ISSN 2616-5775 Vol. 5, Issue 11: 10-17, DOI: 10.25236/AJCIS.2022.051102

Stock Price Prediction Based on Discrete Hidden Markov Chain

Wenxuan Qiu, Hongye Cai

College of Mathematics and Statistic, Shenzhen University, Shenzhen, 518060, China

Abstract: *Stock forecasting has always been the center of the financial market. This paper uses the Discrete Hidden Markov Model (HMM) to predict stock price. The initial data of stock price is denoised by wavelet, and the denoised data is processed as the input data of HMM, and the number of implicit states that make the model most robust is determined by using OEHS criteria. The model is trained by Baum- Welch-Algo algorithm to obtain the correlation matrix parameters, and then the implicit state with the maximum probability is obtained by the Viterbi algorithm. This paper also introduces the Voting Strategy to improve the probability of accurate prediction. Finally, according to the Observation emission matrix, the probability distribution of the return rate of the next day is obtained, and the stock price is obtained. The accuracy of the predicted value is tested by MAPE. The experimental results show that the Discrete Hidden Markov model can better predict the future stock price trend. The accuracy of the predicted value was tested by MAPE, and the value was 0.0284.*

Keywords: *Vanke A; HMM model; Wavelet denoising; Baumwelchalgo algorithm*

1. Introduction

As one of the important financial instruments in the financial market, stock forecasting has always been the focus of financial research. However, due to the large number of factors affecting stock price, which are difficult to quantify and uncontrollable, it is very difficult to accurately predict the fluctuation of stock price. Therefore, a large number of domestic and foreign scholars have conducted a lot of research. At first, most scholars chose to use Regression Analysis, Autoregressive Model, Autoregressive Moving Average Model, Auto Regressive Integrated Moving Average Model, VAR (Vector Autoregressive Model), and other methods to predict the stock price[1]. However, because there are many factors affecting the stock price, and the volatility is violent in unexpected situations, non-linear and many influencing factors are difficult to quantify, these methods can not accurately predict the stock price. In 1970, BOX[2] an American statistician,n and others proposed the ARIMA model, which greatly improved the accuracy of stock price prediction, but in essence, it can only capture linear information. Subsequently, Engle[3] put forward the autoregressive conditional heteroscedasticity model (ARCH model) to better describe the uncertainty of stock price. However, A series of models such as ARCH and GARCH is often difficult to fit in the highly nonlinear stock price prediction, resulting in high order and the result of overfitting. Most of the fitting methods are based on the stock price and financial-related information itself. We have reason to believe all the information, including the state of the financial market, the form of the national policy, the investor's investment mood, etc. But the future investor's mood is often not considered in the regression prediction. After a sample survey of 125 members of the American Association of individual investors, De Bondt[4] found that the mood of individual investors, that is optimistic, pessimistic, or neutral sentiment has a significant correlation with the overall performance of the Dow Jones industrial average. However, it is often difficult to quantify sentiment.

In recent years, many intelligent algorithms have been applied to stock price prediction. We do not need to fully quantify every data, so Machine Learning and other algorithms have pushed stock price prediction to a new height. In 2008, Huang Li[5] used BP neural network to predict the rise and fall of the stock market. At the same time, she studied the problems of complex network structure and low classification ability of BP neural network in the classification of complex samples, reduced the dimension of input samples with factor analysis (FA), meliorated the BP algorithm with genetic algorithm (GA), and proposed FAGABPNN algorithm. However, the BP neural algorithm needs many parameters and is prone to local extremum and overfitting[6]. Later, Wang Weihong et al[7] proposed the PCA-FOA-SVR algorithm, which uses PCA to eliminate the interference information of stock price characteristics, and uses the Drosophila algorithm to optimize SVR parameters to overcome the defect that neural

ISSN 2616-5775 Vol. 5, Issue 11: 10-17, DOI: 10.25236/AJCIS.2022.051102

network overfitting falls into a local minimum. However, it has not fully demonstrated how to find a suitable kernel function.

Based on the fact that the stock price can be observed, but the influencing factors are difficult to judge, this paper uses Hidden Markov Model (HMM) to predict the stock price. HMM is a statistical model, which was first proposed by Baum et al[8]. It is to predict the actual unobservable sequence through the observable sequence. When applied to stock prediction, the relevant probability is obtained through learning, and the most likely situation is obtained by using the probability. In this paper, I use HMM model to predict the stock price in the next month and get a good prediction result.

