Bi-level Rare Temporal Pattern Detection

Dawei Zhou, Jingrui He, Yu Cao, Jae-sun Seo
Arizona State University
Tempe, AZ 85281
Email: {dzhou23, jingrui.he, ycao, jaesun.seo}@asu.edu

***Abstract*—**

Nowadays, temporal data is generated at an unprecedented speed from a variety of applications, such as wearable devices, sensor networks, wireless networks and etc. In contrast to such large amount of temporal data, it is usually the case that only a small portion of them contains information of interest. For example, for the ECG signals collected by wearable devices, most of them collected from healthy people are normal, and only a small number of them collected from people with certain heart diseases are abnormal. Furthermore, even for the abnormal temporal sequences, the abnormal patterns may only be present in a few time segments and are similar among themselves, forming a rare category of temporal patterns. For example, the ECG signal collected from an individual with a certain heart disease may be normal in most time segments, and abnormal in only a few time segments, exhibiting similar patterns. What is even more challenging is that such rare temporal patterns are often non-separable from the normal ones. Existing works on outlier detection for temporal data focus on detecting either the abnormal sequences as a whole, or the abnormal time segments directly, ignoring the relationship between abnormal sequences and abnormal time segments. Moreover, the abnormal patterns are typically treated as isolated outliers instead of a rare category with self-similarity.

In this paper, for the first time, we propose a bi-level (sequence-level/ segment-level) model for rare temporal pattern detection. It is based on an optimization framework that fully exploits the bi-level structure in the data, i.e., the relationship between abnormal sequences and abnormal time segments. Furthermore, it uses sequence-specific simple hidden Markov models to obtain segment-level labels, and leverages the similarity among abnormal time segments to estimate the model parameters. To solve the optimization framework, we propose the unsupervised algorithm *BIRAD*, and also the semi-supervised version *BIRAD-K* which learns from a single labeled example. Experimental results on both synthetic and real data sets demonstrate the performance of the proposed algorithms from multiple aspects, outperforming state-of-the-art techniques on both temporal outlier detection and rare category analysis.

***Keywords*-rare category detection; temporal data mining; time series; time segments;**

I. Introduction

In the era of big data, we are exposed to large amount of temporal data, such as biomedical signals [19], financial transaction records [13], and network traffic [20]. Besides the large volume of data, we are also facing the following challenges: (1) the class membership is often highly skewed in the sense that the minority classes (rare temporal patterns) are overwhelmed by the majority classes (normal temporal patterns); (2) it is usually the case that identifying the minority classes is more important than identifying the majority classes in the temporal data; (3) the minority classes are often non-separable from the majority classes. For example, most of the ECG signals collected by wearable devices are normal, generated by healthy people, and only a small number of them are abnormal, generated by people with certain heart diseases such as arrhythmia. Without domain specific knowledge, it can be very difficult to distinguish between abnormal ECG signals and normal ones. In malicious insider identification, the daily activities of most employees are normal, and only a small number of employees are malicious insiders with abnormal activities. Since these guileful insiders usually try to camouflage as normal employees, these abnormal activities may be very similar to the normal ones. Furthermore, within the abnormal temporal sequences, there may only be a few time segments exhibiting similar abnormal patterns, forming a rare category of temporal patterns. For instance, the ECG signal of an individual with arrhythmia may only show irregular heartbeats in a few time segments; the malicious insiders may behave abnormally every now and then. Fig. 1 illustrates such bi-level structure of the temporal data, where abnormal sequences contain at least one abnormal time segment, and normal sequences only contain normal time segments. In this paper, we aim to detect abnormal sequences and abnormal segments simultaneously, which correspond to the bi-level rare temporal pattern detection.

To the best of our knowledge, such bi-level structure (sequence level vs. segment level) is not exploited in existing works on outlier detection for temporal data, which focus on either the sequence level, or the segment level. Furthermore, they fail to explore the similarity among the abnormal time segments, treating them as isolated outliers. On the other hand, existing works on rare category analysis are mainly focused on static data, which are not readily applicable to temporal data with rare categories of abnormal patterns.

To bridge this gap, in this paper, we study the problem of rare temporal pattern detection by exploiting the bi-level structure in the data. Our proposed model is based on an optimization framework that maximizes the likelihood of observing the data on both the sequence level and the

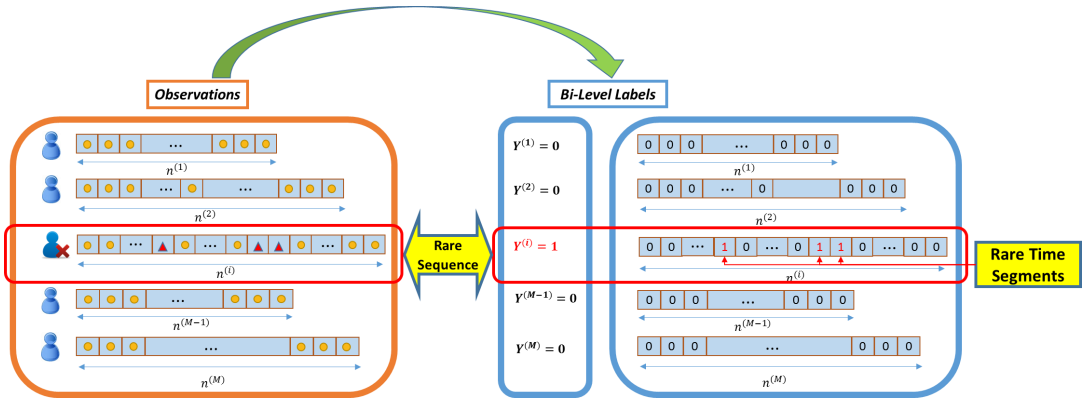


Figure 1: Illustration of the Bi-Level Structure in the Temporal Data.

