

# Anomaly Detection of Clinical Behavior Sequences

Hebiao Yang

School of Computer Science and Telecommunication Engineering, Jiangsu University  
Jiangsu 212013, China

Xiaodong Yuan (Corresponding author)

School of Computer Science and Telecommunication Engineering, Jiangsu University  
Jiangsu 212013, China  
E-mail: yxdzhs@163.com

*This research was financially supported by Jiangsu high-tech research project (Grant No. BG2007028).*

## Abstract

The identification of abnormal clinical behavior during the process of treatments is of great significance for regulating the standard medical behavior. Due to clinical behavior constrained by time, and the timing of subsequence, GSP algorithm was modified in the present paper, and described the timing of subsequence by the introduction of the concept of legal subsequences in order to detect the frequent patterns in sequences; sequence association rules in accordance with the characteristics of territorial behavior were screened using association rule methods in order to establish rule base; Comparing the similarity between the detected frequent patterns and normal behavior rules, anomaly detection of the detected behavior was operated and the validity of the methods was verified through experiments.

**Keywords:** Clinical behavior, Sequence association rules, Anomaly detection, Similarity

## 1. Introduction

Abnormal behavior in the process of clinical diagnosis was the medical behavior that was not in consistent with the diagnosis and treatment (Yang, 2004, PP. 601-603). The nonstandard treatment affected the normal treatment, but also caused damages to the patient's physical and mental health. Anomaly detection in clinical behaviors was primarily to seek unreasonable, or unusual clinical behavior in the process of clinical treatments, with attempts to regulate clinical diagnosis and treatment behavior and prevent the occurrence of illegal actives, and its critical point was to establish normal clinical behavior model, and compare and estimate the current clinical behavior using the model.

Association rule was mainly applied to analyze the links of the same record among different attributes, and one of the earliest technologies applied in the anomaly detection of data mining, such as data mining from the behaviors in the area of intrusion detection systems (Bruno, 2007) and credit card fraud detection (Sanchez, 2009, PP. 3630-3640) in order to find abnormality. Characteristics of such behaviors were that they were independent from one another and occurred in order, and also belonged to Boolean data. In a literature (Khelifa, 2007), continuous data was processed with association rule mining, and all these data was not continuous in order and failed to occur at the same time. The relationship among clinical behaviors not only contained independent parts, but also continuous ones. As for the intraoperative blood transfusion process, it was always operated with other behaviors together. Since clinical behavior occurring during the process of single disease clinical diagnosis and treatment in clinical path followed a certain sequence, and the sequence was relatively constant and stable. For example, clinical pathway for acute appendicitis surgery treatment included simple routine examination, application of preventive medication, anesthesia, surgery, pathological examination, post-operative check-up and so on. The relationship among these behaviors was enormously complicated with strong correlation, and therefore, association patters among behavioral sequences were observed using association rule methods.

Consequently, aimed at the characteristics of clinical behaviors, we attempted to develop an anomaly detection method of clinical behaviors based on association rule. We sought and found correlations among these clinical behavior sequences under the normal condition, obtained sequence association rules, and finally ascertained whether the behaviors were abnormal or not according to the similarity of rule set between the detected behavior and normal behavior.

## 2. Anomaly detection model of clinical behaviors

Figure 1 was the anomaly detection model of clinical behaviors, and this model based on behavior rules during the process of clinical diagnosis and treatment. In the studying stage of modeling, normal clinical data was required to collect, and therefore, the rule we learned was the one under normal conditions. During detection, if several behaviors didn't conform to these rules, the rules might be abnormal.

## 3. Anomaly detection based on sequence association rules

### 3.1 Analysis

Behaviors during the process of single disease clinical diagnosis and treatment based on clinical path were related to the diagnosis and treatment behaviors for this disease. Clinical behavior sequence association rule during the process of clinical diagnosis and treatment possessed the following form:  $\{p_1, p_2, \dots, p_n\} \rightarrow p$ .

The rules could be understood that after clinical behavior  $p_1, p_2, \dots, p_n$  occurred during the process of treatment, the diagnosis and treatment behavior P was always conducted. If clinical diagnosis and treatment behavior conformed to this rule in certain process, it indicated that the diagnosis and treatment behavior might be normal, and otherwise it was abnormal.

