huber regression to fit the closing price and obtain the trending. Specifically, using a single linear regression, the expression is: $y=ax+b$. y represents the predicted value of the dependent variable, x represents a single independent variable, and a and b are the undetermined parameters of the regression model, where a is also called the regression coefficient. The loss function of Huber regression is Huber loss, and it is calculated as follows:

$$L_{\delta}\left(y, f\left(x\right)\right)=\begin{cases}\frac{1}{2}(y-f\left(x\right))^{2} & \text{for}|y-f\left(x\right)|\leq\delta\\ \delta\left(\left|y-f\left(x\right)\right|-\frac{1}{2}\delta\right), & \text{otherwise}\end{cases} \tag{1}$$

where $\boldsymbol{\delta}$ is the hyperparameter that needs to be adjusted.

Through continuous iteration, the loss value is minimized to obtain the optimal fitting function, and the regression coefficient a in the function is obtained. According to the normal distribution, the distribution of a is counted, and finally the label category is obtained according to the distribution interval of a.

FEATURE SELECTION

In the RF algorithm, the contribution of each feature on each classification tree can be compared and the Gini index can be calculated to judge the importance of each feature. The Gini index of the node indicates the impurity of the node. The contribution of each risk indicator to the total risk is assessed as the percentage of the average Gini reduction of the indicator to the sum of the average Gini reduction of all indicators. The formula for calculating the Gini index is as follows:

$$GI_{m}=\sum_{k=1}^{K}\hat{p}_{mk}\left(1-\hat{p}_{mk}\right) \tag{2}$$

K is the number of categories of the sample set, $\hat{p}_{mk}$ is the probability that the node m belongs to the k-th sample, and, when it is a binary classification (K=2), the Gini index of the node m is:

$$GI_m = 2\hat{p}_m\left(1-\hat{p}_m\right) \tag{3}$$

$\hat{p}_m$ is the probability that the sample point belongs to any category at node *m*. The importance of variable *Xj* at node *m* is as follows:

$$VIM_{jm}^{(Gini)} = GI_m - GI_l - GI_r \tag{4}$$

where GI_l and GI_r respectively represent the Gini indices of two new nodes split by node *m*.

If the variable *Xj* appears *M* times in the tree *i*, the importance of the variable *Xj* in the tree *i* is:

$$VIM_{ij}^{(Gini)} = \sum_{m=1}^{M} VIM_{jm}^{(Gini)} \tag{5}$$

The Gini importance of variable *Xj* in the RF is defined as follows:

$$VIM_{j}^{(Gini)} = \frac{1}{n}\sum_{i=1}^{n} VIM_{ij}^{(Gini)} \tag{6}$$

where *n* is the number of classification trees in the RF.

CALCULATION OF THE FOUR SIGNS AND GENERATION OF HEXAGRAMS

By sorting the importance of each feature to the Gini index of the label, the top 30 features with Gini index importance are finally selected. It is necessary to map and classify the physical meaning of stock characteristics with the thoughts of the Three Vitals and then divide all the features into three categories. After filtering out 30 features through the RF algorithm, the 30 features can be mapped to the three categories of the Three Vitals according to the correspondence table between the features and the Three Vitals categories.

Finally, it is necessary to select the two features with the highest importance in each category, which contains a total of six features, and arrange the filtered features in the order of Chi, Jin, and Ten. After obtaining the six features, according to the I Ching theory, the values of the six features can be converted into the values of the Four Signs. The authors utilize the sliding windows mechanism to reflect the dynamic characteristics of I Ching and calculate the Four Signs value.

```
Algorithm One: The Four Signs Generator Algorithm
Input: Windows size k, Four Signs ratio old_rate = 1/8, young_
rate = 3/8, time series data list vec={vl,...vn}, ordered list
sortedvec, and the subscript of the element earliest added to the
list m.
Output: Four Signs value list sixiang_list.
1: for i = 0 to k-1 do
2: inivec[i] = vec[i]
3: for j = k to n-1 do
4: sortedvec = sort(inivec)
5: maxv = max(sortedvec), minv = min(sortedvec)
```

```
6: cha = maxv-manv
7: A = old_rate*cha, B = young_rate*cha.
8: laoyin=[minv,minv+A), shaoyin=[minv+A,min+A+B),
shaoyang=[min+A+B,min+A+2B), laoyang=[min+A+2B, maxv)
9: if vec[j] ∈ laoyin
10:  sixiang_list[j-k] = 6
11:  else if vec[j] ∈ shaoyin
12:  sixiang_list[j-k] = 8
13:  else if vec[j] ∈ shaoyang
14:  sixiang_list[j-k] = 7
15:  else
16:  sixiang_list[j-k] =9
17:  del sortedvec[m],sortedvec[k] = vec[j]
18: end if
19: end for
```