2. Introducing of HMM

2.1 Overview of Model

HMM is mainly composed of hidden state sequences $(S_1, S_2, \cdots, S_n)$ and observation sequences $(O_1, O_2, \cdots, O_n)$. The hidden state sequence is a state sequence containing unknown variables. The states are not visible but can be obtained through the observation sequence O. Each observation value is expressed as various states through a specific probability density distribution. Each observation vector is generated by a state sequence with a corresponding probability density distribution, as shown in Fig. 1. Therefore, Hidden Markov is a double stochastic process. At the same time, each hidden state S_{n-1} is only related to the previous state S_n and has nothing to do with other states. The observation sequence O is a directly observable value.

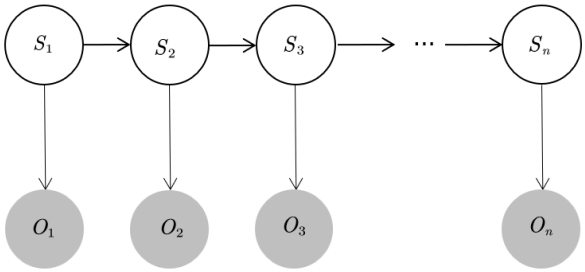

Figure 1: Hidden Markov process

2.2 The parameters of HMM

HMM can be described by five elements: $\lambda = (N, M, A, B, \pi)$

(1) Number of hidden states N: number of states not observed $N = \{q_1, q_2, \cdots, q_n\}$

(2) Number of observable states M: the number of states that can be observed $M = \{v_1, v_2, \cdots, v_m\}$

(3) Transition matrix(A) $A = \{a_{ij}\}, a_{ij} = P(q_t = S_j \mid q_{t-1} = S_i)$

(4) Observation emission matrix(B) $B = \{b_{jk}\}, b_{jk} = P(O_t = v_k | q_t = S_j)$

(5) Prior probability matrix (Π) $\Pi = \{\pi_i\}, \pi_i = P\{q_1 = S_i\}$

2.3 Three basic problems of HMM

Let $\lambda = (A, B, \pi)$ be a given HMM parameter and $O = (O_1, O_2, ..., O_n)$ be the observation sequence, then the three problems are described as:

(1) Evaluation： Given the observation sequence $O = (o_t)$ and the relevant parameter $\lambda = (A, B, \Pi)$ of the model, the probability $P(O|\lambda)$ is obtained. In the evaluation, we can get the

ISSN 2616-5775 Vol. 5, Issue 11: 10-17, DOI: 10.25236/AJCIS.2022.051102

matching degree between the observation sequence and the model. Through the evaluation, we can select the most matching model among the trained models. Generally, we use Forward-Backward Algorithm to solve this problem.

(2) Decoding: Given the observation sequence $O=(o_t)$ and the relevant parameters $\lambda=(A,B,\Pi)$ of the model, try to get the hidden relation between the model $\lambda=(A,B,\Pi)$ and the observation sequence $O=(o_t)$, and obtain the hidden state sequence $S=(q_1,q_2,\cdots q_n)$ with the largest probability. This problem can be solved by the Viterbi algorithm.

(3) Learning: The process of learning is the most core problem in HMM. through the given observation sequence $O=(o_t)$ and the initial slave model parameter $\lambda=(A,B,\Pi)$, the model parameters are adjusted and strengthened by learning the observation sequence. And then make $P(O|\lambda)$ reach the maximum under the parameter λ. this problem is generally solved by the Baum-Welch algorithm.

2.4 Algorithm flow

Step 1: Collect the original data. In this paper, we refer to the observable data such as the price and trading volume of the stock on that day.

Step 2: Determine the number of hidden states. By setting or applying relevant criteria, such as AIC, BIC, or OEHS, the models with a different number of states are measured, and finally, the most stable number of hidden states of the model is selected.

Step 3: Model parameter estimation. The model iteratively obtains the final model parameters by continuously learning the observation sequence.

Step 4: according to the trained HMM, the Viterbi algorithm is used to identify the pattern and find the most similar log-likelihood value.

Step 5: get the predicted value and predict the accuracy.