segment level. Furthermore, it uses sequence-specific simple hidden Markov models to generate segment-level labels, and leverages the similarity among the abnormal time segments to estimate the model parameters. To solve the optimization problem, we propose an unsupervised algorithm for detecting rare temporal patterns named *BIRAD* and its semi-supervised version named *BIRAD-K*. Both algorithms are based on Block Coordinate Update, which repeatedly update the sequence-level labels, segment-level labels, and the model parameters. We analyze these algorithms in terms of convergence and time complexity, and empirically evaluate their performance on both synthetic and real data sets.

The rest of this paper is organized as follows. After a brief review of the related work in Section 2, we introduce the bi-level model, the optimization framework, and the proposed algorithms with performance analysis in Section 3. In Section 4, we present the experimental results on both synthetic and real data sets, which demonstrate the effectiveness and efficiency of the proposed framework. Finally, we conclude this paper in Section 5.

II. Related Work

In this section, we briefly review the related work on rare category analysis, outlier detection for temporal data, and multi-instance learning.

A. Rare Category Analysis

Rare category detection is the problem of identifying minority classes from the under-represented feature spaces, while minimizing the number of labeling requests. Up until now, several techniques have been developed for rare category detection in different scenarios. [22] introduced the problem setting of rare category detection and experimented with different hint selection strategies to detect useful anomalies. In [10, 11], the authors presented two active learning schemes to detect rare categories via unsupervised local-density-differential sampling strategy. More recently, in [31], the authors studied the problem of rare category detection on multi-view data and proposed a Bayesian framework named MUVIR, which exploited the relationship between multiple views and estimated the overall probability of each example belonging to the minority class. In [32], the authors proposed a fast method for rare category detection on time-evolving graphs, which incrementally updated the detection models based on local updates. In this paper, we further study the problem of rare category detection on temporal data and aim to exploit the bi-level structure of abnormal temporal sequences / time segments.

B. Outlier Detection for Temporal Data

Outlier detection, also called anomaly detection or novelty detection, refers to the problem of finding instances that do not conform to the expected behavior in the data. This problem has been studied in various domains, such as heterogenous networks [5, 20, 15], crowdsourcing [30, 33] and spatiotemporal channels [25, 27]. Prior works mainly focused on two categories of temporal outliers: outliers in time series databases and outliers within the given time series [9]. For the first category of outliers, the previous methods aim to identify a few time series as outliers, such as clustering methods [21], parametric methods [4], window-based methods [8]. For the second category of outliers, the methods aim to find particular elements or subsequences on the given time series. For example, in [12], the authors presented an autoregressive data-driven model to identify outliers in environmental data streams; in [3], the authors studied a more challenging problem that outlier detection faced with a never-ending data stream. Different from existing works on outlier detection for temporal data, our work focuses on the more challenging case where the abnormal temporal patterns are non-separable from the normal ones, and we propose to leverage the relationship between abnormal temporal sequences and abnormal time segments for the sake of improving the detection accuracy.

C. Multi-Instance Learning

Multi-instance learning is a variation of supervised learning, where examples are considered as bags consisting of

multiple individual instances. [7] is the earliest literature that introduced and showed the importance of multi-instance learning. In the past decades, various techniques were proposed targeting multi-instance learning. In [17, 29], diverse density based frameworks are proposed for solving the multi-instance learning problem, by measuring the intersection of the positive bags minus the union of the negative bags. [1] presented an extended version of support vector machine on multi-instance learning, and developed a heuristic method to solve the mixed integer quadratic programs. [34] is the first study on the problem of multi-instance learning under the condition that examples are not independent and identically distributed (i.i.d) by constructing an undirected graph of each bag and designing a graph kernel to classify the positive and negative examples. Similar to multi-instance learning, in our model, the segment-level labels collectively determine the corresponding sequence-level label. However, here we assume that the relationship among adjacent time segments is governed by segment-specific simple hidden Markov models, and many existing works on multi-instance learning can be seen as special cases of our proposed model.

III. Bi-Level Model for Rare Temporal Pattern Detection

In this section, we propose a bi-level model for detecting the rare temporal patterns. We start with notation and problem definition. Then we present the model formulation, followed by the optimization techniques. Finally, we introduce both the unsupervised algorithm *BIRAD* for detecting the rare temporal patterns and its semi-supervised version *BIRAD-K*.