The hexagram is composed of Yaos. According to the I Ching thought, the YangYao is defined as 1, and the YinYao is defined as 0. Then, mapping the Four Signs value with 1 and 0, the values 6 and 8 represent YinYao, and 7 and 9 are YangYao. With the sliding windows mechanism, it is possible to generate the quarterly 6-bit sequence for each stock. The original hexagram is simply obtained by sorting the sequence in order of Chi, Jin, and Ten.

Algorithm Two: Hexagrams Explainer
Input: Original hexagram $Orig_i$, changed hexagram $Change_i$, changed Yao count $YbNum_i=0$, 64 hexagrams and explanations mapping dictionary *Gua_dict*, 384 Yao and explanations mapping dictionary *Yao_dict*, the unchangeed Yao of the Origi *StaY*, Yao when the high position will not change of the Origi *Hp_StaY*, and Yao when the low position will change of the Origi *Lp_UstaY* .
Output: Explanation list of original and changed hexagrams.

```
1: for k = 0 to 5 do
2: if Orig_i[k] != Change_i[k]
3: flag_i = 1, YbNum_i= YbNum_i+1
4: end if
5: end for
6: if flag_i = 0
7: res_i = Gua_dicti[Orig_i]
8: end if
9: else
10: if YbNum_i = 1
11: res_i  = Yao_dict[StaY]
12: else if YbNumi = 2
13: res_i  = Yao_dict[Hp_StaY]
14: else if YbNumi = 3
15: res_i  = Gua_dict[Orig_i]
16: else if YbNum_i = 4
17: res_i = Yao_dict[Lp_UstaY]
18: else if YbNumi = 5
19: res_i = Yao_dict[StaY]
20: else
21: res_i  = Gua_dict[Change_i]
22: end if
```

According to the I Ching mechanism, the hexagrams will change when there are Old Yang Yao and Old Yin Yao in the Six Yao sequence. The rule is: Every "1" in Old Yang Yao change to "0" and every "0" in Old Yin Yao change to "1". Finally, the hexagrams generator outputs are determined as an original hexagram and a changed hexagram.

The decoder is designed to define the categories corresponding to the 64 hexagrams in the I Ching and explanations in each hexagram. There are 64 different hexagrams explanations and 384 Yao explanations which are strictly connected with each hexagram and the position of Yao. To follow the original explaining process, the authors extract the hexagrams explainer algorithm (Algorithm Two) from the classic I Ching mechanism.

The order of Yaos in the Three Vitals is uncertain and need 2*2*2=8 times permutation, totally. During training, the authors use the eight different explainer results as the predicted values and compare them with the label values, respectively. Finally, the authors choose the best performing Three Vitals order to build the I Ching prediction model in current scenario.

