International Journal of Security and Its Applications
Vol. 10, No. 9 (2016), pp.19-28
http://dx.doi.org/10.14257/ijsia.2016.10.9.03

Detecting Fraudulent Financial Information of a Company Using Hidden Markov Model

Ruicheng Yang[1], Ailing Zuo[2] and Qing Shen[3]

Inner Mongolia University of Finance and Economics, Hohhot, 010051, China
[1]yang-ruicheng@163.com, [2]610479072@qq.com, [3]1508387502@qq.com

Abstract

Using Hidden Markov Model (HMM), this paper detects some fraudulent financial information of a company. The research index MBSPM (Main Business Service Profit Margin) ratio of company is used to test the effectiveness of HMM. Due to the shortage of the fraudulent data, based on the MBSPM index data we generate the artificial fraudulent financial information data by Poisson process and Uniform distribution, that is, we use the Poisson process to simulate the arrival times of the fraudulent financial information and the Uniform distribution to simulate their fraudulent sizes. By embedding the artificial fraudulent financial information data into the non-fraudulent MBSPM data, we form the sample data series. Applying Henderson filter method, we derive the research sample data by getting rid of the sample data series trends. We do some experiments to test the validity of HMM for detecting the fraudulent financial information, and further make some comparisons with other detecting techniques: logistic regression and ANNs. The detected results show that the HMM approach can significantly improve the identification accuracy.

Keywords: *Fraudulent information detection; Hidden Markov Model; MBSPM ratio*

1. Introduction

Fraudulent financial information often occurs when the manager of a company intentionally misleads users of financial reports by manipulating the financial information in a way that is outside the constraints imposed by generally accepted accounting principles. Detecting fraudulent financial information is difficult and continues to be an important concern for financial auditors and academic researchers [1]. In fact, a fraudulent financial data of a company can be regarded as an outlier because it is very different from normal non-fraud financial data, this has drawn a lot of research interests and formed a number of techniques, with special emphasis on some outlier identification models of machine learning techniques, such as logistic regression, ANNs (Artificial Neural Networks), HMM (Hidden Markov Model), and so on. Logistic regression is regarded as the baseline model in detecting the financial information fraud. Notable ones among these works include a study of the relation between the board of director composition and financial information fraud [2], a study of the relation between fraud type and auditor litigation [3] and another that throw light on the link between earnings and operating cash flows and incidence of financial reporting fraud [4]. In the real-life world, the series data of financial ratios often have some nonlinear features, then, ANNs is chosen as our primary tool to perform better in fitting these situations as they are a nonlinear, nonparametric function [5] [6], for examples Green and Choi [7], Lin *et al.* [8], Fanning and Cogger [9] and Feroz *et al.* [10]. Of course, ANNs is not an unmixed blessing methodology. A potential drawback of ANNs lies in the use of simulated rather than actual financial data. Unfortunately, it is extremely difficult to obtain enough quantity of real-life data that contains some degree of "contamination" (see [11] [12]). In contrast, HMM is used in a new way to detect the outliers in many areas such as in the

ISSN: 1738-9976 IJSIA

works [13] [14], T. Nguyen *et al.* [13] use HMM to discover the key features for the earthquake generation which are not accessible to direct observation, I. Votsi *et al.* [14] introduce an approach to cancer classification through gene expression profiles by designing supervised learning hidden Markov models (HMMs). Although these works don't directly explore the fraudulent financial information, but we can borrow their ideals and attempt to use it to detect the fraudulent fraud financial information. Essentially, HMM is a statistical model in which the system being modeled and it is assumed to be a Markov process with unobserved state, it is a type of stochastic modeling appropriate for non-stationary stochastic sequences, with statistical properties that undergo distinct random transitions among a set of different stationary processes. Different from a regular Markov model in which the state is directly visible to the observer, the state in a HMM is not directly visible, but output, dependent on the state, is visible. Here, HMM is used in a new way to detect and forecast the fraudulent financial information of a company. It provides a probabilistic framework for modeling a time series of observations. Details of the proposed method are provided in Section 2.

The remainder of this paper is organized as follows: Section 2 provides a brief overview on HMM, Section 3 gives the detecting method using HMM, Section 4 lists some experimental results, and finally, Section 5 concludes the paper.