As shown in Figure 2:

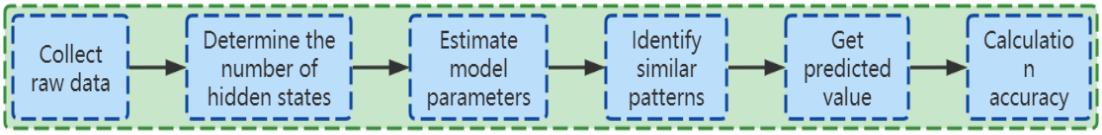

Figure 2: The flow of the Algorithm

3. Modeling process

3.1 Data selection and processing

To avoid the situation that the stock price is affected by the large-scale exchange due to the manipulation of the stock by the market makers, and violating the market state many times, we select the stock price of the larger enterprise - Vanke A (00000 2) for analysis. It can be seen from Figure 3 that the skewness of the closing price sequence of Vanke A within the selected time range is 0.166284, the skewness is greater than zero, and it is slightly right biased, the kurtosis is 3.339824, the value is greater than 3, and the tail of the sequence is thick. Therefore, the sequence distribution is right-biased and has a sharp peak and thick tail. By analyzing the other three groups of price series statistics, the distribution pattern of each series is similar to its corresponding closing price.

Hassan and Nath[8] select the data of multiple variables—opening price, closing price, highest price and lowest price rather than single variable data, which can improve the prediction accuracy through voting strategy.

In this paper, 1189 sets of data (excluding the closing date) of Vanke A: from January 3, 2017, to November 30, 2021, are selected as the training data to determine the HMM model parameters; 23 sets of data from December 1, 2021, to December 31, 202,1 are used as test data.

ISSN 2616-5775 Vol. 5, Issue 11: 10-17, DOI: 10.25236/AJCIS.2022.051102

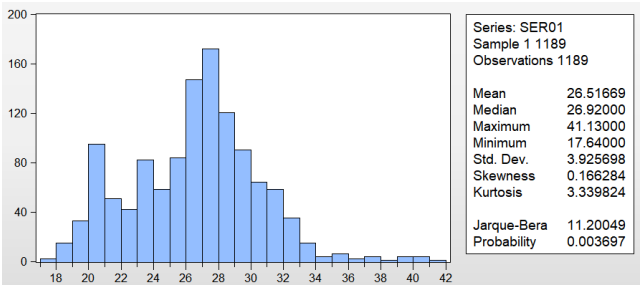

Figure 3: Distribution and characteristics of the closing price of Vanke A

3.2 Wavelet Domain Denoising

Stock data are susceptible to various random disturbances, including political changes and irrational investments. The noise composed of random disturbances can not help us better predict, and even affect the accuracy of our prediction. Therefore, in data processing, we should first remove the noise containing less financial information.

Most of the traditional methods are moving average, traditional filtering, and Kalman filtering[9]. The moving average is rough to remove noise. It places effective information and invalid information on the same status, and it is easy to dispose of effective information. At the same time, considering the high coincidence rate of the high-frequency signal and the low-frequency signal of the stock price, it is impossible to remove the noise by the traditional filtering method of separating the high and low frequencies. Most stock prices are complex and nonlinear changes, so it is impossible to remove noise by the Kalman filter method to determine variance and equation. For non-linear and non-stationary time series data, it is appropriate to use adaptive wavelet analysis[7]. Because the stock price is often unstable, high-frequency information accounts for a high proportion, the information contained in large fluctuations is highly effective, and small fluctuations are mostly composed of random disturbances, so we use wavelet denoising to remove the noise in small fluctuations and ensure the smoothness and similarity of signals. Figure 4 shows the denoising of the closing price of Vanke A shares. The red line is the original sequence and the blue line is the sequence after denoising.

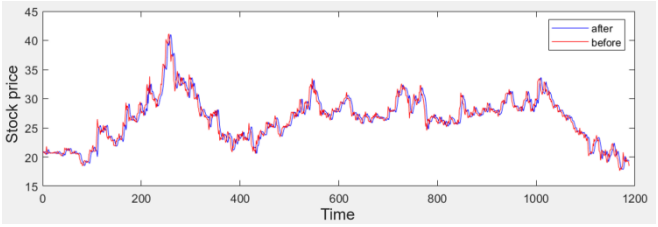

Figure 4: Share price of Vanke A after wavelet denoising

Figure 5: It can be seen that the low-frequency signal (a3) contains a high amount of information, so it is more reserved, while the high-frequency signal (d1, d2, d3) contains more disturbance items, so it is more discarded.

