Prediction of stock price by Hidden Markov model

Xin An[1, *]

[1]Department of mathematics, College London University, London, England

*Corresponding author: zcahxan@ucl.ac.uk

Abstract. Hidden Markov model is widely used in different fields, such as speech decoding, weather prediction, biometric, medical diagnosis and prediction. This paper explores the application of hidden Markov model in quantitative investment in the financial field, mainly in the field of securities trading. The main idea of this paper is to crack the different market states reflected in the price changes of securities. Price changes are affected by different market states at the same time, but this state is hidden and needs to be inferred. If people can calculate the relevant parameters of the state and its transition matrix, people can predict the stock price trend and establish the own trading strategy at the same time. The hidden Markov model is used as a theoretical basis in this paper, through the analysis of the fit between the stock market and the model, as well as various statistical knowledge as a tool. Through EM algorithm and machine learning as the programming basis, people finally selected Pingan and Moutai as the research objects, predicted the stock price changes through training. Therefore, when the training set is sufficiently large, hidden Markov model will have a bright performance in stock price prediction.

Keywords: Hidden Markov model; Stock price forecast; EM algorithm; machine learning.

1. Introduction

Speaking of the probability model used for statistical analysis in mathematics, there must be hidden Markov model [1]. It was proposed as early as the 1970s. Since then, it has been continuously improved and developed. In the 1980s, due to the convenience of information dissemination [2], it has been further spread, and has been involved and applied in many fields. Due to the late start of world finance, its application is also relatively late. By the early 1990s, with the rise of the financial sector, the application of this model in the financial field has gradually become common. Nowadays, the application of this model has been very extensive [3]. For example, in the daily life, computers, mobile phones, home appliances, speech recognition, fingerprint recognition, character recognition, etc., hidden Markov models are often used [4,5]. Hidden Markov model is derived from Markov process. Hidden Markov model mainly describes two closely related stochastic processes [6]. Its characteristic is the process with double randomness, and these two processes mainly include the hidden state random process, and its other corresponding is called the observed value random process.

Under specific circumstances, each invisible state will correspond to several or more variables, including observable variables, and the required observations will be distributed according to specific circumstances [7]. Usually, it is a random sequence that cannot see the state, because it is usually hidden. People can observe the random sequence value of the trial observation value, so people can use the hidden Markov model, and then calculate by using the observation value sequence as the input bias of the model. Therefore, people often can't see the hidden state in the actual process, and only infer the approximate sequence of the hidden state through the sequence of observations. Generally, the operation of stocks can reflect the economic state of a country, and it can be seen that they also play an essential role in contributing to the national economic construction. Therefore, it is necessary to analyze the general trend or specific trend of stock prices in the financial market. Investors and researchers will regard it as an important research goal and direction. At the beginning, people's analysis of stock investment is viewed and analyzed from a superficial and macro perspective. In most cases, they analyzed the general situation of the industry and company, then make a rough estimation of the stock, and then select the stocks that are considered to be very suitable for investment. In other words, through people's daily observation of the stock price, the relative standard of purchasing stocks is determined by focusing on the trading volume and the rise range.

In general, at the beginning, people only paid attention to the basic analysis and investment. Nowadays, with the rapid development of Finance and the intersection and integration of Finance and other disciplines such as statistics, computers and so on, finance has gradually shown the general trend of quantification, accuracy. In the process of cross integration with other fields, quantitative investment has gradually entered people's vision. The investment of stocks has also become predictable to a certain extent, which is based on the comprehensive application of mathematical and computer models [8]. The hidden Markov model people mentioned above also has great application and development in other fields. In statistical analysis, it is displayed as a probability graph, which is not limited to the great success in speech recognition and sound processing, but the application value of hidden Markov model does not stop here [9]. Many experts and scholars believe that hidden Markov models have a good development space in the financial direction, and can bring great help to the financial field and promote its development to a great extent. The central idea of hidden Markov model is to determine the required price by studying the different states of financial asset prices [10]. This state is generally hidden. Only by decoding and mining through special methods, when people decode all the internal parameters related to the state through a certain method, people can accurately analyze the changes of its parameters [11]. Everyone can also establish their own appropriate trading strategy through this method. Here people describe an application of hidden Markov model in the field of financial investment. The visibility of this method is described, which has great prospects for development. Therefore, the problems studied in this paper are of great practical significance.