2. HMM Reviews

To define a HMM, we introduce the following notations derived from Hassan and Nath's paper [15]:

- Number N: The number of hidden states in the model.
- Number M: The number of distinct observation symbols per state (observation symbols correspond to the physical output of the system being modeled).
- Length T : The length of observation sequence, *i.e.*, the number of observations.
- States $Q=\{q_1, q_2, \ldots\ldots, q_T\}$: hidden states at time t=1, 2, …, T.
- Observations $O=\{o_1, o_2, \ldots\ldots, o_T\}$: In any HMM, during time till T, there is a sequence of observations as $\{ o_1, o_2, \ldots\ldots, o_T \}$.
- Transition Matrix $A_{N\times N} = \{a_{ij}\}_{N\times N}$: Each element $a_{ij} \geq 0$ denotes the probability of transition from state i to state j such that $\sum_j a_{ij} = 1$.
- Observation Emission Matrix $B= \{b_j(O_t)\}$: $b_j(O_t) \geq 0$ represents the probability of observing O_t at state j such that $\sum_{t=1}^{T} b_j(O_t) = 1$.
- Prior Probability $\pi = \{\pi_i\}: \pi_i \geq 0$ represents the probability of being in state i at the beginning of the experiment such that $\sum_{i=1}^{N} \pi_i = 1$, *i.e.*, at time t = 1.

Technically, a HMM denoted as λ is considered as the triple $\lambda = (\pi, A, B)$. To work with HMM, the following three fundamental problems should be resolved:

1. Given the model $\lambda = (\pi, A, B)$ how do we compute $P(O|\lambda)$, the probability of occurrence of the observation sequence $O=\{ o_1, o_2, \ldots\ldots, o_T \}$.

2. Given the observation sequence O and a model λ, how do we choose a state sequence $q_1, q_2, \ldots\ldots, q_T$ that best explains the observations.

3. Given the observation sequence O and a space of models found by varying the model parameters A, B and π, how do we find the model that best explains the observed data.

There are established algorithms to solve the above questions. In our work we have used the forward-backward algorithm to compute $P(O|\lambda)$ (problem 1), Viterbi algorithm to resolve problem 2, and Baum-Welch algorithm to train the HMM (problem 3). The details of these algorithms are given in the tutorial by Rabiner [16].

3. Detecting Method using HMM

Detecting fraudulent financial information of a company will be complex and lengthy, a series of actions and activities in the form of several phases must be considered to break down the problem to conquer the complexity. Followed the steps in [16], Figure 1 depicts the overall process that is considered when solving the problem. Additionally, the following subsections will discuss each of the phases in more details.

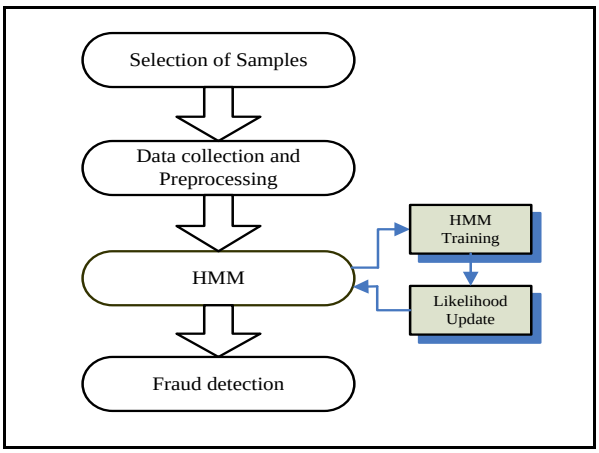

Figure 1. Detecting Process Review

3.1. Selection of Samples

Based on expert's knowledge, intuition, and previous research, it is important to identify some key financial items that are relevant for detecting financial fraud information. In this paper, we will focus on the fraudulent financial information of the listed company in China securities market. In China, some accounting items such as inflated revenue and profits are two main indicators of fraud behaviors, and these two items are usually reflected by the MBSPM (Main Business Service Profit Margin). So, we select MBSPM ratio as the sample to explore the fraudulent information of a company.