The data in the stocks market is selected to verify the validity of this model, because the data in the financial field are easy to obtain and the amount of data is large enough. The authors' proposed model proposed is also applicable for data in other fields. The construction steps of other domain models are described below.

```
Algorithm Three: Overall Model Construction
Input: All features allfeature_list, meaning of category tian
tian_label, meaning of category di di_label, meaning of category
class ren ren_label, four Signs generator slideWin_algorithm,
the formation rules of this hexagram guaXiango, rules for
the formation of change of fortune guaXiangc, and divination
Interpretation rules Hexagram_Explain.
Parameter: tian_list,di_list,ren_list,featurek,sixfeature_
list,labelSet
Output: Prediction results of divination jiagua_result.
1: for i = 0 to len(allfeature_list) do
2: if allfeature_list[i]∈tian_label
3: tian_list.append(allfeature_list[i])
4:         else if allfeature_list[i]∈di_label
5:         di_list.append(allfeature_list[i])
6:         else
7:         ren_list.append(allfeature_list[i])
8:         end if
9:         end for
10: linearRegression(featurek)->labelSet
11: randomForest(allfeature_list) and (tian_list,di_list,ren_
list)->sixfeature_list
12: fourSigns = slideWin_algorithm(sixfeature_list)
13: benGua = guaXiango(fourSigns)
14: bianGua = guaXiangc(bengua,fourSigns)
15: jiagua_result = Hexagram_Explain(benGua,bianGua)
16: comparison between jiagua_result and label
17: get model evaluation results
```

The above algorithm process is explained as follows:

1. All the features are divided into three categories called the Three Vitals, namely Ten, Chi, and Jin. The Ten corresponds to the characteristics that can reflect the development prospects of the field, the Chi corresponds to the characteristics that can reflect the current market development

of the field, and the Jin correspond to the characteristics that can reflect the development level of the field itself.
2. It is necessary to find out the features that can reflect the development trend of the field, and use the linear regression algorithm to construct the label set based on those features.
3. The RF algorithm is used to calculated the importance of all features, and the top two features in each Three Vital are selected. Finally, six features are obtained.
4. The data corresponding to the selected six features are converted into the Four Signs values by the sliding window method.
5. According to the generation principle of the hexagrams in the I Ching, the original hexagram is generated through the Four Signs values.
6. According to the principle of changing hexagrams in the I Ching, the changed hexagrams are generated from the original hexagrams and the prediction results are obtained by comprehensively judging the original hexagrams and the changing hexagrams. Finally, The prediction accuracy is obtained by comparing the prediction result with the label value.

EXPERIMENTS

Experimental Dataset

It is necessary to set up a specific prediction scenario to verify the performance of the I Ching prediction model in real-world application. The following points have to be considered: Open datasets with easy access, large scale and easy to maintain and update datasets, and a scenario as figurative as possible. Thus, the authors finally choose stock market as experimental scenario. The trends of stocks are divided into three categories: Falling, stable, and rising. Linear regression constructs a linear representation by finding the relationship between independent and dependent variables which are often used for prediction tasks. Optimized linear regression algorithm can significantly reduce predict errors and perform even better than multiple deep learning algorithms in specific scenario. Therefore, the authors also construct the label set using linear regression. Specifically, they fit the trend of closing prices over a certain period of time by linear regression and get regression coefficient. Label sets are constructed from categories, according to the range of regression coefficients.

The stock market datasets the authors used in the experiment are crawled from the NetEase Finance Web site, which collected 3000 different stocks data from 2010 March to 2020 March. The datasets contain two parts, namely financial data and money flow data. Specifically, financial data report the company's net profit margin, debt ratio, and other data reflecting the company's performance in each quarter, and money flow data report the company's opening price, closing price, turnover rate, and other capital flow data daily. All stocks are divided into 10 sectors according to industry, and the experiment predicts the stock trending in the next year based on the historical data of the past quarter.

Evaluation Indicators

Commonly used evaluation indicators for classification problems are precision, recall, and F1 score. Table 1 shows the confusion matrix.

Precision refers to the proportion of all correct predictions to the total, and the formula is as follows:

Table 1.
Confusion matrix

	Positive	Negative
True	True positive	True negative
False	False positive	False negative

$$P = \frac{TP}{TP + FP} \tag{7}$$

Recall refers to the proportion of correct positive predictions to all correct predictions. The formula is as follows:

$$R = \frac{TP}{TP + FN} \tag{8}$$

F1 score is an indicator that combines P and R. The formula is as follows:

$$F1 = \frac{2PR}{P + R} = \frac{2TP}{2TP + FP + FN} \tag{9}$$

The F1 score can evaluate the classifier performance comprehensively. To verify the capacity of the model, the authors mainly conducted two groups of comparative experiments: (1) They compared the I Ching prediction model to other classic algorithms with and without industry differences; (2) they tested the influence of industry differences on the I Ching prediction model.