A. Notation and Problem Definition

Suppose that we are given a set of M temporal sequences $S = \{\boldsymbol{x^{(1)}}, \ldots, \boldsymbol{x^{(M)}}\}$, and, in temporal sequence $\boldsymbol{x^{(m)}}$ where $m = 1, \ldots, M$, there are $n^{(m)}$ temporal segments, i.e., $\boldsymbol{x^{(m)}} = \{x_1^{(m)}, \ldots, x_{n^{(m)}}^{(m)}\}$. Let $\boldsymbol{y^{(m)}} = \{y_1^{(m)}, \ldots, y_{n^{(m)}}^{(m)}\} \in \{0,1\}^{1\times n^{(m)}}$ denote the segment-level labels, or hidden states of temporal segments in $\boldsymbol{x^{(m)}}$, and $Y^{(m)} \in \{0,1\}$ denote the sequence-level label, or hidden state of $\boldsymbol{x^{(m)}}$. Without loss of generality, we assume that: (1) $y_i^{(m)} = 1$ corresponds to abnormal segments, and $y_i^{(m)} = 0$ corresponds to normal segments; (2) $Y^{(m)} = 1$ corresponds to abnormal temporal sequences, and $Y^{(m)} = 0$ corresponds to normal sequences. As the bi-level structure illustrated in Fig. 1, only a small portion of temporal sequences in S correspond to abnormal sequences, in which only a small portion of temporal segments are abnormal segments. Therefore, the abnormal segments would be extremely rare when considering the whole data set S. Our goal is to identify anomalies in the sequence level as well as the segment level. For the sake of clarity, we also introduce the following indicator function $I(\boldsymbol{y^{(m)}}) = \max_{i=1}^{n^{(m)}} y_i^{(m)}$.

B. Model Formulation

Our model lies in inference about the bi-level hidden state process given the observations S, which involves calculating the following posterior distribution.

$$\begin{aligned} Pr(\boldsymbol{y^{(m)}}|\boldsymbol{x^{(m)}}) &\propto Pr(\boldsymbol{y^{(m)}}, \boldsymbol{x^{(m)}}) \\ &= Pr(\boldsymbol{x^{(m)}}) Pr(\boldsymbol{y^{(m)}}|\boldsymbol{x^{(m)}}) \end{aligned} \tag{1}$$

Thus, we propose the objective of our model as follows.

$$\underset{\boldsymbol{y^{(1:M)}}}{\operatorname{argmax}} \sum_{m=1}^{M} \ln Pr(\boldsymbol{y^{(m)}}, \boldsymbol{x^{(m)}}) \tag{2}$$

As the data could be categorized into normal and abnormal temporal sequences, we can rewrite Eq. 2 as follows.

$$\begin{aligned} &\underset{\boldsymbol{y^{(1:M)}}}{\operatorname{argmax}} \sum_{Y^{(m)}=1} \ln Pr(\boldsymbol{y^{(m)}}, \boldsymbol{x^{(m)}}) \\ &+ \sum_{Y^{(m)}=0} \ln Pr(\boldsymbol{y^{(m)}}, \boldsymbol{x^{(m)}}) \end{aligned} \tag{3}$$

By introducing sequence-level label $Y^{(m)}$ to the preceding equation, we have

$$\begin{aligned} &\underset{\boldsymbol{y^{(1:M)}}}{\operatorname{argmax}} \sum_{m=1}^{M} \ln[Pr(\boldsymbol{y^{(m)}}, \boldsymbol{x^{(m)}}|Y^{(m)} = 1) \\ &\times Pr(Y^{(m)} = 1)]^{Y^{(m)}} + \ln[Pr(\boldsymbol{y^{(m)}}, \boldsymbol{x^{(m)}}|Y^{(m)} = 0) \\ &\times Pr(Y^{(m)} = 0)]^{(1-Y^{(m)})} \\ &\quad \text{s.t.} \quad Y^{(m)} = \max_i y_i^{(m)} \\ &\qquad\quad m = 1, \ldots, M, i = 1, \ldots, n^{(t)} \end{aligned} \tag{4}$$

Let L_0 denote $\ln Pr(\boldsymbol{y^{(m)}}, \boldsymbol{x^{(m)}}|Y^{(m)} = 0)$, and L_1 denote $\ln Pr(\boldsymbol{y^{(m)}}, \boldsymbol{x^{(m)}}|Y^{(m)} = 1)$. We can rewrite Eq. 4 as follows.

$$\begin{aligned} &\underset{\boldsymbol{y^{(1:M)}}}{\operatorname{argmax}} \sum_{m=1}^{M} (1 - Y^{(m)})[L_0(\boldsymbol{x^{(m)}}) + \ln Pr(Y^{(m)} = 0)] \\ &+ Y^{(m)}[L_1(\boldsymbol{x^{(m)}}) + \ln Pr(Y^{(m)} = 1)] \\ &\quad \text{s.t.} \quad Y^{(m)} = \max_i y_i^{(m)} \\ &\qquad\quad m = 1, \ldots, M, i = 1, \ldots, n^{(t)} \end{aligned} \tag{5}$$

In order to model the joint probability of the segment-level labels $\boldsymbol{y}^{(m)}$ and the temporal data $\boldsymbol{x}^{(m)}$, we propose to use simple Hidden Markov Models (HMM) [2]. In particular, we have the following three assumptions: (1) the Markov assumption, i.e., the next state is dependent only upon the current state, where the state corresponds to the segment-level label $y_i^{(m)}$; (2) the stationarity assumption, i.e., state transition probabilities are independent of the actual time at which the transitions take place; (3) the output independence assumption, i.e., current output (observation) is statistically independent of the previous outputs (observations). Next we elaborate on modeling normal and abnormal temporal

sequences.