3.2. Data Collection and Preprocessing

Detecting the fraudulent financial information of a company with real-life data is a difficult task. Generally, research dataset should include two parts: training dataset and testing dataset. Testing dataset is easier to get, but the training dataset is relatively difficult to get because we can't get enough amounts of training data with many fraud behaviors. So, the collection of real training data with enough fraud behaviors is almost impossible. However, we can manually add some fraud information to the non-fraudulent data to form the research sample dataset. Since getting a large amount of real data for training without any fraud is relative easier, so, based on these real non-fraudulent data, we can create some artificial fraudulent data by Poisson process and Uniform

distribution, and combine the real non-fraudulent data and artificial fraud data into the research dataset. That is, the research dataset is generated by two parts, *i.e.*,

Research dataset= Real non-fraudulent dataset + Artificial fraudulent dataset (1)

Since the real non-fraudulent data series usually has some trends or periodical components, so, we must preprocess them by some statistical approach before these data enter the HMM training process. Here, we apply the Henderson filter to get rid of the series trend, thus, the above research dataset in (1) should be rewritten as follows:

Research dataset= Real filtered non-fraudulent dataset+ Artificial fraudulent dataset (2)

3.3. HMM Training

HMM training is the key issue to solve problem 3 in Section 2. For matching the preprocessed research data of financial ratio, we estimate the parameters of model $\lambda = (\pi, A, B)$, and give the following settings and considerations as follows:

- States: In the experiments, we have N = 5; intuitively, it is denoting the stages in time that are allowed in different transitions in the HMM training.
- Mixtures: Let M = 2.
- Left-Right Delta: Experimentally, we have tried Δ=1 and Δ= 3.
- Prior Probability: Adhering to the left-right HMM, we have π = (1; 0; 0; 0; 0).

HMM Training: According to [15] [16], we need to use Baum-Welch learning algorithm to train the HMM. Baum-Welch training algorithm does its job based on a number of iterations to tune and update the parameters of the model, namely the transition probability *B*.

3.4. Likelihood Update and Fraud Detection

In the path to detect the fraud information of a company, we must process the following three phases:

Firstly, there is a need to compute the likelihood of the historical dataset (training dataset) in the past time. When having the data at some specific time, it is straightforward to compute the likelihood from the HMM. This is Problem I in Section 2 which is computed using forward backward algorithm that is proposed in [15] [16].

Secondly, we locate the past behaviors which would be similar to that of current time (denoted the current time by symbol '*c*'). To do this, we obtain the likelihood value for the observation sequence on time '*c*'. Each of the observation sequence is built by using the MBSPM ratio. For better explaining this, let us assume that the likelihood value for observation sequence on time '*c*' is 'L_c'. Now, from the historical dataset, using the HMM, the observation sequence is located in which data would produce the same or close to the 'L_c' value. Say, the HMM find the *m*-th time observation which produces the same likelihood value or nearest to the 'L_c'. As we know, the financial ratio of the current time '*c*' should follow the same or similar past pattern. Denoting the data value at the *m*-th time is V_m, then, we can predict that the data value V_c at the current time *c* is V_m.

Finally, we set a threshold value ε (in this paper $\varepsilon = 0.1$) and compute the relative difference $\left|\frac{V_c - V_m}{V_c}\right|$ (here V_c is the real value of the current time '*c*'), and the decision rule is given as follows:

$$\begin{cases} \text{if } \left|\frac{V_c - V_m}{V_c}\right| > \varepsilon, & \text{the data includes fraud information} \\ \text{if } \left|\frac{V_c - V_m}{V_c}\right| \le \varepsilon, & \text{the data doesn't include fraud information} \end{cases} \quad (3)$$

Based on the above steps, we can successfully derive the actual statue of the current time 'c .

4. Numerical Experiments

4.1. Data Preprocessing Numerical Analysis

Now we select the MBSPM ratio of GZMT (GuiZhou MaoTai) company as the real non-fraudulent sample data, the data series comes from the web site http://finance.sina.com.cn. A total of 48 quarterly data is chosen from years 2001 to 2013. Based on the non-fraudulent sample data, we generate the artificial fraudulent data by the Poisson process and Uniform distribution. Here, the Poisson process $P(\lambda)$ with λ=0.25 is to simulate the arrival time (or embedded time) of the fraudulent financial information, and the Uniform distribution $U(a, b)$ with a=0.1 and b=0.3 is to simulate the corresponding fraudulent sizes at the embedded time. For convenience, we renumber the quarterly times from years 2001 to 2013 as 1-48 times, and the real non-fraudulent data and the synthetic data are shown in Figure 2.