EXPERIMENTAL RESULTS AND ANALYSIS

The authors selected classic machine learning algorithms for comparison experiments, including SVM, XGBoost, and k-nearest neighbor (KNN). XGBoost is an implementation of boosting algorithm (Tianqi, Chen et al., 2016).

Figure 3 shows that the I Ching model performs better than other classic algorithms in the stock market prediction task (Figure 3a), accuracy increases with prediction length and converges when the prediction length reaches 12 months (Figure 3b), and, unlike other algorithms, the I Ching prediction model shows significant performance improvement in industry modeling and Id.x represents the industry the authors mentioned above in the same order of dimensional disaster problem and prone to overfitting and underfitting (Figure 3c).

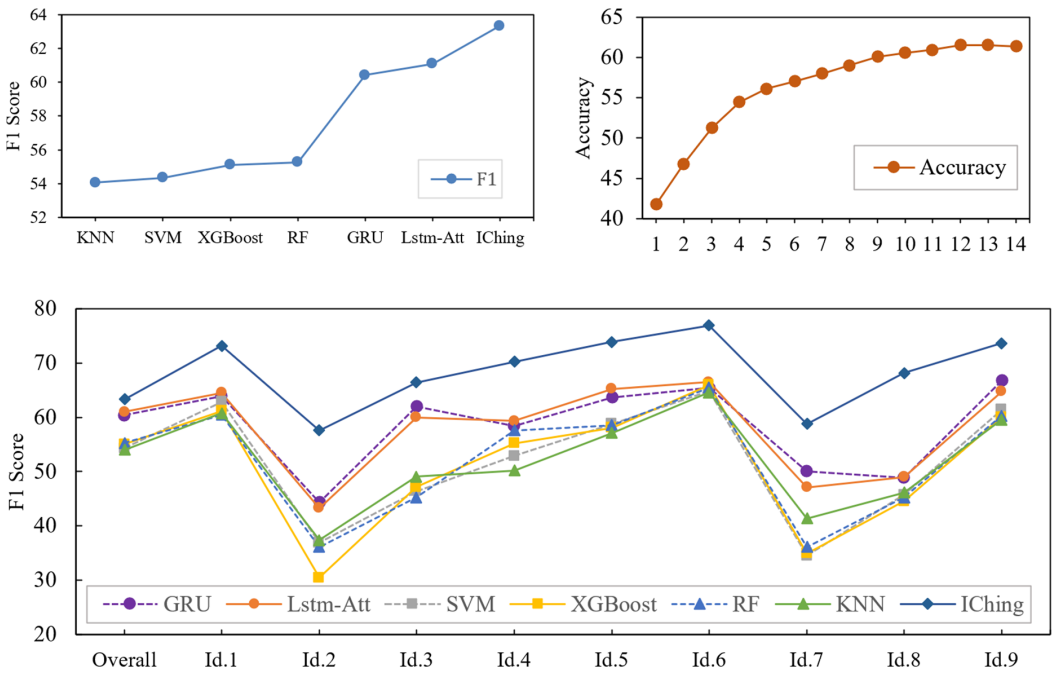
XGBoost is composed of multiple basic learners, and each iteration is based on the difference between the previous value and the target value. The predicted value of XGBoost consists of all the previous values, which gives the model good antioverfitting ability. Another advantage of XGBoost is parallel computing, which improved the training speed.

SVM is a data-oriented classification algorithm (Cortes & Vapnik, 1995). Based on linear features of training data, there are three types of SVM learning model: Linear SVM, linearly separable SVM, and nonlinear SVM. The advantages of SVM are it has good performance in small samples and it solves the dimensional disaster and computational complexity problems. However, SVM is not suitable for solving multiclass problems.