Modeling Normal Temporal Sequences: For the sake of exposition, we first model the normal temporal sequence, i.e., $Y^{(m)} = 0$. The log likelihood of any normal temporal sequence $\boldsymbol{x^{(m)}}$ is defined by

$$\begin{aligned} L_0 &= \ln Pr(\boldsymbol{y^{(m)}}, \boldsymbol{x^{(m)}} | Y^{(m)} = 0) \\ &= \ln[Pr(\boldsymbol{x^{(m)}} | \boldsymbol{y^{(m)}}, Y^{(m)} = 0) \times Pr(\boldsymbol{y^{(m)}} | Y^{(m)} = 0)] \end{aligned} \tag{6}$$

Based on the Markov assumption and output independence assumption, we have

$$\begin{aligned} Pr(\boldsymbol{y^{(m)}} | Y^{(m)} = 0)] = Pr(y_1^{(m)} | Y^{(m)} = 0) \\ \times \prod_{j=2}^{n^{(m)}} Pr(y_j^{(m)} | y_{j-1}^{(m)}, Y^{(m)} = 0) \end{aligned} \tag{7}$$

By applying the stationarity assumption, we have

$$Pr(\boldsymbol{x^{(m)}} | \boldsymbol{y^{(m)}}, Y^{(m)} = 0) = \prod_{i=1}^{n^{(m)}} Pr(x_i^{(m)} | y_i^{(m)}, Y^{(m)} = 0) \tag{8}$$

Plugging Eq. 7 and Eq. 8 into Eq. 6, we have

$$\begin{aligned} L_0 = \ln[\prod_{i=1}^{n^{(m)}} Pr(x_i^{(m)} | y_i^{(m)}, Y^{(m)} = 0) \\ \times Pr(y_1^{(m)} | Y^{(m)} = 0) \times \prod_{j=2}^{n^{(m)}} Pr(y_j^{(m)} | y_{j-1}^{(m)}, Y^{(m)} = 0)] \end{aligned} \tag{9}$$

On the other hand, we assume that any normal temporal segment $x_i^{(m)}$ is drawn from an unknown Gaussian distribution, although the proposed model can be generalized to other parametric distributions:

$$Pr(x_i^{(m)} | y_i^{(m)}, Y^{(m)} = 0) \sim \mathcal{N}(x_i^{(m)}, \mu_0, \sigma_0)$$

where mean μ_0 and variance σ_0 are not given. Hence, $\mathcal{N}(x_i^{(m)}, \mu_0, \sigma_0)$ could be interpreted as the emission probability of $x_i^{(m)}$ given hidden state 0.

Then, Eq. 9 can be rewritten as follows.

$$\begin{aligned} L_0 = \sum_{i=1}^{n^{(m)}} \ln \mathcal{N}(x_i^{(m)}, \mu_0, \sigma_0) + \ln Pr(y_1^{(m)} | Y^{(m)} = 0) \\ + \sum_{j=2}^{n^{(m)}} \ln Pr(y_j^{(m)} | y_{j-1}^{(m)}, Y^{(m)} = 0) \end{aligned} \tag{10}$$

For any normal temporal sequences $\boldsymbol{x^{(m)}}$, there is no segment-level state transition, i.e., all temporal segments are normal. Therefore, L_0 could be simplified as follows.

$$L_0 = \sum_{i=1}^{n^{(m)}} \ln \mathcal{N}(x_i^{(m)}, \mu_0, \sigma_0) \tag{11}$$

Modeling Abnormal Temporal Sequences: As we mentioned before, if temporal sequence $\boldsymbol{x^{(m)}}$ contains at least one abnormal segment $x_i^{(m)}$ with $y_i^{(m)} = 1$, then we claim that temporal sequence $\boldsymbol{x^{(m)}}$ is abnormal, i.e., $Y^{(m)} = 1$. Similar as before, we have the following log likelihood.

$$L_1 = \ln Pr(\boldsymbol{y^{(m)}}, \boldsymbol{x^{(m)}} | Y^{(m)} = 1) \tag{12}$$

In our model, we assume that the features from abnormal time segments are generated from the same compact distribution across all abnormal temporal sequences. Similar to Eq. 9, by taking advantage of the HMM assumptions and Bayes' Rule, we can rewrite Eq. 12 as follows.

$$\begin{aligned} L_1 = \ln[\prod_{(i=1)}^{n^{(m)}} Pr(x_i^{(m)} | y_i^{(m)}, Y^{(m)} = 1) \\ \times Pr(y_1^{(m)} | Y^{(m)} = 1) \times \prod_{j=2}^{n^{(m)}} Pr(y_j^{(m)} | y_{j-1}^{(m)}, Y^{(m)} = 1)] \end{aligned} \tag{13}$$

For any abnormal temporal sequence $\boldsymbol{x}^{(m)}$, we define the corresponding Hidden Markov Model [2] $\lambda = (N, O, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi})$ as follows.

1) N, the number of hidden states in the model. In this paper, $N = 2$, i.e., normal and abnormal states.
2) O, the number of distinct observations. In our model, for any temporal sequence $\boldsymbol{x^{(m)}}$, the number of observations is the length of $\boldsymbol{x^{(m)}}$.
3) $\boldsymbol{A}$, $N \times N$, the state transition probability distribution. $\boldsymbol{A}$ is an $N \times N$ matrix. In this paper, we have:

$$\boldsymbol{A} = \begin{bmatrix} a_{00} & a_{01} \\ a_{10} & a_{11} \end{bmatrix}$$

   where a_{ij} denotes the transition probability from state i to state j. And we have $a_{ij} \in [0, 1]$ and $a_{i0} + a_{i1} = 1$, $i \in \{0, 1\}$.
4) $\boldsymbol{B}$, the observation emission probability distribution, which is an $N \times O$ matrix. We assume that normal time segments meet distribution $\mathcal{N}(\mu_0, \sigma_0)$, while abnormal time segments meet distribution $\mathcal{N}(\mu_1, \sigma_1)$.
5) $\boldsymbol{\pi}$, the initial state probability distribution, of which the length is N. In our model, for any temporal sequence $\boldsymbol{x^{(m)}}$, we define $a_0^{(m)}$ as the probability that the initial temporal segment $x_1^{(m)}$ is abnormal. Then, we can write the initial state probability distribution of sequence $\boldsymbol{x^{(m)}}$ as $[1 - a_0^{(m)}, a_0^{(m)}]$.