2. Methodology

Definition 1: a double discrete stochastic process$\{C_t\}$，$\{X_t\}$ is called hidden Markov chain if it meets the following two conditions:

$$P(C_t = c_t|C_1^{t-1} = c_1^{t-1}) = P(C_t = c_t|C_{t-1} = c_{t-1}) \quad (1)$$

$$P(X_t = x_t|C_1^{t-1} = c_1^{t-1}, X_1^{t-1} = x_1^{t-1}) = P(X_t = x_t|C_t = c_t) \quad (2)$$

It should be pointed out here thatC_1^trepresentative$(C_1, C_2, \cdots, C_t)$，Namely：C_1^t=$(C_1, C_2, \cdots, C_t)$；It can be seen from this that$(C_1^t = c_1^t) = (C_1 = c_1, C_2 = c_2, \cdots\cdots, C_t = c_t)$，Later in this paper, the meaning of the simultaneous occurrence of superscripts and subscripts of letters is similar. A list of state sets representing states from subscripts to superscripts uses simple symbols, and the definition of hidden Markov chain can also be abbreviated as

$$\begin{cases} p(c_t|c_1^{t-1}) = p(c_t|c_{t-1}) \\ p(x_t|c_1^t, x_1^{t-1}) = p(x_t|c_t) \end{cases} \quad (3)$$

In the definition of hidden Markov chain，$\{C_t\}$is called a state process, which is unobservable. Equation (1) shows that it is a Markov chain;$\{X_t\}$ It is called the observed value process, Equation (2) shows that it is only related to the current state, and has nothing to do with the previous state and observations. On this basis, people can further define the enhanced version of hidden Markov chain.

Definition 2: the enhanced hidden Markov chain meets the following two conditions:

$$p(c_t|x_1^{t-1}, c_1^{t-1}) = p(c_t|c_{t-1}) \quad (4)$$

$$p(x_t|x_1, \cdots, x_{t-1}, x_{t+1}, \cdots, x_T, c_1, \cdots, c_t) = p(x_t = c_t) \quad (5)$$

The enhanced version of hidden Markov chain (1) shows that given T-1 state variables, the T state variable is not related to all other variables; Equation (2) shows that given the T state variable, the T observation is irrelevant to all other variables. In general, there are two random processes in the framework of hidden Markov chain,$\{C_t\}$and$\{X_t\}$；$\{C_t\}$ is Markov chain，given C_{t-1}，C_t Independent of all other variables，$\{X_t\}$ Is the observed value process，given C_t，X_t It is independent of all other variables. In essence, expectation maximization algorithm can be seen as an extension of maximum likelihood estimation method, which is mainly used to solve the problem of parameter estimation with unobservable variables. Suppose there are two sets of variables in the statistical model: $\{x_1, x_2, \cdots, x_T\}$ and $\{y_1, y_2, \cdots, y_T\}$ among $\{x_1, x_2, \cdots, x_T\}$ is an observable variable,$\{y_1, y_2, \cdots, y_T\}$ Is an unobservable variable, EM algorithm is mainly divided into two steps:

E-step: The expectation step is used to find the expression of expectation to be maximized in the current iteration round.

M-step: In the process of maximizing expectations, the iterative formulas of all parameters that maximize the expectation in this round are obtained by making its partial derivative 0.