KNN can be used for classification or regression. KNN judges sample characteristics according to the neighbors of the sample in the feature space. The advantages of KNN are it is easy to implement, requires no training, and is especially suitable for multiclassification problems. Nevertheless, KNN will show evident error when the sample is unbalanced.

Attention LSTM uses the attention mechanism in LSTM, and leads to better results. In real-world application, the feature importance of long-term series is mutative and cannot be distinguished by LSTM. The attention mechanism can solve this problem by weighting the hidden layer output of LSTM. In model building, the attention layer and the fully connected layer need to be added after the LSTM layer.

Figure 3.
F1 score of classic algorithms, accuracy of increasing prediction length, and F1 score of industry modeling



GRU is a type of recurrent neural network, and is proposed to solve problems such as long-term memory and gradients in backpropagation. Compared with LSTM, GRU uses an update gate to control the amount of data that the previous memory information can continue to retain until the current moment, and a reset gate to control how much past information to forget.

Classic machine learning methods perform better than deep learning because stock closing price is affected by many factors, and it is difficult for the neural network model to learn data with weak regularity. On the other hand, classic machine learning algorithms are more difficult to solve.

The performance of the I Ching prediction model is higher than other algorithms in all evaluation indicators, except recall on the stock datasets. The I Ching prediction model can select different indicators as the basis for predicting according to the different conditions, and has good adaptability in various scenarios.

Table 2.
Model performance comparison in stock market with classic algorithm

Algorithm	Precision	Recall	F1 score
SVM	45.51	67.46	54.35
XGBoost	46.31	68.05	55.11
RF	46.46	68.16	55.26
GRU	60.15	60.71	60.43
LSTM-Att	61.85	60.33	61.08
KNN	52.00	56.77	54.05
I Ching	65.41	61.51	63.33

Table 3.
Experimental results of I Ching industry modeling

Industry	P	R	F1 score
Overall	65.41	61.51	63.33
Transportation logistics	78.81	68.65	73.20
Information technology	56.64	58.46	57.53
Chemicals	70.14	63.36	66.43
Construction	74.15	67.13	70.22
Real estate	80.49	68.87	73.99
Wholesale and retail	71.61	64.16	67.56
Power and energy	84.56	70.84	76.94
Communication device	58.99	58.45	58.72
General manufacturing	71.80	65.09	68.17
Mining	79.99	68.73	73.69

After modeling by industry, the precision of most industries is higher than overall modeling, and the improvement is not stable. Considering the characteristics of each industry are different, the I Ching model can build hexagrams based on the characteristics of each industry and achieve better prediction performance and flexibility. However, two industries perform worse than overall modeling: Information technology and general manufacturing. The reason might be the parameter *win_len* is not the optimal value. Considering the rapid development in these two industries, the development of the industry can have dramatic change in just a few months. Also, as an industry that is now receiving much attention, it will be affected by more uncertain factors, such as national policies and public orientation.

The advantages of the method the authors proposed in this paper compared with traditional data mining algorithms mainly include two aspects:

1. The I Ching model avoids the risk of overfitting to a large extent, because the model constructed in this paper leverages feature dimensionality reduction, linear and nonlinear fitting to accomplish the feature of activation function. The model itself is not complicated, and the experimental results can show that the performance of the I Ching model on the test set is better than that of the traditional data mining model in the comparative verification.
2. Traditional data mining models require large-scale datasets for training and have high requirements for the quality of datasets. The authors' method can quickly select key features by taking advantage of the I Ching knowledge. The I Ching model only needs to focus on the Three Vitals feature selection, the Four Signs values, and changed Yao, which requires fewer parameters, can quickly converge the model through fewer datasets, and can obtain better prediction results. Besides, the I Ching model has strong versatility. When it is applied to other fields, it only needs to focus on constructing the mapping relationship between features and the Three Vitals beings under the background knowledge, and there is no need to make adjustments to the structure of the model.