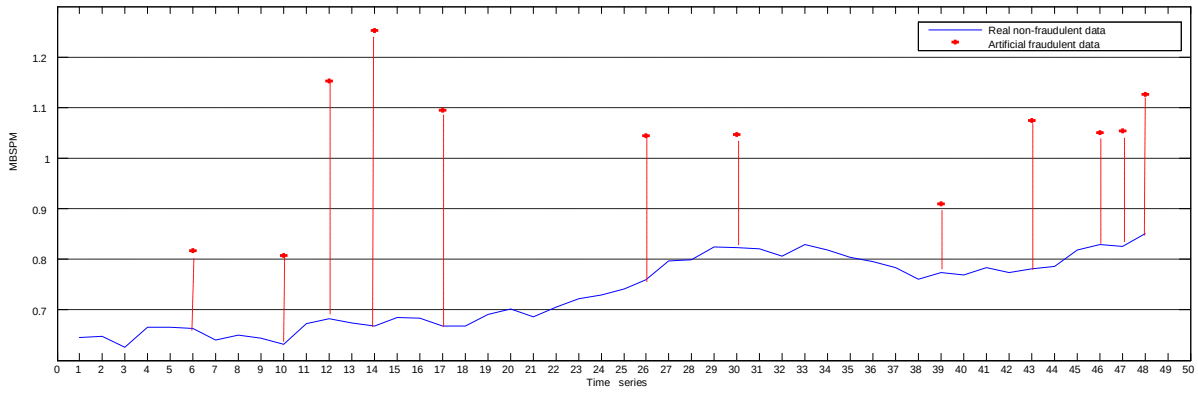

Figure 2. Real Non-fraudulent Data and Artificial Fraudulent Data

The sample data in Figure 2 are divided into two parts: the first 36 samples as training data, and the other 12 samples as testing samples. From Figure 2, we know the sequential non-fraudulent data has some trends, so, we first get rid of the trend before the sample data enter the HMM process. The Henderson filters are a widely used set of trend filters, or smoothers, then, we apply the Henderson filter technique to estimate the data series trend. But, in the real-life world, we don't know whether the testing data is embedded fraudulent information or not before we detect them, so, the data which we will use to estimate the trend should be the combined data with training sample data (not including non-fraud information) and test samples (containing some fraudulent information). The trend details can see Figure 3, and the data series that has been got rid of trends by Henderson filters technique see Figure 4. So far, we have formed the training dataset and testing dataset (see Figure 4).