The execution process of EM algorithm is to start from the initial round, and execute E and M steps in sequence in each round until the parameters converge, so $\{x_1, x_2, \cdots, x_T; y_1, y_2, \cdots, y_T\}$ called complete data, its likelihood function is

$$L(\theta|x_1, x_2, \cdots, x_T; y_1, y_2, \cdots, y_T) = p(x_1, x_2, \cdots, x_T; y_1, y_2, \cdots, y_T|\theta)$$
$$= p(x_1^T|\theta) \cdot p(y_1^T|x_1^T, \theta) \quad (6)$$

When there are unobservable variables, the likelihood function is random, because$y_1, y_2, \cdots, y_T$ Is random data, Therefore, the likelihood function is

$$L(\theta|x_1^T, y_1^T) = p(x_1^T|\theta) \cdot p(y_1^T|x_1^T, \theta) \quad (7)$$

Given time t, consider $x_1, x_2, \cdots, x_T$ as function of $y_1, y_2, \cdots, y_T$.

Accordingly,

$$L(\theta|x_1, x_2, \cdots, x_T) = p(x_1^T|\theta) \quad (8)$$

It is called the likelihood function of incomplete data. According to the likelihood function of the perfect data, the log likelihood function of the complete data is defined as

$$logL(\theta|x_1^T, y_1^T) = logp(x_1^T, y_1^T|\theta) = log[p(x_1^T|\theta) \cdot p(y_1^T|x_1^T, \theta)] \quad (9)$$

Obviously, people can also regard CDLL as a random variable, because it is a random variable function of $y_1, y_2, \cdots, y_T$.

The EM algorithm, as its name suggests, there are two steps: expectation (E-step) and maximization (M-step). First, the core idea of E-step is to find the conditional expectation of the log likelihood function CDLL of complete data. Then， define the conditional expectation of CDLL as Q function, that is:

$$Q(\theta, \tilde{\theta}) = E\{logp(x_1^T, y_1^T|\theta)|x_1^T, \tilde{\theta}\} \quad (10)$$

3. Results and Discussion

After According to the rational hypothesis, everyone in the market will make a rational judgment on the future income of the company based on historical information, market public information, grapevine news, expert opinions, etc., and discount the estimated future income based on this judgment, which involves the judgment of the company's future growth rate, duration, discount rate, etc. from the above, he can calculate the internal value of the company. Like all commodity transactions of houses and televisions, the actual buying and selling behavior generated by the expectation of valuation judgment forms the supply and demand in the market. Countless people buy and sell, and when a price reaches equilibrium, the current market price (share price) is formed. This market price will in turn affect people's valuation and supply and demand.

Value investors believe that the real stock price may deviate from this intrinsic value. If you judge that the intrinsic value of a company exceeds the current stock price, then buy it. Do not buy or sell it if it is lower than the current share price. Value investors believe that stock prices will return to intrinsic value sooner or later. In other words, in a sense, if the intrinsic value of an enterprise can be obtained from various market information, its stock price can be predicted.

In addition, whether the stock price can be predicted is also related to whether human behavior is completely rational. If people are completely rational, then the stock price cannot be predicted, while some behavioral finance experts say that if people are not completely rational, the stock price can be predicted instead. Its argument is based on these psychologists and the subsequent emergence of Behavioral Finance on human nature. Therefore, to some extent, when most people who participate in the stock market are not completely rational, the stock price can be predicted.

Data prepossessing is required before programming. First, a time series is a certain amount measured in a time series within a certain time interval. Time series analysis is mainly based on the

trend and data characteristics of historical data. Time series can be divided into stationary series and non-stationary series. Stationary sequences are characterized by sequences that have no trend but are random, rather than sequences that contain trend, seasonality, and randomness, that is, stationary sequences. In its broadest form, time series analysis is about inferring what happened to a range of data points in the past and trying to predict what will happen in the future. Time series analysis attempts to understand the past and predict the future. Generally, time series usually include the following types:

Trend - a trend is a consistent directional movement in a time series. These trends will be random or deterministic. The long-term rise or fall of the time series over a long period of time.

Seasonal changes – amount of time series contain seasonal changes. This is especially true in series that show commercial sales or climate levels. People often see seasonal changes in commodities, especially those related to annual temperature changes or growing seasons.