ACKNOWLEDGMENT

This research was completed under the guidance of professor Guozi Sun and HuaKang Li. Their professional knowledge, scientific attitude, and rigorous academic spirit have all had a significant influence on this work.

CONCLUSION

In this paper, the authors investigated the prediction task and proposed a prediction model based on the I Ching mechanism, specifically, the fine-tuned I Ching prediction model, the RF, and the sliding windows mechanism to rank all the features in importance. The authors' I Ching model incorporates the classic I Ching hexagram generation process and decodes the hexagrams according to the original text of ZhouYi, which deeply integrates the I Ching knowledge. Specifically, the authors used the Three Vitals for feature dimensionality reduction. The mapping process of the Four Signs is similar to a function which knows the independent variable to get the dependent variable, because dynamic features are the key features of the I Ching and the result of the Four Signs mapping is relative, rather than absolute. Therefore, the authors used the dynamic sliding window method. According to the concept of the original hexagram, the original hexagram is formed from the Four Signs values, which completes the work of the generator. In addition, the I Ching have hexagrams changing, which is formed on the basis of original hexagram according to the rules of Yao change. The process of changing the hexagram is basically a linear transformation on the learned features. The decoder completes the decoding and obtains the prediction result after the model obtains the original hexagrams and the changed hexagrams, and the number and position of Yao changes.

To verify the real-world application ability, the authors conducted the experiment in the stock market scenario. Th experiments showed that the I Ching prediction model has a good performance on the task of predicting stock trends. However, on the other hand, the I Ching prediction model does not dynamically adjust the sliding window due to the development speed and growth cycle of different industries.

The trend prediction is only one aspect of mining the value of big data in the stock market. In future work, the research on problem diagnosis and early warning can also be conducted based on the prediction method the authors illustrated in this paper. When it is predicted that the future development of stocks is unfavorable, the model determines the cause of the unfavorable situation, and guides the company to adjust its business strategy according to the reason, so as to achieve the role of diagnosis and early warning.

In addition, the authors did not use the news text data in the experiment; the news data can reflect the development trend of the industry, which is beneficial to predicting the future development of the company. In the follow-up work, the text data should also be added for training.

CONFLICT OF INTEREST

The authors declare that they do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

FUNDING AGENCY

This research was supported by the National Natural Science Foundation of China [grant numbers 61906099] and by the Open Fund of Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Natural Resources [grant numbers KF-2019- 04-065].

REFERENCES

Audemard, G., Bellart, S., Bounia, L., Koriche, F., Lagniez, J. M., & Marquis, P. (2022). Trading complexity for sparsity in random forest explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*. doi:10.1609/aaai.v36i5.20484

Bartley, C., Liu, W., & Reynolds, M. (2019). Enhanced random forest algorithms for partially monotone ordinal classification. *Proceedings of the AAAI Conference on Artificial Intelligence*. doi:10.1609/aaai.v33i01.33013224

Bartol, K., Bojanić, D., Petković, T., Peharec, S., & Pribanić, T. (2022). Linear regression vs. deep learning: A simple yet effective baseline for human body measurement. *Sensors (Basel)*, *22*(5), 1885. doi:10.3390/s22051885 PMID:35271032

Breiman. (2001). Random forests. *Mach Learn*.

Brunatom, B. (2016). X-MIFS: Exact mutual information for feature selection. In *Proceedings of 2016 International Joint Conference on Neural NetWorks (IJCNN)*. IEEE. doi:10.1109/IJCNN.2016.7727644

Cao, T. T., Zhang, M. J., & Andreae, P. (2016). A wrapper feature selection approach to classification with missing data. *Proceedings of the 19th European Conference on the Applications of Evolutionary Computation.*

Chandrashekar, Girish, & Ferat. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*.

Chen, Y. W. (2021). *How do enterprises face difficulties? From I Ching enlightenment of the "Four Difficult Trigrams."* Tsinghua Management Review.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. doi:10.1007/BF00994018

Fang, X. H. (2021). Priori I Ching introduction: A I Ching phenomenological investigation. *I Ching Research*.