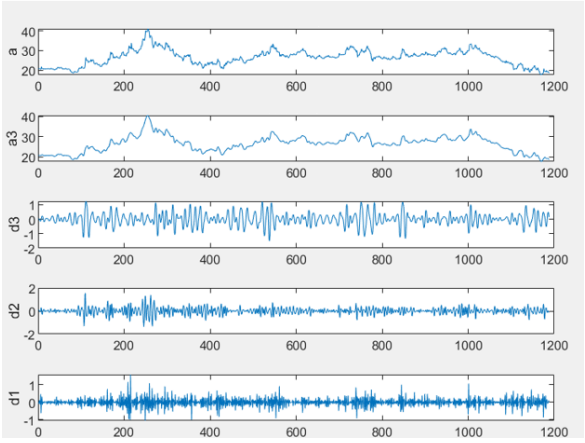

Figure 5: Raw/low frequency, high-frequency signal

ISSN 2616-5775 Vol. 5, Issue 11: 10-17, DOI: 10.25236/AJCIS.2022.051102

3.3 Establishment of HMM model

3.3.1 Classification of observation data

In the problem of predicting stock prices, although it is more valuable to predict an accurate value, it is often difficult to predict an accurate value due to the characteristics of the stock itself. At the same time, it is easy to fall into the problem of overfitting, and the prediction ability is reduced. Therefore, compared with the continuous HMM model, the discrete HMM model can predict the stock price state more stably[10]. Therefore, the observation data are divided into four states (soaring, slowly rising, slow decline, and plummeting) in this paper. The specific data are shown in the TABLE.1 below.

Table 1: Classification of observation data O_n

	Soaring	Slowly rising	Slow decline	Plummet
Range	6%~10%	0%~6%	-6%~0%	-10%~-6%

3.3.2 Determine the number of hidden states

Two, three, four, and five hidden states are selected respectively, and the model is tested by using the OEHS criteria to determine the number of hidden states of the model. The observation sequence is divided into two groups odd bits and even bits. 15 groups of initial distribution, transition probability matrix, and mixed normal distribution are randomly generated for odd and even sequences, and the parameters are trained by the EM algorithm. The results are shown in the following TABLE 2:

Table 2: Determine the number of hidden states

Number	Loglik 1	Loglik 2	Loglik 3	Rate of change
2	-3952.25	-1604.01	-1610.28	0.1080%
3	-3787.76	-1623.20	-1608.23	0.0512%
4	-3767.57	-1626.04	-1600.50	0.0142%
5	-3751.58	-1612.60	-1610.49	0.0562%

It can be seen that when the number of hidden states is 4, the change rate is the lowest, so it can be determined that the number of hidden states is 4. Through consulting the papers, relevant literature, and drawing the K-line chart, this paper confirms the hidden state as bull market, bear market, shock and rebound. The following Figure 6. is the hidden state representation taking the K-line chart of Vanke A in recent years as an example.

Figure 6: Example of hidden state

3.4 Model parameter estimation

Since the HMM model is dependent on the initial parameters, we randomly generate ten groups of data and one group of data generated by the law of large numbers, including Transition matrix(A)、Observation emission matrix(B)、Prior probability matrix(Π), The model parameters are trained by Baum-Welch algorithm. When the set threshold is reached, the iteration is stopped, and the parameters gradually tend to be stable. Thus, the effect of reducing the dependence on the initial value is achieved. The specific parameters including TABLE 3. TABLE 4. TABLE 5. are as follows:

ISSN 2616-5775 Vol. 5, Issue 11: 10-17, DOI: 10.25236/AJCIS.2022.051102

Table 3: Transition matrix (A)

	Bull	Bear	Shock	Rebound
Bull	0.4093	0.1592	0.4214	0.0101
Bear	0.0401	0.4011	0.411	0.1478
Shock	0.1449	0.1618	0.6218	0.0715
Rebound	0.4039	0.0816	0.415	0.0995

Table 4: Observation emission matrix (B)