Sequence dependence - one of the most significant features of time series, especially financial series, is sequence correlation. This happens. When time series observations that are close to each other tend to correlate.

For the change of a specific stock price in the financial market people studied, there is no doubt that it is a time series in the strict sense. For the traditional time series prediction, it can do a better prediction for the stable time series. However, people found that the two stocks are unstable time series through ADF test, which brings difficulties to the traditional time series prediction. Therefore, people choose the hidden Markov model with stronger prediction accuracy for this unstable time series. The implementation of this model relies on the machine learning algorithm mentioned above.

HMM training is as follows:

Given the training sets X and y, estimate what kind of hidden Markov model this is.

The following assumes that the data is complete:

$$\hat{a}_{ij} = \frac{A_{i,j}}{\sum_{j=1}^{N} A_{i,j}} \tag{11}$$

Normalize each row, a above $A_{i,j}$ represents the number of times that state I is transferred to state j in the sample.

For the initial state, people can directly look at the statistics of the starting state in the sample.

$$\hat{\pi}_i = \frac{c_i}{\sum_{i=1}^{N} c_i} \tag{12}$$

Similarly, the launch probability can also be estimated.

$$\hat{b}_{i,j} = \frac{B_{i,j}}{\sum_{i=1}^{N} B_{i,j}} \tag{13}$$

HMM prediction is as follows:

Sequence prediction problem:

Given the triplet, determine the hidden Markov model of π, A, B, and given the sequence x, find the most likely state sequence y.

Approximation Algorithm:

The core idea of the algorithm is to choose the most likely state at each time t is i_t^*, So people can get a sequence of states $I^* = (i_1^*, i_2^*, \cdots, i_T^*)$, Take it as the prediction result.

Given the specific parameters and observation sequence of Markov model. State q_i at time t the probability of I is $\gamma_t(i)$:

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{P(O|\lambda)} = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N} \alpha_t(j)\beta_t(j)} \tag{14}$$

At every moment, t is the most likely state i_t^* is:

$$i_t^* = \arg\max_{1 \leqslant i \leqslant N} [\gamma_t(i)], \quad t = 1,2,\cdots,T \tag{15}$$

So as to get the final state order

$$I^* = (i_1^*, i_2^*, \cdots, i_T^*) \tag{16}$$

In this paper, 571 sets of data from China Ping An and Guizhou Moutai stocks from December 31, 2019 to May 13, 2022 are selected. The first 541 sets are the training set and the last 30 sets are the test set. The specific results of learning and prediction are as follows:

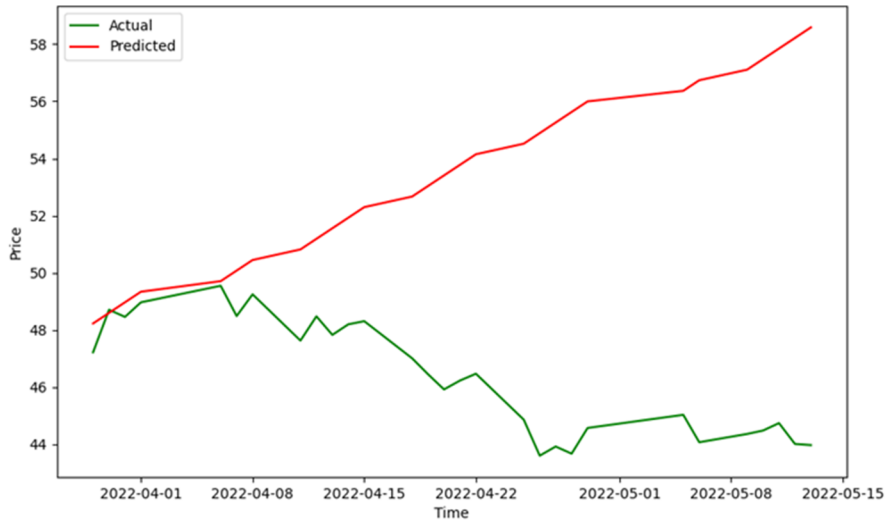

Fig. 1 Comparison Chart of Pingan forecast in China

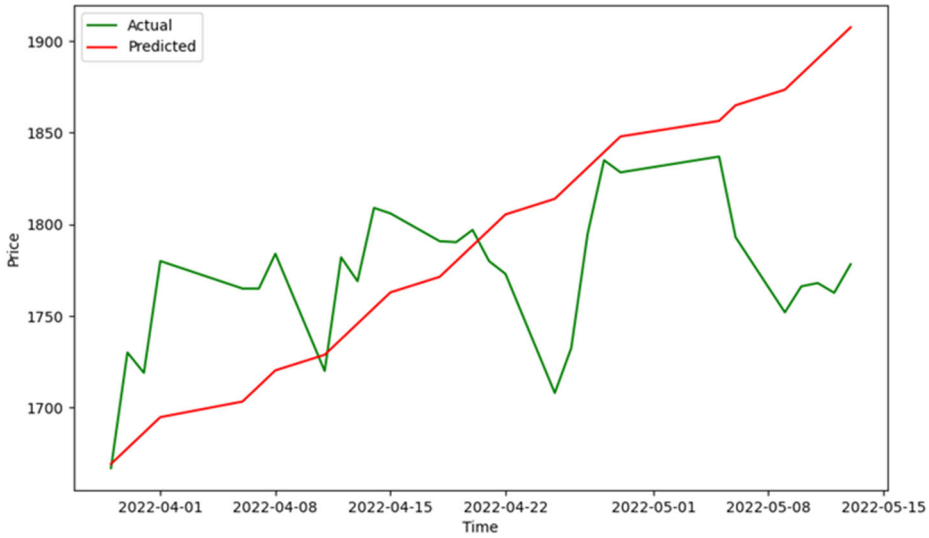

Fig. 2 Comparison Chart of the prediction of Moutai, Guizhou

From the results, the effect of hidden Markov model in predicting the change of Guizhou Moutai stock price is not very accurate, but the basic trend and the emergence of key points are relatively accurate. But in contrast, the prediction of PingAn in China is the same at the key points, but the trend is the opposite. This fully shows the instability of hidden Markov model in machine learning effect. The above Ping An prediction problem in China may be caused by the lack of training sets. People will expand the training set for one year and observe the effect again.