Gorade, S. M., Deo, A., & Purohit, P. (2017). A study of some data mining classification techniques. *International Research Journal of Engineering and Technology*, *4*(4), 3112–3115.

Guo, Y., & Lu, X. (2020). *Analysis the application of I Ching numerical hexagram model in stock market prediction*. Modern Marketing.

Hu, Y. (2013). *A prediction model for stock market: A comparison of the world's top investors with data mining method*. WHICEB.

Huang, X., & Liu, W. P. (2021). Research on near infrared feature selecting based on variable importance and partial least squares. *Journal of Hunan City University (Natural Science)*.

Ismail, F. (2019). Deep learning for time series classification: A review. *Data Mining and Knowledge Discovery*, *33*(4), 917–963. doi:10.1007/s10618-019-00619-1

Kannan, K. S., Sekar, P. S., & Mohamed, S. M. (2010). *Financial stock market forecast using data mining techniques*. Lecture Notes in Engineering & Computer Science.

Kesavaraj, G., & Sreekumar, S. (2013). A study on classification techniques in data mining. In *Proceedings of the 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*. IEEE. doi:10.1109/ICCCNT.2013.6726842

Kim, S. (2021). Pipeline construction cost forecasting using multivariate time series methods. *Journal of Pipeline Systems Engineering and Practice*.

Kim. (2011). Customer need type classification model using data mining techniques for recommender systems. *International Journal of Economics and Management Engineering*.

Kui, Y., Xianjie, G., Lin, L., Jiuyong, L., Hao, W., Zhaolong, L., & Xindong, W. (2020). Causality-based feature selection: Methods and evaluations. *ACM Computing Surveys*, *53*(5).

Li, C. Y. (2014). "Unity of man and nature" or "Ten, Chi, Jin" Three Vitals: The basic framework of Confucian environmental philosophy. *I Ching Research*.

Liu, C., Hoi, S. C., Zhao, P., & Sun, J. (2016). Online ARIMA algorithms for time series prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*. doi:10.1609/aaai.v30i1.10257

Liu, Y. X., Chen, B., & Zhou, Z. Y. (2019). *A improved feature selecting algorithm based on random forest*. Modern Electronic Technology.

Makiya, N., Alex, S., Youngsoo, K., Jonghyun, K., & Jinoh, K. (2021). Automated feature selection for anomaly detection in network traffic data. *ACM Transactions on Management Information Systems*, *12*(3).

Nascimento Da Silva, P., Plastino, A., Fabris, F., & Freitas, A. A. (2021). A novel feature selection method for uncertain features: An application to the prediction of pro-/anti-longevity genes. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *18*(6), 2230–2238. doi:10.1109/TCBB.2020.2988450 PMID:32324561

Neelamegam, S., & Ramaraj, E. (2013). Classification algorithm in data mining: An overview. *Network Trends and Technology: International Journal*.

Rana, A. (2021). A review of data mining techniques for the analyzation of big data. *Asian Journal of Multidimensional Research*.

Ratanamahatana, C. A., & Eamonn, K. (2005). Three myths about dynamic time warping data mining. In *Proceedings of the 2005 SIAM International Conference on Data Mining*. Society for Industrial and Applied Mathematics. doi:10.1137/1.9781611972757.50

Ruiz, A., Flynn, M., Large, J., Middlehurst, M., & Bagnall, A. (2021). The great multivariate time series classification bake off: A review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, *35*(2), 401–449. doi:10.1007/s10618-020-00727-3 PMID:33679210

Ruiz, A., Pasos, M., Flynn, & Bagnall, A. (2020). *Benchmarking multivariate time series classification algorithms*. Academic Press.

Shi, Q., Yin, J., Cai, J., Cichocki, A., Yokota, T., Chen, L., Yuan, M., & Zeng, J. (2020). Block hankel tensor ARIMA for multiple short time series forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*. doi:10.1609/aaai.v34i04.6032

Silvestrini, A., & Veredas, D. (2008). Temporal aggregation of univariate and multivariate time series models: A survey. *Journal of Economic Surveys*, *22*(3), 458–497. doi:10.1111/j.1467-6419.2007.00538.x

Tang, J., Salem, A., & Huan, L. (2014). Feature selection for classification: A review. *Data classification: Algorithms and applications*.