In a literature (Wang, 2008), clinical sequent obtained from sequence generation algorithm was expressed as  $s = \langle e_1(st_1, et_1)e_2(st_2, et_2)\dots e_n(st_n, et_n) \rangle$ , where  $e_i$  was item set,  $e_i = \{i_1, \dots, i_k\}$  and  $st_i$  was the starting time of item set, and  $et_i$  was the closing time of item set. Therefore, clinical behavior sequence data had the following features:

(1) Each event had a starting and closing time attributes as a constraint, and there were sequence and parallel relations among events.

(2) Behaviors were continuous.

As the clinical behaviors were constrained by time, therefore, continuous behaviors with the starting and closing time imposed legal constraints on subsequences. For example, as for subsequence  $s_1 = \langle ABD \rangle$  of  $s = \langle A\{B,C\}\{B,D\} \rangle$ , B and D were concurrent in S, namely  $et(B) < st(D)$ , but  $s_1$  represented the relationship  $et(B) < st(D)$  which was a group of sequence behaviors, and partly distorted in the real sense. Consequently,  $s_1$  was a subsequence but not a legal one that conformed to practical clinical behavior characteristics. Whether a sequence was legal or not depended on the presence of intersection parts between two item sets, and the legal subsequence among item sets strictly complied with  $et(e_i) \leq st(e_{i+1})$ .

If a sequence was not a valid sequence, and thus association rules deduced from it were all invalid. For example, as for previous-described illegal subsequence  $s_1 = \langle ABD \rangle$ , we could deduct rules  $R_1: \langle AB \rangle \rightarrow \langle D \rangle$  and  $R_2: \langle A \rangle \rightarrow \langle BD \rangle$  from it; they were all legal subsequences before or after  $R_1$ , and thus  $R_1$  was a legal rule, but  $R_2$  was not. This was because that the resultant illegal items were assigned to the both sides of rule, which eliminated all factors producing illegal subsequences and guaranteed that sequences at the both sides of rule were legal subsequences. Accordingly, sequences that had merely one illegal part could produce legal rules.

### 3.2 Correlated definitions

According to the characteristics and form of clinical behavior sequence data, in order to describe the right association rules for clinical behavior sequences, the definitions were supposed as follows:

Definition 1: given a sequence s, the number containing s affairs in data set was called the support of s sequence, and denoted as support (s).

Definition 2:  $s, s'$  represented two sequences in the data set, if their support was larger than the given threshold,

and then the confidence was defined as follows:

$$confidence(s, s') = \frac{support(s, s')}{support(s)}$$

If  $confidence(s, s')$  was also larger than a given value,  $s \rightarrow s'$  would form a sequence association rules.

Definition 3 (5): If there was an integer  $i_1 < i_2 < \dots < i_n$  that could make  $a_1 \subseteq b_{i_1}, a_2 \subseteq b_{i_2}, \dots, a_n \subseteq b_{i_n}$  without continuous item sets of  $a_j$  and  $a_{j+1}$ ,  $u \in a_{j+1}$  and  $v \notin a_{j+1}$ , and  $\{u, v\} \subseteq b_{i_j}$  or

$\{u, v\} \subseteq b_{j+1}$ , the sequence  $s' = \langle a_1 a_2 \dots a_m \rangle$  was the legal subsequence of  $s = \langle b_1 b_2 \dots b_n \rangle$ , and also called legal support  $s'$ , denoted as  $s' \propto s$ .

Definition 4: The number of affairs that sequence was supported in the data set was called the legal support number of  $s$ .

Definition 5: As for  $s \in D, s' \subseteq s$ , if  $s' \propto s$  and  $s - s' \propto s$ , the rule  $s' \rightarrow s - s'$  was legal.

### 3.3 Sequence association rule discovery algorithm of clinical behaviors

#### 3.3.1 The generation of frequent sequence set

As sequence association rule was sought, we should take the temporal relations among clinical behaviors, and modify GSP algorithm. Therefore, we proposed a CSAR-GSP algorithm (Clinical Sequence Association Rules by GSP).

A sequence could be supported simultaneously by legal and illegal sequences. For example, as for sequences  $s' = \langle \{A, C\} \{B\} \rangle$ ,  $s_1 = \langle \{A, C\} \{B, C, D\} \{E\} \rangle$  and  $s_2 = \langle \{A, C\} \{B, D\} E \rangle$ ,  $s'$  was illegally supported by  $s_1$ , and legally by  $s_2$ . If the number of  $s'$  supported by  $s_1$  was larger than by  $s_2$ ,  $s$  was illegal, and vice versa.