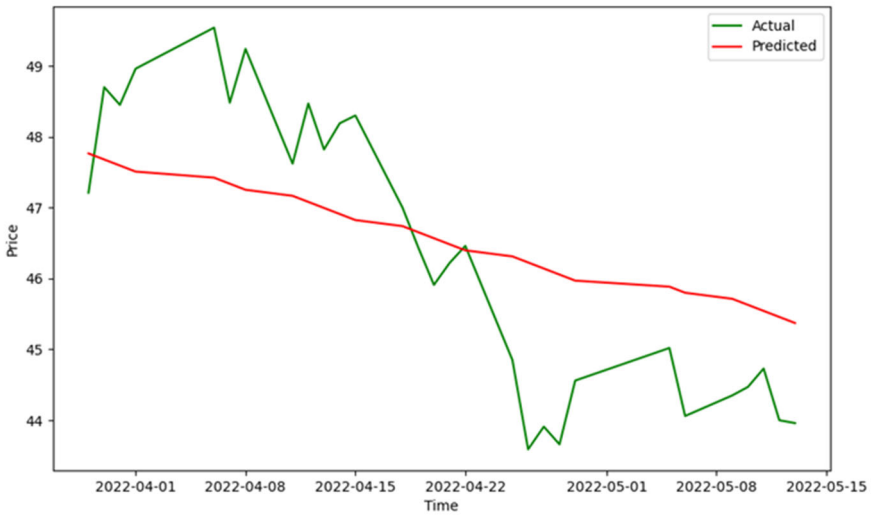

Fig. 3 Prediction of Pingan in China after adjusting the training set

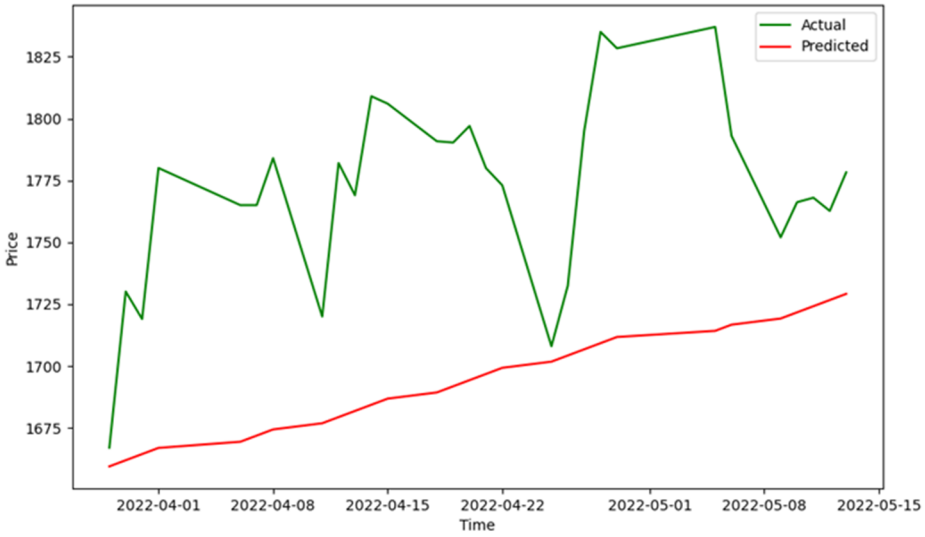

Fig. 4 Prediction of Moutai in Guizhou after adjusting the training set

By expanding the training set to December 31, 2018, the learning effect is significantly stronger, which also verifies that the previous inaccurate prediction may be the reason why there are too few training sets.

In the field of machine learning, a fundamental theorem exists is that there is “no free lunch”. In other words, no algorithm can solve all of the things very perfectly, especially in the field of supervised learning. For example, you can't say that neural networks are better than decision trees in any case, and vice versa. They are influenced by many factors, such as the structure or size of the dataset. Therefore, when using a given test set to evaluate performance and select algorithms, different algorithms should be selected based on the specific problem. Of course, for specific problems, the selected algorithm must be applicable to your own problems, which requires selecting the correct machine learning task. For example, when you need to clean the house, you can use a vacuum cleaner, broom, or mop, but you should never take out a shovel and dig. First, the connection between the training set and the test set is known. In short, one needs to learn a model on a training set and then get a test set to use. The effect should be measured according to the error rate of the test set. But many times, people can only assume that the test set and the training set conform to the same data distribution, but people can't get the real test data. At this time, how can people measure the test

error rate when people only see the training error rate? Because the training samples are not enough (at least not enough), the accuracy of the model obtained from the training set is actually not high. (Even if the training set is 100% accurate, this does not express the fact that it shows the distribution of data. The aim is to show the real data distribution, not just the limited data points of the training set). Besides, in fact, training samples often exist some noise errors. Therefore, the model obtains erroneous data distribution estimates by taking the errors in the training set as the true data distribution features. If people excessively pursue the perfection of the training set, they will support a very complicated model. In this way, the real test set will be in a state of confusion (a phenomenon known as overfitting). However, one should not use overly simple models, otherwise the model will not be enough to represent the data distribution when the data distribution is complicated (reflected in the high error rate even on the training set, which is called lower fitting). Overfitting means that the model used is more tough than the factual data distribution, while underfitting means that the model used is clearer than the factual data distribution.