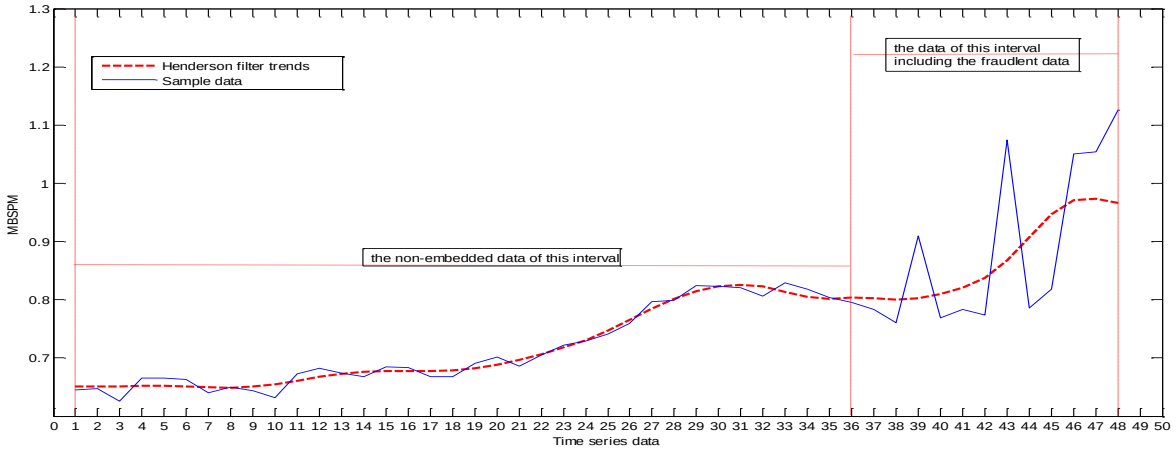

Figure 3. Trend Analysis

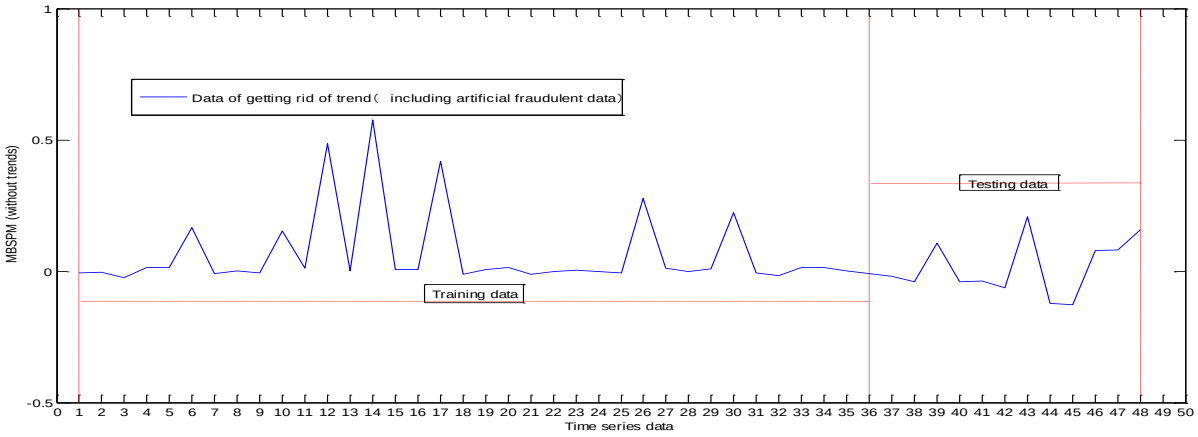

Figure 4. Sample Data Series without Trends

4.2. Numerical Analysis

Now we use the training dataset to train the HMM model. After this, we test the trained model with testing dataset. For convenience, we introduce the symbol '+' to represent whether the real data is embedded fraud information or not, more explicitly, symbol '+' is defined as follows:

$$\begin{cases} + = 0, \text{if the real data has no embedded fraud information} \\ + = 1, \text{if the real data has been embed fraud information} \end{cases} \tag{4}$$

Similarly, introducing another symbol 'O' for the detected result, define the symbol 'O' as follows:

$$\begin{cases} O = 0, \text{if the judgement is no embedded fraud information} \\ O = 1, \text{if the judgement is embed fraud information} \end{cases} \tag{5}$$

Therefore, according to (4) and (5), if 'O' is in coincidence with '+', that is, the two symbols 'O' and '+' overlaps together into one symbol '⊕', this means the judgment is right; otherwise, if 'O' is separated from '+' , this shows the judgment is wrong.

Now we give one experiment, in which the experiment is embedded fraud information for 5 times, *i.e.*, there are 5 times '+=1'. Using the above HMM method, we can successfully identified 4 times, *i.e.*, we get 4 times '⊕=1', the detected results see Figure 5.

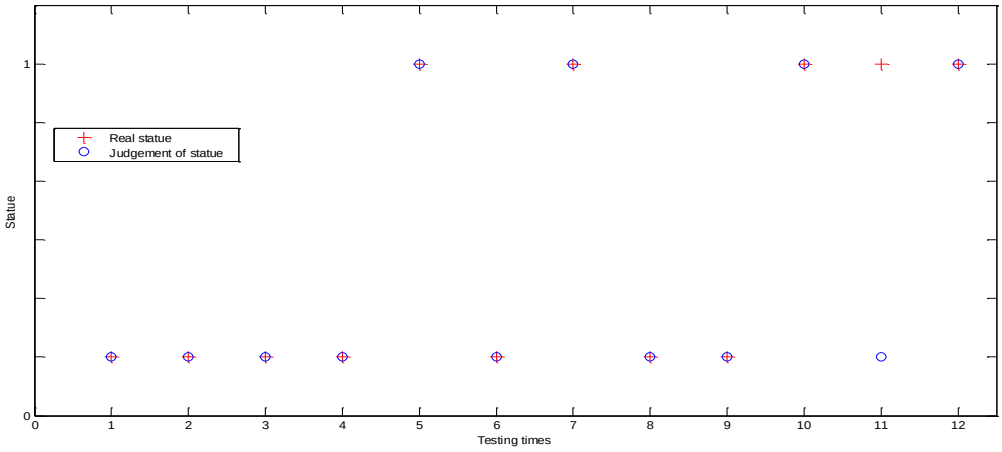

Figure 5. Detected Results using HMM

Furthermore, using the trained HMM model we can forecast the MBSPM index data, and give the contrast results in Figure 6. Figure 6 shows that only the value of the 11-th time (the third quarter of 2013) is far from the true value, and the other forecasting values are very close to the true values.