Based on the HMM model $\lambda = (N, O, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi})$, Eq. 13

can be rewritten as follows.

$$
\begin{aligned}
L_1 = & \sum_{i=1}^{n^{(m)}} [y_i^{(m)} \ln \mathcal{N}(x_i^{(m)}, \mu_1, \sigma_1) \\
& + (1 - y_i^{(m)}) \ln \mathcal{N}(x_i^{(m)}, \mu_0, \sigma_0)] + [y_1^{(m)} \ln a_0 \\
& + (1 - y_1^{(m)}) \ln(1 - a_0)] + \sum_{j=2}^{n^{(m)}} [y_{j-1}^{(m)} y_j^{(m)} \ln a_{11} \\
& + y_{j-1}^{(m)}(1 - y_j^{(m)}) \ln(1 - a_{11}) + (1 - y_{j-1}^{(m)}) y_j^{(m)} \ln a_{01} \\
& + (1 - y_{j-1}^{(m)})(1 - y_j^{(m)}) \ln(1 - a_{01})]
\end{aligned} \tag{14}
$$

Overall Objective Function: Plugging Eq. 11 and Eq. 14 into the objective function in Eq. 5, we have

$$
\begin{aligned}
& \underset{\substack{\boldsymbol{y^{(1:M)}}, a_0, a_{11}, \\ \mu_1, \sigma_1, \mu_0, \sigma_0}}{\operatorname{argmax}} \sum_{m=1}^{M} Y^{(m)} \{ \sum_{i=1}^{n^{(m)}} [y_i^{(m)} \ln \mathcal{N}(x_i^{(m)}, \mu_1, \sigma_1) \\
& + (1 - y_i^{(m)}) \ln \mathcal{N}(x_i^{(m)}, \mu_0, \sigma_0)] + [y_1^{(m)} \ln a_0 \\
& + (1 - y_1^{(m)}) \ln(1 - a_0)] + \sum_{j=2}^{n^{(m)}} [y_{j-1}^{(m)} y_j^{(m)} \ln a_{11} \\
& + y_{j-1}^{(m)}(1 - y_j^{(m)}) \ln(1 - a_{11}) + (1 - y_{j-1}^{(m)}) y_j^{(m)} \ln a_{01} \\
& + (1 - y_{j-1}^{(m)})(1 - y_j^{(m)}) \ln(1 - a_{01})] + \ln Pr(Y^{(m)} = 1)\} \\
& + (1 - Y^{(m)}) \{ \sum_{i=1}^{n^{(m)}} \ln \mathcal{N}(x_i^{(m)}, \mu_0, \sigma_0) + \ln Pr(Y^{(m)} = 0)\} \\
& \quad \text{s.t.} \quad a_0, a_{11}, a_{01} \in [0, 1] \\
& \qquad\quad Y^{(m)} = \max_i y_i^{(m)} \\
& \qquad\quad m = 1, \ldots, M, i = 1, \ldots, n^{(t)}
\end{aligned} \tag{15}
$$

C. Optimization

Given any finite observation sequence, it is challenging to maximize the posterior probability by adjusting the HMM model parameters $(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi})$. In fact, there is not a practical method to exactly solve this problem. However, a number of iterative procedures, such as EM based methods [14] and gradient based methods [14], have been proposed to obtain a local maximum of this problem. In the following two subsections, we will introduce two simple and fast algorithms, i.e., *BIRAD* and *BIRAD-K*, targeting the novel setting of bi-level rare temporal pattern detection. Both of these two algorithms are built upon Block Coordinate Update (BCU) method [16, 26, 28], which divides all the variables into multiple blocks and iteratively updates them. To be specific,

Updating Initial State Probability Distribution: By taking the partial derivative of Eq. 15 with respect to a_0, and letting it equal to zero, we have the following closed form update rule.

$$
a_0^{(m)} = y_1^{(m)} \tag{16}
$$

Updating State Transition Probability Distribution: By taking the partial derivative of Eq. 15 with respect to a_{11} and a_{01}, and letting them equal to zero, we have the following closed form update rules.

$$
a_{11} = \frac{\sum_{m=1}^{M} \sum_{j=2}^{n^{(m)}} Y^{(m)} y_{j-1}^{(m)} y_j^{(m)}}{\sum_{m=1}^{M} \sum_{j=2}^{n^{(m)}} Y^{(m)} y_{j-1}^{(m)}} \tag{17}
$$

$$
a_{01} = \frac{\sum_{m=1}^{M} \sum_{j=2}^{n^{(m)}} Y^{(m)} (1 - y_{j-1}^{(m)}) y_j^{(m)}}{\sum_{m=1}^{M} \sum_{j=2}^{n^{(m)}} Y^{(m)} (1 - y_{j-1}^{(m)})} \tag{18}
$$

Updating Observation Emission Probability Distribution: By taking the partial derivation of Eq 15 with respect to μ_1, σ_1, μ_0, σ_0, and letting them equal to zero, we have the following closed form update rules.