Tianqi, C., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Clinical Orthopaedics and Related Research*.

W, X. N. (2015). Composition of I Ching Gua&Yao Words: The triad of literature, history, and philosophy. *Jiangxi Social Sciences*.

Wang, W., & Carreira-Perpinan, M. (2014). The role of dimensionality reduction in classification. *Proceedings of the AAAI Conference on Artificial Intelligence*. doi:10.1609/aaai.v28i1.8975

Xu, D., Cheng, W., Zong, B., Song, D., Ni, J., Yu, W., Liu, Y., Chen, H., & Zhang, X. (2020). Tensorized LSTM with adaptive shared memory for learning trends in multivariate time series. *Proceedings of the AAAI Conference on Artificial Intelligence*. doi:10.1609/aaai.v34i02.5496

Y, H., & H, Y. D. (2011). The discussion of "changing" concept in I Ching. *Chuanshan Academic Journal*.

Yue, Z., Wang, Y., Duan, J., Yang, T., Huang, C., Tong, Y., & Xu, B. (2022). TS2Vec: Towards universal representation of time series. *Proceedings of the AAAI Conference on Artificial Intelligence.* doi:10.1609/aaai.v36i8.20881

Zhan, Y., & Yu, X. H. (2019). Research on macro influencing factors based on stock market stability. *Proceedings of the 2019 Annual Meeting on Management Engineering*. doi:10.1145/3377672.3378037

Zhao, P., & Lai, L. (2021). Efficient classification with adaptive KNN. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Zheng, H. Y., Zhou, Z. Q., & Chen, J. Y. (2021). RLSTM: A new framework of stock prediction by using random noise for overfitting prevention. *Computational Intelligence and Neuroscience*, *2021*(3), 1–14. doi:10.1155/2021/8865816 PMID:34113377

Zhou, Z., & Hooker, G. (2021). Unbiased measurement of feature importance in tree-based methods. *ACM Transactions on Knowledge Discovery from Data*, *15*(2), 1–21. doi:10.1145/3429445

Wenjie Liu is a master's student of Nanjing University of Posts and Telecommunications. Wenjie's research interest is in data mining and natural language processing.

Chen Sai is a master's student of Nanjing University of Posts and Telecommunications. Chen's research interest is in data mining and natural language processing.

Guoyao Huang is a student of Nanjing University of Posts and Telecommunications.

Lingfeng Lu is a master's student of Nanjing University of Posts and Telecommunications. Lingfeng's research interest is in natural language processing.

Huakang Li received his PhD degree in Computer Science from School of Computer Science and Engineering, University of Aizu in 2011. Currently. Huakang Li works as an associate professor at the School of Artificial Intelligence and Advanced Computing, XJTLU Entrepreneur College (Taicang). From May 2011 to September 2013, Huakang Li worked as a postdoctoral research fellow at the Department of Computer Science and Engineering, Shanghai Jiaotong University. During that period, Huakang Li went to Ali Cloud Computing Ltd. as a visiting researcher to develop the National 863 Project. From September 2013 to July 2020, Huakang Li was a lecturer at School of Computer Science, Nanjing University of Posts and Telecommunications. His research is focused on artificial intelligence and big data mining, specifically natural language processing, social computing, knowledge engineering, knowledge graphs, and the related domain applications.

Guozi Sun is a professor in the School of Computer Science and Technology at Nanjing University of Posts and Telecommunications, China. His research interests include blockchain forensics, digital forensics, and digital investigation. Sun received his Ph.D. in mechanical engineering and automation from Nanjing University of Aeronautics and Astronautics, China. He is a member of the IEEE Computer Society, ACM, China Computer Federation (CCF), Chinese Institute of Electronics (CIE), and Information Security and Forensics Society (ISFS), China.