Adaptive Anomaly Detection of Coupled Activity Sequences

Yuming Ou, Longbing Cao and Chengqi Zhang

***Abstract*—Many real-life applications often involve multiple sequences, which are coupled with each other. It is unreasonable to either study the multiple coupled sequences separately or simply merge them into one sequence, because the information about their interacting relationships would be lost. Furthermore, such coupled sequences also have frequently significant changes which are likely to degrade the performance of trained model. Taking the detection of abnormal trading activity patterns in stock markets as an example, this paper proposes a Hidden Markov Model-based approach to address the above two issues. Our approach is suitable for sequence analysis on multiple coupled sequences and can adapt to the significant sequence changes automatically. Substantial experiments conducted on a real dataset show that our approach is effective.**

***Index Terms*—Multiple coupled sequences, Anomaly, HMM, Adaptation, Stock market.**

I. Introduction

TYPICAL sequence analysis [4], [9], [7], [1], [10] mainly focuses on identifying patterns on one sequence. However, dealing with the real-life problems, we often have to face multiple interacting sequences rather than only one single sequence. For example, in stock markets there are three coupled sequences including buy orders, sell orders and trades by matching orders from both buy and sell sides. These three sequences are coupled with each other in terms of many aspects such as timing, price and volume. The interacting relationships among them contain rich information which is very valuable to stock market surveillance. As price manipulators may deliberately place their buy orders and/or sell orders and indirectly affect the trade price through manipulating the interaction between them, the interaction is an important clue to identifying stock price manipulations. If we study the three sequences separately or simply merge them into one sequence, the valuable information about their interacting relationships would be of course lost.

In real-life applications, we also often face another issue that is the significant changes in sequences. For instance, the trading activities in stock markets change frequently due to the investors' sentiment and the external market environment, resulting in the potential significant changes in the three coupled sequences. Thus it is necessary for sequence analysis methods to identify the significant changes and adapt to the new environment.

In this paper, we employ agent technology to develop a pattern mining system to detect abnormal trading activity patterns in the three coupled sequences including buy orders, sell orders and trades. The system uses six Hidden Markov Model(HMM)-based models to model the trading activity sequences in different ways: three standard HMMs for modeling single sequences respectively; an integrated HMM model combining all individual sequence-oriented HMMs; a Coupled HMM reflecting coupled relationships among sequences; and an Adaptive Coupled HMM to automatically capture the significant changes of activity sequences. The above six HMM-based models compete with each other. The outputs generated by the best model are used as the final outputs of system.

The rest of this paper is organized as follows. We present the system framework in Section II. After Section III introduces the modeling of trading activity sequences by HMM-based methods, Section IV provides the approach to identify abnormal activity patterns using six HMM-based models. The model selection and evaluation are introduced in Section V, and the experimental results are given in Section VI. Finally, Section VII concludes this paper.

II. Agent-Based Framework for Discovering Abnormal Patterns in Coupled Sequences

To make our system autonomous, we use agent technology to build the system. As shown in Figure 1, the system consists of the following main agents: Activity Extraction Agent, Anomaly Detection Agent, Change Detection Agent, Model Adjusting Agent, and Planning Agent. They collaborate with each other to find out the best model, and then deploy this best one for activity pattern discovery. In particular, to adapt to the source data dynamics, Change Detection Agent detects changes in the outputs of CHMM, and then the Planning Agent triggers the adjustment and retraining of the CHMM model (More details are introduced in Section III-C).

III. Modeling Activity Sequences by Hidden Markov Model-Base Methods

In this section, we first introduce the approaches to build Hidden Markov Model (HMM) [8], [5] for single activity sequence and Coupled Hidden Markov Model (CHMM) [2], [6] for multiple coupled activity sequences respectively, and then improve the CHMM by adding an automatically adaptive mechanism to it to create an Adaptive CHMM (ACHMM).

Yuming Ou is with the Faculty of Engineering and Information Technology, University of Technology Sydney.

Longbing Cao is with the Faculty of Engineering and Information Technology, University of Technology Sydney.

Chengqi Zhang is with the Faculty of Engineering and Information Technology, University of Technology Sydney.