$$
\mu_1 = \frac{\sum_{m=1}^{M} \sum_{t=1}^{n^{(m)}} y_t^{(m)} x_t^{(m)}}{\sum_{m=1}^{M} \sum_{t=1}^{n^{(m)}} y_t^{(m)}} \tag{19}
$$

$$
\sigma_1 = \frac{\sum_{m=1}^{M} \sum_{t=1}^{n^{(m)}} y_t^{(m)} |x_t^{(m)} - \mu_1|_2^2}{\sum_{m=1}^{M} \sum_{t=1}^{n^{(m)}} y_t^{(m)} - 1} \tag{20}
$$

$$
\mu_0 = \frac{\sum_{m=1}^{M} \sum_{t=1}^{n^{(m)}} (1 - y_t^{(m)}) x_t^{(m)}}{\sum_{m=1}^{M} \sum_{t=1}^{n^{(m)}} (1 - y_t^{(m)})} \tag{21}
$$

$$
\sigma_0 = \frac{\sum_{m=1}^{M} \sum_{t=1}^{n^{(m)}} (1 - y_t^{(m)}) |x_t^{(m)} - \mu_0|_2^2}{\sum_{m=1}^{M} \sum_{t=1}^{n^{(m)}} (1 - y_t^{(m)}) - 1} \tag{22}
$$

Updating Bi-level Labels: In this part, we give an easy and fast update strategy for updating bi-level labels. For updating the sequence-level labels, we first score each temporal sequence by comparing the log likelihood of the sequence being labeled as abnormal vs. normal in each iteration. Then, the sequences with higher scores would be labeled as abnormal and the rests will be labeled as normal. The details will be illustrated in *BIRAD* and *BIRAD-K*. For updating the segment-level labels, there are the following two cases. When the sequence-level label $Y^{(m)} = 0$, we can directly label each segment in $\boldsymbol{y}^{(m)}$ as $0_{1 \times n^{(m)}}$. When the sequence-level label $Y^{(m)} = 1$, we apply Viterbi algorithm [23] to iteratively update the most likely hidden states, or segment-level labels, $\boldsymbol{y^{(m)}}$, which maximizes the objective function in Eq. 15.

D. BIRAD Algorithm

Based on the update rules introduced in the previous subsection, we first introduce the unsupervised method — Bi-level Rare Temporal Anomaly Detection (*BIRAD*) algorithm. It is given an unlabeled temporal sequence data set S and the proportion of abnormal sequences P as inputs. It outputs the hidden states of all temporal sequences and temporal segments in S. The algorithm iteratively updates the HMM parameters $\lambda = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi})$ and the bi-level hidden states until convergence, or a certain stopping criterion is satisfied. The details of *BIRAD* are presented in Algorithm 1.

Algorithm 1 Bi-level Rare Temporal Anomaly Detection (*BIRAD*)

Input:
Temporal sequence data set $\boldsymbol{x^{(1)}}, \ldots, \boldsymbol{x^{(M)}}$
Proportion of abnormal sequences P.

Output:
$Y^{(1)}, \ldots, Y^{(M)}$; $\boldsymbol{y^{(1)}}, \ldots, \boldsymbol{y^{(M)}}$

1: Initialize sequence-level and segment-level labels.
2: **while** stopping criterion is not satisfied **do**
3: Update initial state probability distribution $\boldsymbol{\pi}$ by Eq. 16.
4: Update transition probability distribution $\boldsymbol{A}$ by Eq. 17 to Eq. 18.
5: Update emission probability distribution $\boldsymbol{B}$ by Eq. 19 to Eq.22.
6: **for** m =1: M **do**
7: Update hidden states $\boldsymbol{y}^{(m)}$ of $\boldsymbol{x^{(m)}}$ by Viterbi Algorithm.
8: Compute $L_1(\boldsymbol{x}^{(m)})$ in Eq. 14 based on updated $\boldsymbol{y^{(m)}}$.
9: Compute $L_0(\boldsymbol{x}^{(m)})$ in Eq.11 based on updated $\boldsymbol{y^{(m)}}$.
10: Compute $score^{(m)} = L_1(\boldsymbol{x}^{(m)}) + \ln P - L_0(\boldsymbol{x}^{(m)}) - \ln(1-P)$
11: **end for**
12: Label the temporal sequences with positive scores as abnormal, i.e., $Y^{(m)} = 1$, and keep the updated prediction labels $\boldsymbol{y}^{(m)}$. Label the remaining temporal sequences as normal, i.e., $Y^{(m)} = 0$, and label the segments in these sequences as normal, i.e., $\boldsymbol{y}^{(m)} = 0_{1\times n^{(m)}}$.
13: **for** $m = 1 : M$ **do**
14: **if** $I(\boldsymbol{y^{(m)}}) \neq Y^{(m)}$ **then**
15: Let $Y^{(m)} = 0$ and $\boldsymbol{y^{(m)}} = 0_{1\times n^{(m)}}$.
16: **end if**
17: **end for**
18: **end while**
19: **return** $Y^{(1)}, \ldots, Y^{(M)}$; $\boldsymbol{y^{(1)}}, \ldots, \boldsymbol{y^{(M)}}$.