	Soaring	Slowly rising	Slow decline	Plummet
Bull	0.0692	0.5825	0.3426	0.0057
Bear	0.0066	0.3049	0.6824	0.0061
Shock	0.0132	0.4534	0.5216	0.0118
Rebound	0.0431	0.4868	0.4651	0.005

Table 5: Prior probability matrix (Π)

	Bull	Bear	Shock	Rebound
P	0.0166	0.2786	0.6957	0.0091

3.5 Pattern recognition and prediction

Hassan finds the L_j closest to the likelihood value $L_n = \log P(O_n|\lambda)$ generated by the observation value O_i under $\lambda = (A, B, \Pi)$ the historical observation value $O = (O_1, O_2, ..., O_n)$, so as to find the two dates with the most similar observation value States, predict S_{i+1} according to S_{j+1}, and calculate O_{i+1} according to the observation emission matrix (B) [11]. Because Hassan holds the following view: on the two dates with similar likelihood values, the change of the observed values on the second day is similar, that is, the rise and fall are similar. However, with the continuous development of the stock market, it is impossible to complete the task of stock price prediction simply by copying the stock price path. So this paper adopts Viterbi algorithm: The hidden state sequence $S = (q_1, q_2, \cdots q_n)$ with the largest probability is obtained by observing the sequence $O = (O_1, O_2, ..., O_n)$. That is, after the implicit state S_n of the day is obtained, the transition matrix is used to obtain S_{n+1}, and then the observation emission matrix (B) is used to obtain O_{n+1}. In this paper, four observation values are used to obtain four prediction values, and the final prediction results are obtained through the voting strategy. The specific voting strategy is shown in the table below:

The hidden state sequence s with the largest probability is obtained by observing sequence a. That is, after the implicit state D of the day is obtained, the transition matrix is used to obtain F, and then the observation emission matrix is used to obtain F. In this paper, four observation values are used to obtain four prediction values, and the final prediction results are obtained through the voting strategy. The specific voting strategy is shown in TABLE 6. below:

Table 6: Voting strategy

	Opening prices	Closing prices	Highest prices	Bottom prices	O
Soaring	√	√	√		√
Slowly rising				√	
Slow decline					
Plummet					

In particular, if the effective value cannot be obtained by voting, the weighting formula is as follows:

ISSN 2616-5775 Vol. 5, Issue 11: 10-17, DOI: 10.25236/AJCIS.2022.051102

$$O_{n+1=}0.15 * O_{\substack{opening \\ prices}} + 0.5 * O_{\substack{closing \\ price}} + 0.15 * O_{\substack{Highest \\ price}} + 0.2 * O_{\substack{Bottom \\ price}} \quad (1)$$

In this paper, 23 groups of data from December 1, 2021, to December 31, 2021, are used as the test data. The specific prediction results are as follows. According to Figure 7., it can be seen that the accuracy rate of the prediction results reaches 82.61%:

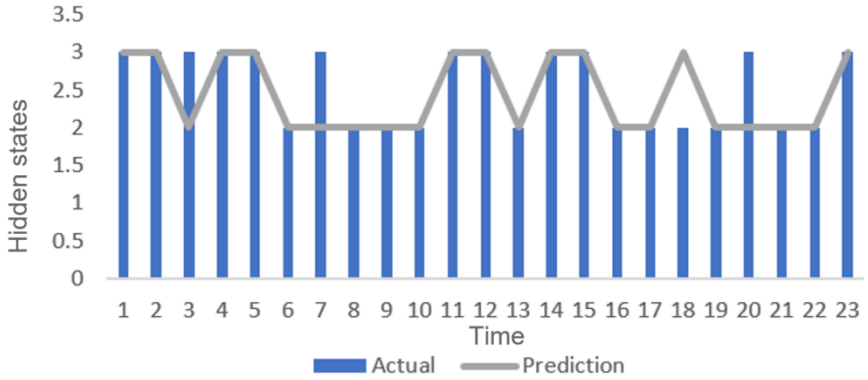

Figure 7: Comparison of actual and predicted results

3.6 Model prediction improvement

The four observation states can be extended to twenty observation States, that is, one point is taken as an interval for prediction. The specific prediction results are as follows Figure 8. :