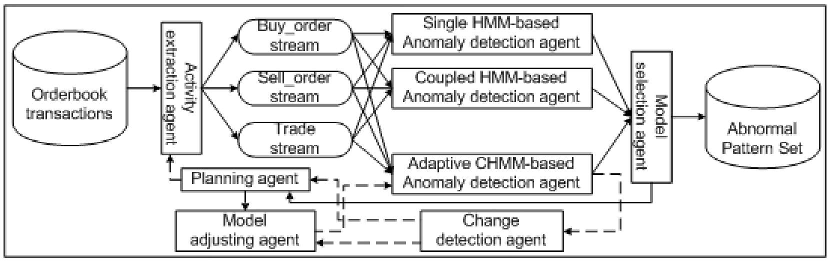


Fig. 1. Agent-Based Framework for Identifying Abnormal Activity Patterns

A. Modeling Single Activity Sequence by HMM

To model the single activity sequences including buy-order, sell-order and trade sequences separately, we build three models *HMM-B*, *HMM-S*, and *HMM-T* for them based on the standard HMM. The hidden states and observation sequences of the three models are defined as follows:

- In model *HMM-B*, the hidden states S^{buy} represent the investors' belief, desire and intention (BDI) on buy side, S^{buy} =\{*Positive Buy, Neutral Buy, Negative Buy*\}. In model *HMM-S*, the hidden states S^{sell} denote the investors' BDI on sell side, S^{sell} =\{*Positive Sell, Neutral Sell, Negative Sell*\}. In model *HMM-T*, the hidden states S^{trade} stand for the market states, S^{trade} =\{*Market Up, Market Down*\}. The exact values of the hidden states are unknown, while they can change from one to another with particular probabilities. For example, we cannot know the investors' BDI is *Positive Buy*, *Neural Buy* or *Negative Buy* actually.
- The observation sequences IA^{buy}, IA^{sell} and IA^{trade} stand for the activity sequences of buy-order, sell-order and trade respectively. The values of these activity sequences of buy-order, sell-order and trade can be observed. In the following, we will detail the method for constructing trading activity sequences.

The construction of trading activity sequences is based on two concepts: $activity$ (A) and $interval\ activity$ (IA), which involve human intention information including prices and volumes in stock markets.

Definition 1: Activity (A) is an action (a) and it is associated with BDI information (represented in p and v).

$$A = (a, p, v) \tag{1}$$

$$a = \begin{cases} buy\ order, & at\ time\ t \\ sell\ order, & at\ time\ t \\ trade, & at\ time\ t \end{cases} \tag{2}$$

$$p = \begin{cases} trade\ price, & of\ trade\ at\ time\ t \\ order\ price, & of\ buy\ or\ sell\ order\ at\ time\ t \end{cases} \tag{3}$$

$$v = \begin{cases} trade\ volume, & of\ trade\ at\ time\ t \\ order\ volume, & of\ buy\ or\ sell\ order\ at\ time\ t \end{cases} \tag{4}$$

Definition 2: Interval Activity (IA) represents the actions and BDI information associated with the activity sequence taking place during a window l (the window size is denoted by w).

$$IA_l = (A_l, P_l, V_l, W_l) \tag{5}$$

which is calculated as follows:

$$A_l = \{A_{l1}, A_{l2}, \ldots, A_{ln}\} \tag{6}$$

$$P_l = \frac{\sum_{i=1}^{n} p_i}{W_l} \tag{7}$$

$$V_l = \frac{\sum_{i=1}^{n} v_i}{W_l} \tag{8}$$

$$W_l = n \tag{9}$$

where n is the number of activities in the window l.

In stock markets, orders normally do not last for more than one day. Order are placed by investors after market opens and are expired after market closes if they have not been traded. Trades are also based on the orders placed on the same day only. This market mechanism indicates that all orders and trades on a same day are closely related. Thus we construct the sequences for buy order, sell order and trades respectively by grouping the IAs that fall into a same trading day together.

B. Modeling Multiple Coupled Activity Sequences by CHMM

In order to reflect the interacting relationship among the three activity sequences, we use a CHMM consisting of three chains of HMM to model the buy-order, sell-order and trade processes together. As shown in Figure 2, the circles denote the hidden states of the three processes while the squares stand for their observation sequences. The three chains are fully coupled with each other reflecting their interactions.