BIRAD works as follows. First, Step 1 initializes the sequence-level/ segment-level labels. Specifically, one potential way to initialize the bi-level hidden states is to randomly select $M \times P$ temporal sequences and label them as 1, while the rest are labeled as 0. Then, we can initialize any hidden states of temporal segments to be identical as the hidden state of the corresponding temporal sequence. Next, Step 2 to Step 18 applies the BCU optimization process. From Step 3 to Step 5, *BIRAD* updates the initial probability vector $\boldsymbol{\pi}$, transition probability distribution $\boldsymbol{A}$ and emission probability distribution $\boldsymbol{B}$ based on the updated labels $\boldsymbol{y^{(1)}}, \ldots, \boldsymbol{y^{(M)}}$. In Step 7 to Step 10, *BIRAD* updates the segment-level hidden states of $\boldsymbol{x^{(m)}}$ and calculates the scores for each temporal sequence $\boldsymbol{x^{(m)}}$, which estimate the probability of a sequence being abnormal rather than normal. Step 12 updates the sequence-level/ segment-level labels based on $score^{(m)}$. Step 13 to Step 17 checks if there is any inconsistency between $\boldsymbol{y^{(m)}}$ and $Y^{(m)}$. If any inconsistency exists, these temporal sequences are labeled as normal. At last, in Step 19, *BIRAD* returns the predicted bi-level labels.

Next, we analyze the convergence of the proposed *BIRAD* algorithm. We first derive Lemma 1 and Lemma 2, which show that the update rules in Algorithm 1 are upper-bounded and non-decreasing. Lemma 1 and Lemma 2 lead to Theorem 1, which shows the convergence of *BIRAD*.

Lemma 1 (Upper-bounded)**.** *The overall objective function in Eq. 15 is upper-bounded.*

Proof Sketch: Due to properties of the parametric distributions of normal and abnormal time segments, as well as the transition probabilities, it is easy to see that Eq. 15 is upper-bounded. ■

Lemma 2 (Non-decreasing)**.** *The objective function in Eq. 15 is non-decreasing in general under the update rules in Algorithm 1.*

Proof: By separately taking second-order derivatives of Eq. 15 with respect to the variables of initial probability $\boldsymbol{\pi}$, transition probability distribution $\boldsymbol{A}$ and emission probability distribution $\boldsymbol{B}$, it is easy to see that the three Hessian matrices we obtain are negative semi-definite. Thus, when all but one block are fixed, Eq. 15 is a concave function with respect to the free block. In other words, the overall objective function Eq. 15 is non-decreasing when we only update the blocks of the initial probability, the transition probability and the emission probability.

The same conclusion could also be reached when we update the segment-level labels with other blocks fixed, as the Viterbi algorithm always returns the optimal labels $\boldsymbol{y}^{(m)}$ for any input sequence $\boldsymbol{x}^{(m)}$. On the sequence-level, the *BIRAD* algorithm firstly scores each temporal sequence by comparing the log likelihood of the sequence being labeled as abnormal vs. normal. Then all the temporal sequences with positive scores are labeled $Y^{(m)} = 1$, and the ones with negative scores are labeled $Y^{(m)} = 0$. At last, *BIRAD* algorithm corrects the inconsistency between sequence-level and segment-level labels for the following two cases: (1)$Y^{(m)} = 1$ and $\boldsymbol{y^{(m)}} = 0^{1\times n^{(m)}}$; (2) $Y^{(m)} = 0$ and $\boldsymbol{y^{(m)}}$ contains at least one segment-level label as 1. For case 1, it is easy to see Eq. 15 increases by $\ln Pr(Y^{(m)} = 0) - \ln Pr(Y^{(m)} = 1)$ after correction of $Y^{(m)}$, where $Pr(Y^{(m)} = 0) \gg Pr(Y^{(m)} = 1)$. For case 2, the overall objective function in Eq. 15 keeps the same value after correction of $\boldsymbol{y^{(m)}}$. In this way, the objective function value with the resulting sequence-level and the associated segment-level labels is no smaller than any alternative label

assignments. Therefore, the objective function in Eq. 15 is non-decreasing under the update rules of Algorithm 1. ■

Theorem 1 (Local Optimum)**.** *The proposed* BIRAD *algorithm converges to the local optimal.*

Proof: According to Lemma 1 and Lemma 2, the objective function is non-decreasing and upper-bounded based on the update rules in Algorithm 1. Therefore, the proposed *BIRAD* algorithm converges to a local optimal. ■

We also analyze the computational complexity of the *BIRAD* algorithm in the following theorem.

Theorem 2 (Time Complexity)**.** *The time complexity of Algorithm 1 (with Viterbi algorithm) is* $O(LMO)$.

Proof: Let L be the required number of iterations for Algorithm 1 to converge. The time complexity of Viterbi Algorithm is $O(N^2O)$, where N is the number of hidden states, and O is the length of a given temporal sequence. In each iteration of Algorithm 1, we call Viterbi Algorithm M times. Thus, we have the time complexity of Algorithm 1 as $O(LMO)$. ■

E. BIRAD-K Algorithm

In some cases, we may be able to start with a few labeled examples, i.e., labeled segments. To accommodate these cases, we introduce a modified semi-supervised version of Algorithm 1 named *BIRAD-K* in Algorithm 2.