In the framework of statistical learning, when people describe the complexity of the model, people have this view that $Error = Bias + Variance$. The error here can be roughly understood as the prediction error rate of the model, which is composed of two parts. One part is the estimation inaccuracy (Bias) caused by the simplicity of the model, and the other part is the larger change space and uncertainty (Variance) caused by the complexity of the model. Therefore, it is easy to analyze naive Bayes. It simply assumes that each data is irrelevant, which is a seriously simplified model. Therefore, for such a simple model, the bias part will be larger than the variance part in most cases, that is, high deviation and low variance. In practice, in order to minimize the error, people need to balance the proportion of bias and variance when selecting models, that is, balance over fitting and under fitting. When the complexity of the model increases, the deviation will gradually decrease and the variance will gradually increase. Therefore, for different tasks, machine learning algorithms need to meet the requirements while giving consideration to efficiency and error probability. Only continuous improvement of machine learning methods can really solve the continuous complex problems.

4. Conclusion

This paper mainly introduces the practical application of hidden Markov model in specific stock forecasting. Firstly, the concept of hidden Markov model is introduced at the beginning of the full text. Then, based on the actual stock market, the stock data of PingAn and Moutai in Guizhou Province are selected to analyze the mathematical model. The biggest difficulty of all models is to predict the future. A good model should be practical, not divorced from reality, and try to match the future trend of the market. Through the estimation and decoding of the collected data, people get a more reasonable hidden Markov model. How to select data is very important.

However, in the overall results, there are still some problems, especially in the prediction of Ping An stock in China, the lack of the first training time led to a large deviation of the prediction model, but after increasing the training time, it has a good prediction performance, which shows that there is still a lot of room for progress in this model. The application of hidden Markov model in quantitative investment in this paper can be a good reference for future research, because the research of hidden Markov model in China is not many, but the application of hidden Markov model in foreign countries has been very extensive. Scholars can make more energy investment in the research of financial investment based on hidden Markov model. However, it is believed that under the influence of the global economy and the continuous development of international trade, the domestic application of hidden Markov model will be more and more widely. At the same time, in the research process of this paper, in order to facilitate the calculation of data and the establishment of the model, only two market states, bear market and bull market, are selected. But in the actual stock market, there will be many states. There will be a whole disk, but if people choose three states for decoding and analysis, although the data will be more complex and the calculation will be more difficult, it is possible. The

model will be more high-quality, and the prediction effect for the future will be better. Or if too much data will lead to over fitting and the decline of prediction effect, these are issues that need to be discussed by future scholars.

References

[1] Eddy S R. Profile hidden Markov models. Bioinformatics, 1998(9):755.

[2] B Schuster‐Böckler, Bateman A. An introduction to hidden Markov models. Curr Protoc Bioinformatics, 2007.

[3] Olivier Cappé, Moulines E, Tobias Rydén. Inference in Hidden Markov Models. Technometrics, 2006, 48(4):574-575.

[4] Mccallum A, Freitag D, Pereira F. Maximum Entropy Markov Models for Information Extraction and Segmentation. icml, 2001.

[5] O Cappé, Moulines E, Ryden T. Inference in Hidden Markov Models (Springer Series in Statistics). Springer-Verlag New York, Inc. 2005.

[6] Atln B A. Speech emotion recognition using hidden Markov models. Speech Communication, 2003, 41(4):603-623.

[7] Starner T E. Visual Recognition of American Sign Language Using Hidden Markov Model. 1995.

[8] Rabiner L R, Juang B H. Hidden Markov Models for Speech Recognition — Strengths and Limitations. Springer Berlin Heidelberg, 1992.

[9] Nielsen H, Krogh A. Prediction of signal peptides and signal anchors by a hidden Markov model. Intelligent Systems for Molecular Biology, 1998, 6:122-130.

[10] Karplus K, Barrett C, Hughey R. Hidden Markov models for detecting remote protein homologies. Bioinformatics, 1998(10):846-856.

[11] Jarrow R A, Lando D, Turnbull S M. A Markov Model for the Term Structure of Credit Risk Spreads. Review of Financial Studies, 2004, 10(2):481-523.