C. Adapting to Significant Activity Sequence Changes

In order to adapt the significant changes that often exist in trading sequences, we involve multi-agent technology to enhance the CHMM, and form an agent-based adaptive CHMM (ACHMM).

The adaptation of ACHMM is mainly based on the detection of change between the current outputs of model and the current benchmark. The current benchmark is defined as the outputs generated after the last update of model. Three agents contribute to the adaption: Change Detection Agent, Model Adjusting Agent and Planning Agent. The Change Detection Agent checks whether there is a significant difference between the current outputs and the current benchmark based on statistical test methods, for instance, t test. The significant

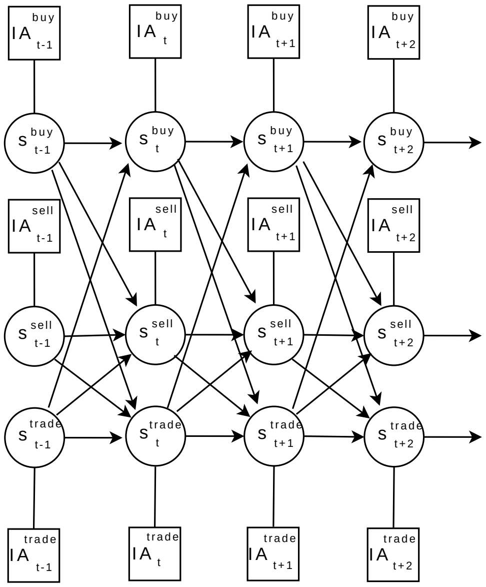


Fig. 2. CHMM for Modeling Multiple Coupled Activity Sequences

difference suggests that there is a change in the trading activities and the CHMM cannot model the trading activities properly, therefore the model needs to be updated. Once the change is detected, the Planning Agent will receive a notice and trigger the Model Adjusting Agent to retrain the model. The outputs generated after this update are the new current benchmark.

IV. Identifying Abnormal Trading Activity Sequences

After the models discussed above are trained, they can be used to identify abnormal trading activity sequences. The basic idea is to calculate the distances from the test sequences to the centroid of model. If the distance of a sequence is larger than a user-specified threshold, then the sequence is considered to be abnormal.

The formulas to compute the centroid (μ) and radius (σ) of model $\mathcal{M} \in \{$*HMM-B, HMM-S, HMM-T, IHMM, CHMM, ACHMM*$\}$ are as follows:

$$\mu = \frac{\sum_{i=1}^{K} Pr(Seq_i|\mathcal{M})}{K} \tag{10}$$

$$\sigma = \sqrt{\frac{1}{K}\sum_{i=1}^{K} Pr(Seq_i|\mathcal{M})) - \mu} \tag{11}$$

where Seq_i is a training sequence and K is the total number of training sequences.

The distance $Dist_i$ from a test sequence $Seq_i^{'}$ to model $\mathcal{M}$ is calculated by the following formula:

$$Dist_i = \frac{\mu - Pr(Seq_i^{'}|\mathcal{M})}{\sigma} \tag{12}$$

Consequently, $Seq_i^{'}$ is an abnormal sequence, if it satisfies:

$$Dist_i > Dist_{max} \tag{13}$$

where $Dist_{max}$ is a given threshold.

V. Model Selection and Evaluation

There are six HMM-based models for modeling the trading activity sequences in the system, including: 1) *HMM-B*: an HMM on buy-order sequences only; 2) *HMM-S*: an HMM on sell-order sequences only; 3) *HMM-T*: an HMM on trade sequences only; 4) *IHMM*: an integrated HMM combining *HMM-B*, *HMM-S* and *HMM-T*. The probability of *IHMM* is the sum of the probability values of the three models. This model does not consider the interactions among the three processes; 5) *CHMM*: a Coupled HMM for trade, buy-order and sell-order sequences, considering their interactions; and 6) *ACHMM*: an Adaptive Coupled HMM which is able to adapt to the significant changes in sequences automatically.