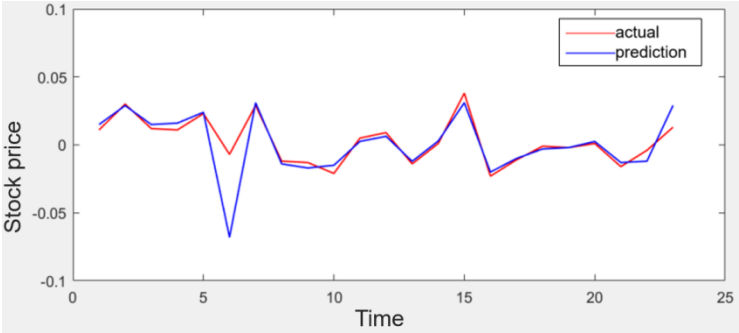

Figure 8: Comparison of actual and predicted results

3.7 Prediction accuracy test

Use Mean Absolute Percentage Error/MAPE[12] for accuracy test, the specific formula is as follows:

$$MAPE = \frac{1}{n}\sum_{t}^{n}\left|\frac{O_{t+1} - T_{T+1}}{O_{t+1}}\right| \quad (2)$$

Where O_{t+1} is the predicted value at time, T_{t+1} is the actual value at t+1, and the MAPE value of the predicted result is 0.0284.

4. Conclusion

In this paper, a discrete HMM model is proposed to train the HMM model by processing the opening price, closing price, highest price, and lowest price as input data[13]. In the training process, ten groups of random initial data and a group of Vanke A-related probability data obtained from the law of large numbers are selected to ensure the most appropriate and robust model. Through the threshold setting, the model reduces the dependence on the initial value as much as possible. Before the input, wavelet denoising is used to remove the noise on the premise of retaining the effective information of the stock price. The number of implicit States is determined by using the OEHS criteria and is finally determined

ISSN 2616-5775 Vol. 5, Issue 11: 10-17, DOI: 10.25236/AJCIS.2022.051102

to be 4. The probability transition matrix is trained by the baumwelchalgo algorithm, the implicit state of the next day is deduced, and then the predicted value is obtained by the observation emission matrix. Finally, MAPE was used to test the accuracy of the predicted value, and the value was 0.0284.

References

[1] Zhang Xuan. Empirical research on stock price prediction based on Hidden Markov model and support vector machine [D]. Shandong University, 2019
[2] Box G, Jenkins G. Time series analysis: forecasting and control[M]. San Francisco: Holden-Day, 1976:2-11.
[3] Engle R F. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation [J]. Economet- Rica, 1982,50(4):987-1007.
[4] Werner F. M. De BONDT, RICHARD H. THALER. Further Evidence on Investor Overreaction and Stock Market Seasonality[J]. The Journal of Finance, 1987,42(3).
[5] Li Huang. BP Neural Network Algorithm Improvement and Application Research[D]. Chongqing Normal University, 2008
[6] Yu Wenli, Liao Jianping, Ma Wenlong. A new time series prediction method of stock price based on Hidden Markov model [J]. Computer application and software, 2010,27 (06): 186-190
[7] Weihong Wang, Pengyu Zhuo, Research on stock price prediction based on PCA-FOA-SVR [J]. Journal of Zhejiang University, 2016, 44(4): 399-404
[8] Bauml E, Petrie T. Statistical inference for probabilistic functions of finite state Markov chains[J]. The annals of mathematical statistics, 1996, 37(6): 1554-1563.
[9] Zhang Yan, Yang Yang. Denoising method and application of financial time series based on wavelet analysis [J]. Journal of Ningbo University (Science and Technology Edition), 2010, 23 (3): 56-59
[10] Xudong Zhang, Yufang Huang, Jiahao Du, Yongwei Liao. Stock price prediction based on Discrete Hidden Markov model [J]. Journal of Zhejiang University of technology, 2020,48 (02): 148-153 + 211
[11] Hassan, M. R., & Nath, B. (2005). Stock market forecasting using hidden Markov model: a new approach. In Proceedings of 5th international conference on intelligent system design and application (pp. 192–196). Poland.
[12] Huang Ran. Prediction and analysis of stock price based on Hidden Markov model [D]. Qingdao University, 2015.
[13] Yu Yongsheng, Xie Tianyidan, Liu Chang, Guo Jingwen, Zhang Weidong. Research on stock price behavior based on feature selection and HMM [J]. Information technology and network security, 2018,37 (08): 96-100. Doi: 10.19358/j.issn.2096-5133.2018.08.022