A sequence after extension could result in the following four situations:

- ① the extension of legal sequence resulted in legal one;
- ② the extension of legal sequence resulted in illegal one;
- ③ the extension of illegal sequence resulted in illegal one;
- ④ the extension of illegal sequence resulted in legal one;

Newly added items should be linked at the most end during the sequence extension process of GSP algorithm, and therefore, they only affected the last two item sets. We merely took the last two items into account. Suppose sequence  $s = \langle e_1 \dots e_{k-1} e_k \rangle$ , new item  $a_i$ , aimed at the above-mentioned four situations, we did as the following manners:

(1) As for ① and ④, due to the consequence of legal sequences, they would not be considered.

(2) If a sequence should have been legal, namely that  $s$  was legal. If it was illegal after extension, newly added items should be responsible for that. Record the location.

(3) If a sequence should have been illegal, namely that  $s$  was illegal,  $et(e_k) \leq st(e_{k-1})$  should be responsible for that. If newly added item  $a_i$  had sequence extension, namely  $s' = \langle e_1 \dots e_{k-1} e_k \{a_i\} \rangle$ , the original location was still illegal. If  $st(a_i) \geq et(e_k)$ , newly added item was legal, and vice versa; If item set extension, namely  $s' = \langle e_1 \dots \{e_k \cup a_i\} \rangle$ , and  $et(e_{k-1}) \leq st(\{e_k \cup a_i\})$  at the right time, illegality was eliminated, or illegal location continuously existed, and new location was only recorded.

As the presence of illegal subsequence could result in illegal rule, it should be marked. We previously analyzed that the generation of rule could eliminate one illegal part, and one illegal sequence could be likely to become legal after extension. As for over two illegal parts in one sequence, if extension, it wouldn't eliminate illegal part. Accordingly, for one item that caused over two illegal parts in a sequence, it was not added to frequent sequence set and deleted directly. Aimed at the previous discussions, we recorded two illegal locations respectively using two flag bits, and the location resulting in illegal items using positive number. 0 represented no illegal locations.

Frequent sequence generation algorithm was as follows:

Input: data set, support threshold

Output: frequent sequence set L

- (1)  $L_1 = \{\text{large 1-sequence}\}$
- (2) For ( $k=2; L_{k-1} \neq \emptyset; k++$ ) do begin
- (3)  $C_k = \text{GSP-generate}(L_{k-1})$  // linked using GSP method, return after pruning
- (4) For each of candidates  $l_k$  in  $C_k$
- (5) Calculate support, delete sequences that failed to conform to the support threshold, and estimate whether  $l_k$  was legal support sequence or not.

- (6) If ( $l_k$  was illegal sequence) then {
- (7)  $e_m, e_{m-1}$  were the last two item sets of  $l_k$
- (8) If ( $l_k.flag2=0 \wedge l_k.flag1=k-1$ ) then
- (9) {Estimate whether it was sequence extension or not. If yes, and previous illegal location couldn't be deleted, estimate whether new location was legal or not again. If new location was illegal, then  $l_k.flag2=k$ , otherwise, estimate the newly added locations, renewedly not or eliminate illegal parts. }
- (10) Else if ( $l_k.flag1=0$ ) then
- (11)  $l_k.flag1=k$
- (12) Else if ( $l_k.flag1>0 \&\& l_k.flag1 \neq k-1 \&\& l_k.flag2=0$ ) then {
- (13) Estimate whether  $l_k$  had sequence extension or not. If only one illegal part existed not in the last end of  $l_{k-1}$ , then  $l_k.flag2=k$ ; }
- (14) Else if ( $l_k.flag>0 \&\& l_k.flag2=k-1$ ) then// two items were all illegal, the second was at the end of  $l_{k-1}$
- (15){Estimate whether it had sequence extension or not. If yes, and it couldn't be legal, delete  $l_k$ , otherwise it had item set extension. If  $st(e_m) \leq et(e_{m-1})$ , then  $l_k.flag2=k$ , otherwise, eliminate the second illegal part  $l_k.flag2=0$ . }
- (16)  $L_k = \bigcup_k l_k$
- (17)endfor
- (18) endfor
- (19) Return  $L_k$

### 3.3.2 The generation of sequence association rules

As for association rules resulted from frequent sequences, the following points should be taken into account:

- (1) Time sequence of sequences on both sides of rules;
- (2) Legality of sequences on both sides of rules;
- (3) The generation of legal and illegal sequence rules

We elucidated the above-mentioned three points using the case of the rule generation of sequence  $S = \langle (A, C) (B, C, D) \rangle$ .

- (1) For time sequence of sequences on both sides of rules, we should always guarantee that the closing time of antecedent behavior was earlier than the starting time of middleware and consequent behaviors.