The selection of the best model amongst these candidates is conducted by the Model Selection Agent. The selection policies conducted by the agent are as follows.

Policy 1 selectBestModel

Rule 1: Select the X ($X > 1$) best candidate models by evaluating the technical performance;
Rule 2: Select the best model from the X ($X > 1$) best models by checking business performance.

The technical performance evaluation of model is based on the following metrics: *accuracy, precision, recall, specificity*. These four technical metrics measure the quality of the models. Furthermore, we introduce a business metric widely used in capital markets to evaluate the business performance of model. This business metric is *return (R)* [3], which refers to the gain or loss for a single security or portfolio over a specific period. It can be calculated by

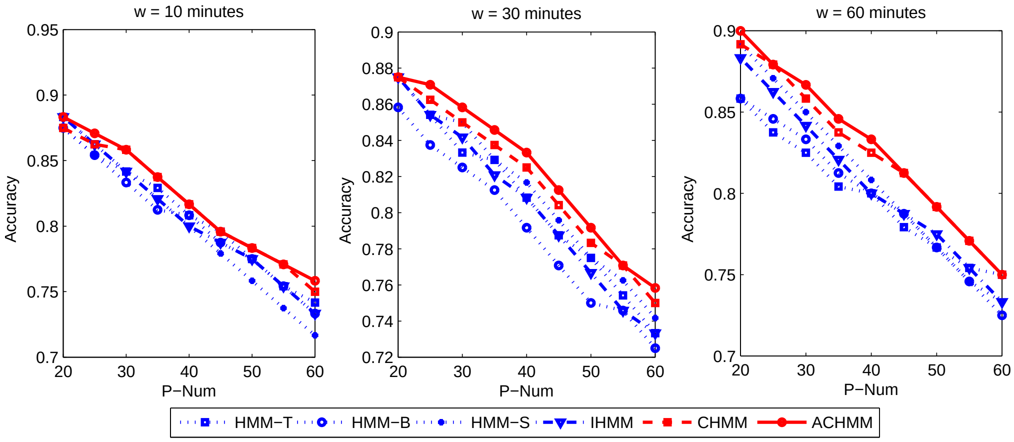
$$R = ln\frac{p_t}{p_{t-1}} \tag{14}$$

where p_t and p_{t-1} are the trade prices at time t and t-1, respectively. Empirically, the trading days with exceptional patterns are more likely to incur higher daily *return* than those without exceptional patterns.

VI. Experimental Results

Our system is tested on a real dataset from a stock exchange, which covers 388 trading days from June 2004 to December 2005. In the dataset, there are some trading days associated with alerts that are generated by the surveillance system used in that stock exchange. These alerts can be used to label the data, that is, the data with alerts is labelled as true anomalies. After labelling the data, we divide the whole dataset into two parts: one is for training models and another is for testing models.

Model *HMM-B*, *HMM-S*, *HMM-T* and *IHMM* are trained by the standard Baum-Welch algorithm [8] respectively, while model *CHMM* and *ACHMM* are trained by the algorithm proposed in [2], which is similar to the Baum-Welch algorithm, respectively. After these six models are trained, they are tested on the test data. The Model Select Agent will choose the best model in terms of their technical and business performance as presented in Section V.



;