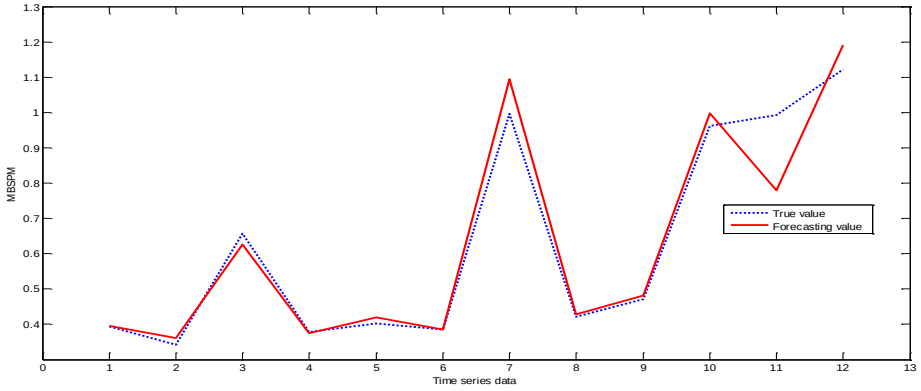

Figure 6. Forecasting results of MBSPM

In order to illustrate the performance for detecting financial fraud information with HMM method, we further do 10 experiments where we embed 37 times fraud information. We summary all the detected results in one picture (see Figure 7). In Figure 7, all the misjudgment time points are labeled with symbol '*' at the time series axis, we can find there are 6 times misjudgments, and the total number of correct rate is 83.8%. In addition, there is another judgment error, that is, at the 110-th time, we can see that the real statue is no embedded fraud information, but the judging result show that the statue is embedded with fraud information, this is another type judgment error. However, the experiment results still show that we have identified successfully 31 times, this means that the HMM method is effective for detecting the financial fraud information.

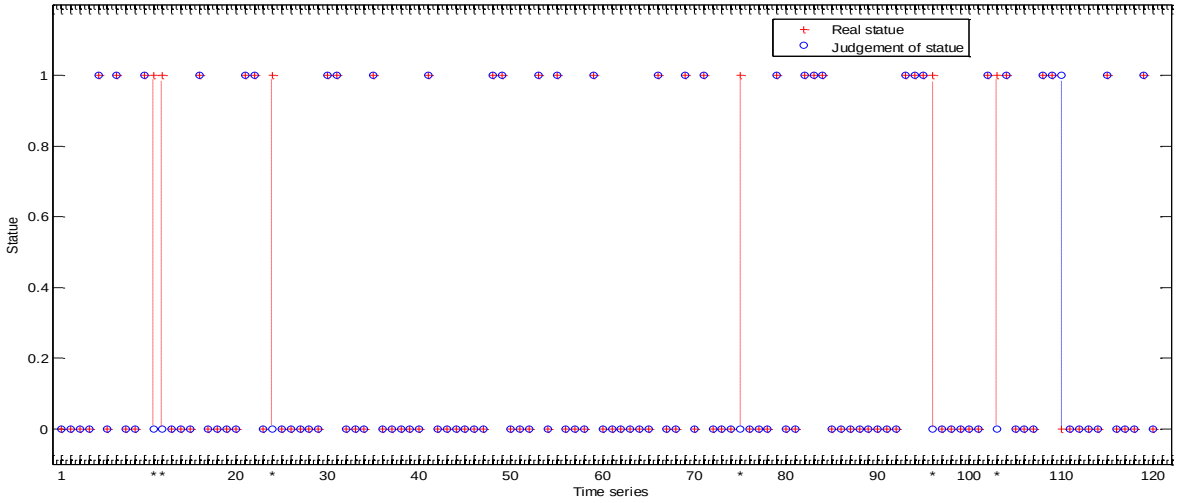

Figure 7. Detected Results of 10 Experiments

4.3. Comparison Analysis

For further illustrating the advantage of the proposed HMM approach, with the same dataset derived from 10 experiments in the above sub-section, we use the logistic regression and ANNs to detect these embed fraud information, and summary all the detected results with logistic regression, ANNs and HMM techniques in Table 1.