For this point, we could take use of this strategy to generate association rules. Firstly, we moved the first item of sequence  $S$  to the antecedent of rules. After that, each generation moved the first item of rule consequent to the last item of rule antecedent. As items in item set were intercurrent and items in the same item set couldn't be classified into different time sequences, the permutations and combinations should be placed on both sides of rules. For example, sequence  $S$  could result in several rules as follows:  $R_1: \langle \{A, C\} B \rangle \rightarrow \langle \{C, D\} E \rangle$ ,  $R_2: \langle \{A, C\} D \rangle \rightarrow \langle \{B, C\} E \rangle$  and so on.

- (2) The legality of sequences on both sides of rules

For example, as above-mentioned, the antecedents of rule  $R_1$  and  $R_2$  were all illegal subsequences of sequence  $S$ , and thus these two rules shouldn't be established and should be deleted. Rule  $R_3: \langle \{A, C\} \rangle \rightarrow \langle \{B, C, D\} E \rangle$  was legal.

- (3) The generation of illegal sequence rules

As long as the illegal items were assigned to both sides of rules, sequences on both sides would be legal. Firstly, we found the item set where illegal sequence existed, took the latter sequence of each item in item set as the back part of rules, and other sequences as the front part of rules. Meanwhile, we should consider the legality of sequences on both sides.

### 3.3 Anomaly detection

Anomaly detection was based on the assumption that the abnormal behavior was different from normal one to

some degree. After obtaining the clinical behavior model by means of association rule mining method, anomaly detection was operated in the clinical behavior model using model comparison methods. The specific point was as follows: first establish behavior model set  $R_1$  under the normal model, and mine behavior model set  $R_2$  from the detected data through calculating the  $similarity(R_1, R_2)$  between two behavior models (Wang, 2000), namely that similarity was applied to show the deviate degree of the current behavior from normal behavior in order to determine whether the detected data was abnormal or not. Similarity ranged from 0 to 1. Higher value showed that the larger matching degree between two compared behavior model sets. 1 denoted total matching, while 0 complete misfit.

#### 4. Results and discussions

We attempted to check the recognition capacity of clinical behavior anomaly detection on abnormal behavior. Data set used in the experiment was a data sample of acute simple appendicitis from 2000 to 2006 sampled from a hospital. Aimed at the goal "clinical cure, treatment integrity", after the previous processing of data, 1200 different cases were sampled from each group at random, and were marked as  $D$ ,  $D_{normal}$  and  $D_{abnormal}$ , respectively. Among them,  $D$  was training set,  $D_{normal}$  (without abnormal data) and  $D_{abnormal}$  (with abnormal data) were test sets.

Using CSAR-GSP algorithm to mine the rule set of training set  $D$ , the uniform confidence threshold was 80%, and listed in Table 1.

To compare the similarity of association rule between abnormal and normal behaviors, test sets  $D_{normal}$  and  $D_{abnormal}$  were processed with association rule mining using the same support threshold, and the corresponding rule sets  $R_{normal}$  and  $R_{abnormal}$  were obtained, respectively. The similarity of the two rule sets between the above-mentioned training set  $D$  or rule set  $R$  was calculated. In order to observe the effects of support on similarity, the similarity of rule sets at different support threshold was assayed, and the results were depicted in Figure 2.

As seen from Figure 2, the mining rule sets from clinical behaviors under normal conditions were different from abnormal conditions, and meanwhile, the difference in similarity was associated with support threshold.

Experiments indicated that the application of data mining technology in the association rule mining of clinical behaviors could estimate whether clinical behaviors was abnormal or not by setting the similarity threshold of rule sets.

#### 5. Conclusions

In the present paper, we elucidated the application of association rule mining in the anomaly detection of clinical behaviors. Due to the huge clinical database, the diagnosis and treatment process of each case was different, and the standardization level of clinical behaviors and detections was not high, which imposed great difficulties to detections. In the present paper, we selected partial data, and could detect clinical abnormal behaviors by means of association rule mining method. Certainly, as an anomaly detection method, it was based on limited statistical data and expert subjective experience, and still had necessary room for modification. For example, modification of anomaly detection algorithm capability and the determination of support and similarity threshold all need further investigation.