Fig. 3. Technical Performance of Six Systems: Accuracy

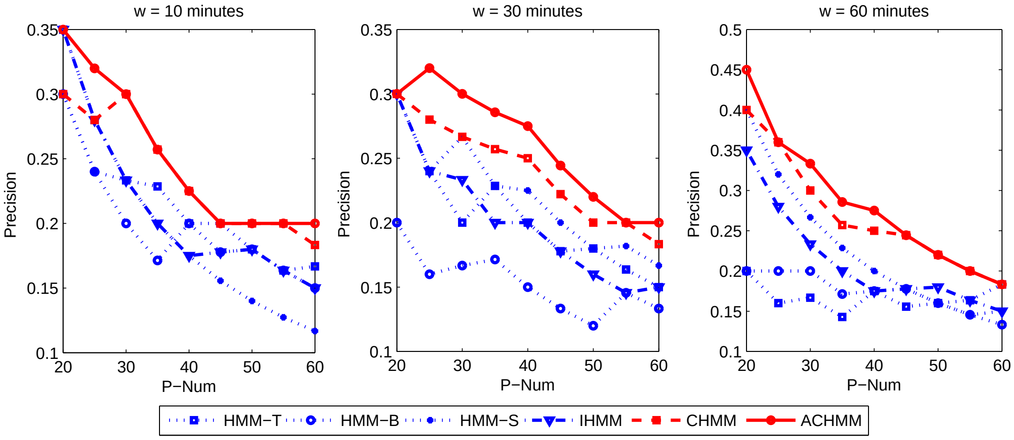


Fig. 4. Technical Performance of Six Systems: Precision

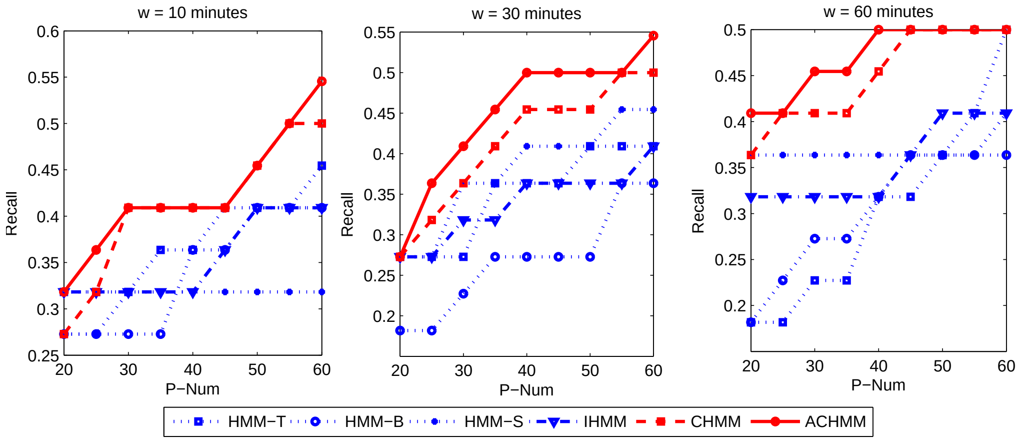


Fig. 5. Technical Performance of Six Systems: Recall

Figures 3, 4, 5 and 6 show the technical performance of the six models, where x axis (P-Num) stands for the number of detected abnormal activity patterns, and y axis represents the values of technical measures. Clearly *ACHMM* outperforms the other five models under different window sizes (w).

In terms of the business performance, Figure 7 shows the business performance of the six models, where y axis represents the values of average daily *return* of trading days in where abnormal activity patterns are detected. We can see that *ACHMM* also outperforms the other five models under different w.