Table 1. Detecting Accuracy with Logistic Regression, ANNs and HMM Techniques

Model	Total amounts of embed fraud information	Amounts of correct identification	Amounts of error identification	Total accuracy rate
Logistic regression	37	24	13	64.86%
ANNs	37	29	8	78.38%
HMM	37	31	6	83.78%

Table 3 shows that the total detecting performance results with different techniques. As can be seen from Table 3, the accuracy rates are 64.86% with logistic regression and 78.38% with ANNs. They are far lower than that of HMM technique. This shows that the HMM approach can significantly improve the identification accuracy.

5. Conclusions

Identifying the financial fraud information of a company is always a hot research in the world. Because the HMM method is relatively mature in the biological information science, fault diagnosis and forecasting fields, this paper attempts to introduce this method to detect the financial fraud information. We take MBSPM ratio as our research index, combine the simulation data and the real data into the research sample data. Using Henderson filter technique and HMM method, we do some experiments to test the detecting ability of HMM. Numerical results and the comparison analysis with logistic regression, and ANNs approaches show that the HMM approach can significantly improve the accuracy for detecting the embedded fraud information. Moreover, this

method can be easily extended to detect the fraudulent information with multidimensional financial indicators, the algorithm only need a transformation from one dimensional algorithm to multidimensional algorithm, but due to the lack of available real-life data, this paper doesn't give further discussion.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No. 71261015), Program for Innovative Research Team in Universities of Inner Mongolia Autonomous Region (No. NMGIT1405), Program for Grassland Talent Engineering of Inner Mongolia Autonomous Region, Program for Young Talents of Science and Technology in Universities of Inner Mongolia Autonomous Region, and Program for Teaching Team in Universities of Inner Mongolia Autonomous Region.

References

[1] J. S. Hammersley, "Auditing: A Journal of Practice & Theory", vol. 30, no. 4 **(2011)**.
[2] M. S. Beasley, "The Accounting Review", vol. 71, no. 4, **(1996)**.
[3] S. E. Bonner, Z. V. Palmrose and S. M. Yong, "The Accounting Review", vol. 73, no. 4, **(1998)**.
[4] T. A. Lee, R. W. Ingram and T. P. Howard, "Contemporary Accounting Research", vol. 16, no. 4, **(1999)**.
[5] S. Bhattacharya, D. Xu and K. Kumar, "Decision Support Systems", vol. 50, no. 3, **(2010)**.
[6] J. L. Perols, "Detecting financial statement fraud-three essays on fraud predictors", multi-classifier combination, and fraud detection using data mining, http://scholarcommons.usf.edu/etd, **(2008)**.
[7] B. P. Green and J. H. Choi, "Auditing: A Journal of Practice & Theory", vol. 16, no. 1, **(1997)**.
[8] J. Lin, M. H. wang and J. Becker, "Managerial Auditing Journal", vol. 18, no. 8, **(2003)**.
[9] K. Fanning and K. Cogger, "International Journal of Intelligent Systems in Accounting", Finance and Management, vol. 7, no. 1, **(1998)**.
[10] E. Feroz, T. Kwon, V. Pastena and K. Park, "International Journal of Intelligent Systems in Accounting", Finance & Management, vol. 9, no. 3, **(2000)**.
[11] S. Bhattacharya, D. Xu and K. Kumar, "Decision Support Systems", vol. 50, no. 3, **(2011)**.
[12] B. Busta and R. Weinberg, "Managerial Auditing Journal", vol. 13, no. 6, **(1998)**.
[13] T. Nguyen, A Khosravi, D. Creighton and S. Nahavandi, "Information Sciences", vol. 316, **(2015)**.
[14] I. Votsi, N. Limnios, G. Tsaklidis and E. Papadimitriou, "Physica A: Statistical Mechanics and its Applications", vol. 392, no. 13, **(2013)**.
[15] Md. R. Hassan and B. Nath, "Stock market forecasting using hidden markov model: a new approach", 5th International Conference on Intelligent Systems Design and Applications, **(2005)**, pp. 192-196.
[16] R. L. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition", Proceedings of the IEEE, vol. 77, no. 2, **(1989)**, pp. 257-286.

Authors

Ruicheng Yang, Dr., Professor, researcher in Inner Mongolia University of Finance and Economics. He is an expert in financial information processing, the theory of financial mathematical modeling and computation.

Ailing Zuo, Librarian of library in Inner Mongolia University of Finance and Economics. She is an expert in the theory information modeling and computation.

Qing Shen, Postgraduate in Inner Mongolia University of Finance and Economics. He majors in financial information processing.