#### References

- Bruno, G., Garza, P., & Quintarelli, E., et al. (2007). Anomaly Detection in XML databases by means of Association Rules. *Database and Expert Systems Applications (DEXA), 2007 18th International Conference*.
- Khelifa, B., & Aomar O. (2007). Mining association rules in temporal sequences. *Proceedings of the 2007 IEEE Symposium on Computational Intelligence and Data Mining (CIDM 2007)*.
- Sanchez, D., Vilam, A., & Cerda, L., et al. (2009). Association rules applied to credit card fraud detection. *Expert Systems with Applications: An International Journal*, 8(36):3630-3640.
- Wang, J., & Yang, H.B. (2008). *Studies on the application of sequence mining in clinical behavior model discovery*. Jiangsu: Jiangsu University Press.
- Wang, W.D., & Bridges, S.M. (2000). Genetic algorithm optimization of membership functions for mining fuzzy association rules. *International Joint Conference on Information Systems, Fuzzy Theory and Technology Conference, Atlantic City, N.J.* March 2, 2000.
- Yang, Z.Y., Kong, L.B., & Yang, Z., et al. (2004). Studies on the establishment of diagnosis and treatment

behavior model. *Chinese Journal of Behavior Medical Science*, 13(6):601-603.

Table 1. Rule set of D

Rule set	Rule
	R1 :< 1>→<2> (s=85.32%, c=83.5%)
R	R2 :<{ 1, 3}>→<4> (s=68.33%, c=92.37%)
	...

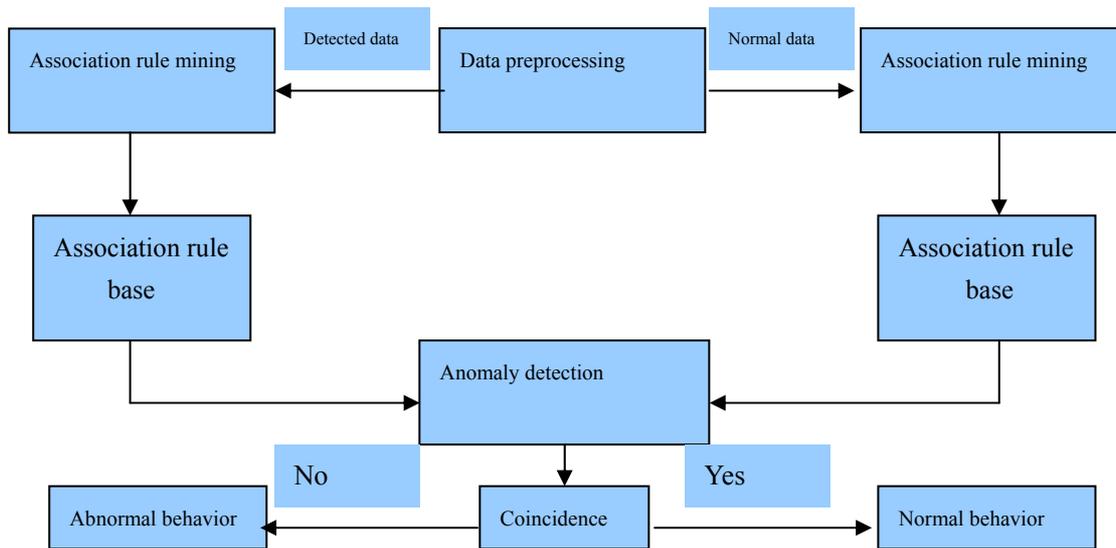


Figure 1. Anomaly detection model of clinical behaviors

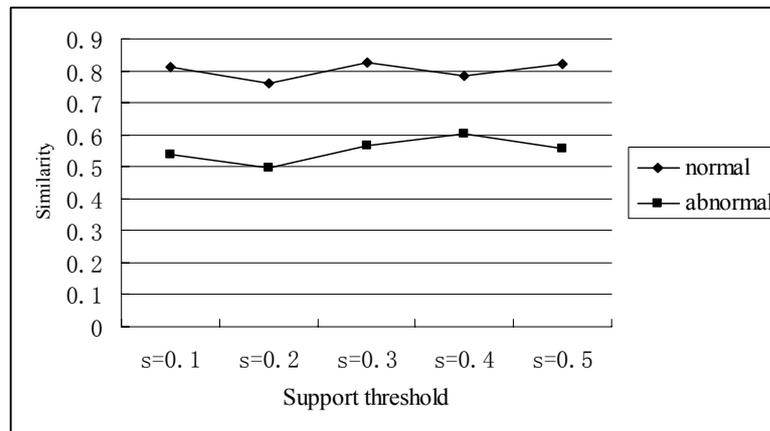


Figure 2. Comparison diagram of rule set similarity