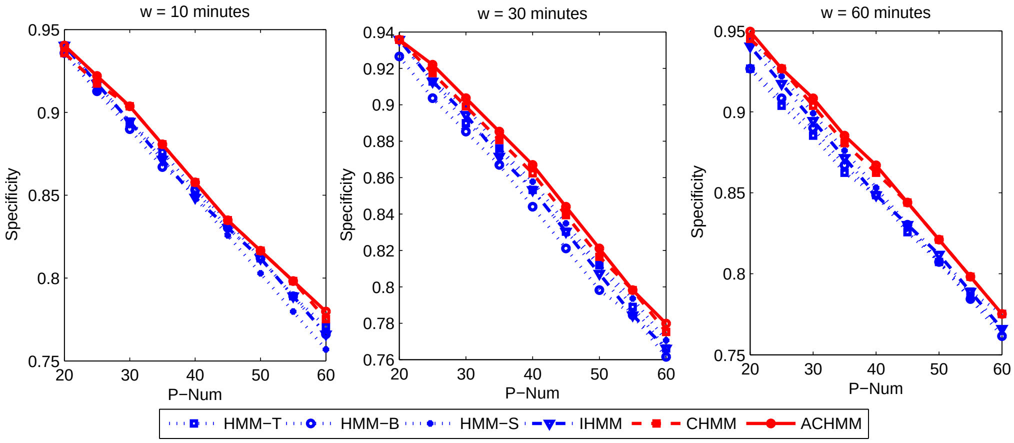


Fig. 6. Technical Performance of Six Systems: Specificity

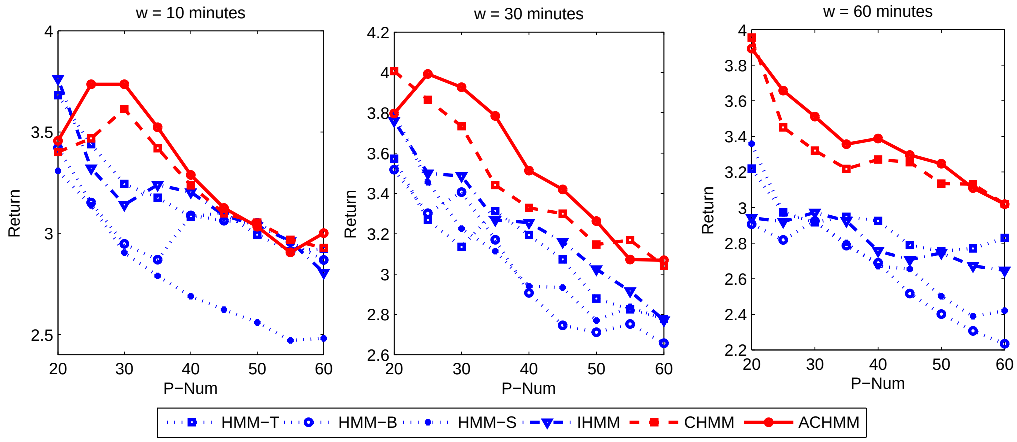


Fig. 7. Business Performance of Six Systems: Return

VII. Conclusion and Future Work.

Many real-life applications, such as detecting abnormal trading activity patterns in stock markets, often involve multiple coupled sequences. Typical existing methods mainly pay attention to only one single sequence. As the interacting relationships among the multiple coupled sequences contain valuable information, it is unreasonable to study the multiple coupled sequences separately by those existing methods. Furthermore, in practice such coupled sequences change frequently, which greatly challenges the trained model.

Taking the detection of abnormal trading activity patterns in stock markets as an example, this paper proposed a HMM-based approach to address the above two issues widely existing in real-life applications. Our approach caters for the sequence analysis on multiple coupled sequences and also can be used under the circumstances in which sequences change frequently. Substantial experiments conducted on a real dataset show that our approach is effective.

Our further work is on generalizing our approach for dealing with other application problems, investigating the update of existing sequence analysis methods for analyzing multiple coupled sequences, and comparing them with our HMM-based models.

References

[1] J. Ayres, J. Flannick, J. Gehrke and T. Yiu, *Sequential Pattern mining using a bitmap representation*, SIGKDD02, pp. 429–435, 2002.

[2] M. Brand, *Coupled hidden Markov models for modeling interacting processes*, Tech. Rep., MIT Media Lab, 1997.

[3] S. J. Brown and J. B. Warner, *Using daily stock returns: the case of event studies*, Journal of Financial Economics, vol. 14, pp. 3–31, 1985.

[4] G. Dongand and J. Pei, *Sequence Data Mining*, Springer, 2007.

[5] R. Durbin, S. R. Eddy, A. Krogh and G. Mitchison, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, 1999.

[6] N. M. Oliver, B. Rosario and A. P. Pentland, *A Bayesian computer vision system for modeling human interactions*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, pp.831–843, 2000.

[7] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal and M. C. Hsu, *Prefixspan: mining sequential patterns efficiently by prefix-projected pattern growth*, ICDE2001, pp. 215-226, 2001

[8] L. R. Rabiner, *A tutorial on hidden Markov models and selected applications in speech recognition*, Proceedings of the IEEE, vol. 77, pp. 275-286, 1989.

[9] R. Srikant and R. Agrawal, *Mining sequential patterns: Generalizations and performance improvements*, EDBT1996, pp. 3–17, 1996.

[10] M. J. Zaki, *SPADE: An efficient algorithm for mining frequent sequences*, Machine Learning, vol. 42, pp. 31–60, 2001.