To be specific, *BIRAD-K* is given a temporal sequence data set S with only one labeled abnormal segment $X_{AG}^{(AQ)}$, where AQ is the sequence-level index and AG is the segment-level index of $X_{AG}^{(AQ)}$, and prior P as input. Compared with *BIRAD*, *BIRAD-K* works better with noisy data, e.g., data with outliers or changing points. The details of *BIRAD-K* are described in Algorithm 2. Step 1 initializes the bi-level hidden states. Step 2 calculates K, which is the number of abnormal temporal sequences in the data set. Step 3 to Step 9 is the BCU process. Identical to *BIRAD*, we first update the initial probability vector $\boldsymbol{\pi}$, transition probability distribution $\boldsymbol{A}$ and emission probability distribution $\boldsymbol{B}$ based on the updated labels from the last iteration. Next, we calculate the scores for identifying abnormal temporal sequences in Step 5. Different from *BIRAD* in Step 6, we label the temporal sequences with the top K scores as abnormal and the rest as normal. Step 7 ensures $Y^{(AQ)}$ and $y_{AG}^{(AQ)}$ are always labeled as 1. Step 8 checks if there is any inconsistency between sequence-level labels and segment-level labels. Finally, in Step 10, *BIRAD-K* returns all the consistent prediction labels upon convergence.

Algorithm 2 Bi-level Rare Temporal Anomaly Detection with K Segments Selected (*BIRAD-K*)

Input:
Temporal sequence data set $\boldsymbol{x^{(1)}}, \ldots, \boldsymbol{x^{(M)}}$ with only one labeled abnormal segment $x_{AG}^{(AQ)}$
Proportion of abnormal sequences P.
Output:
$Y^{(1)}, \ldots, Y^{(M)}$; $\boldsymbol{y^{(1)}}, \ldots, \boldsymbol{y^{(M)}}$

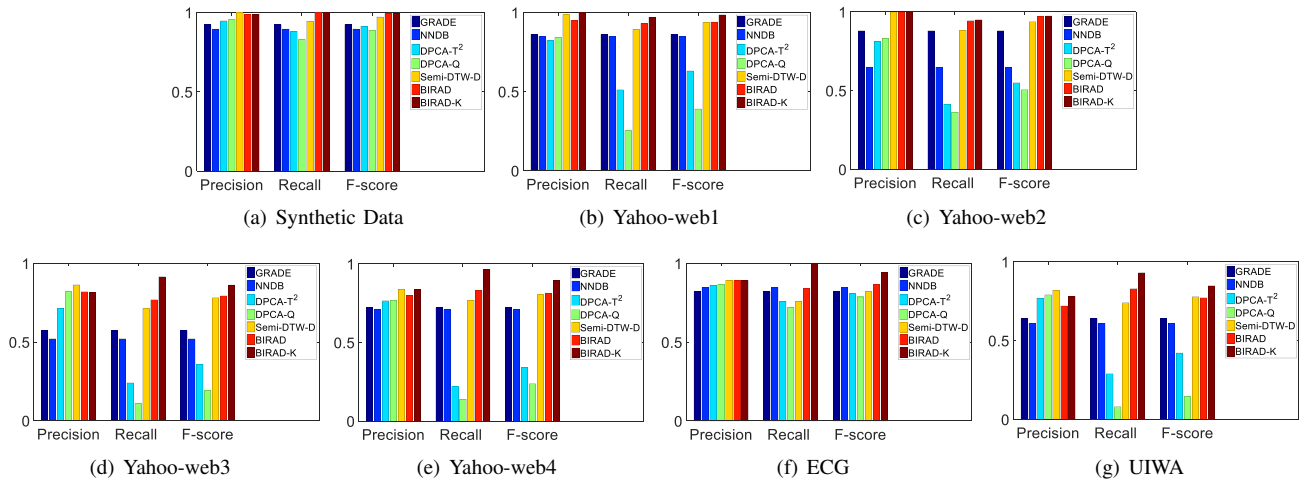
1: Initialize sequence-level and segment-level labels.
2: Compute $K = m \times P$
3: **while** stopping criterion is not satisfied **do**
4: Update HMM model $\lambda = (\boldsymbol{\pi}, A, B)$ as Step 3 to Step 5 in Algorithm 1.
5: Update the segment-level hidden states and anomaly score $score^{(m)}$ for each temporal sequence $\boldsymbol{x^{(m)}}$ as Step 6 to Step 11 in Algorithm 1.
6: Label the temporal sequences with the top K scores as abnormal, i.e., $Y^{(m)} = 1$, and keep the updated prediction labels $y^{(m)}$. Label the remaining temporal sequences as normal, i.e., $Y^{(m)} = 0$, and label the segments in these temporal sequences as normal, i.e., $\boldsymbol{y^{(m)}} = 0_{1 \times n^{(m)}}$.
7: Correct $Y^{(AQ)}$ or $y_{AG}^{(AQ)}$, if either of them are updated as 0.
8: Check and fix the inconsistency between sequence-level labels and segment-level labels as Step 13 to Step 17 in Algorithm 1.
9: **end while**
10: **return** $Y^{(1)}, \ldots, Y^{(M)}$; $\boldsymbol{y^{(1)}}, \ldots, \boldsymbol{y^{(M)}}$.

IV. Experiments

In this section, we evaluate the performance of our proposed algorithms, i.e., *BIRAD* and *BIRAD-K*, on both synthetic and real data sets in comparison with four state-of-the-art unsupervised methods, i.e., NNDB [10], GRADE [11], DPCA-T^2 [24], DPCA-Q [24], and one semi-supervised method, i.e., Semi-DTW-D [6]. The RCD methods, i.e., NNDB and GRADE, require the exact proportion of abnormal time segments in the entire data set. This is the reason why the RCD algorithms produce the same precision and recall rate in the results shown in Fig. 2. For the two PCA methods, the principal components are associated with 95% of the total variance explanation. Semi-DTW-D is a semi-supervised learning method for time series classification. In the comparison experiments, *BIRAD-K* and Semi-DTW-D are given a single labeled abnormal segment as training data.

A. Data Set Description

The synthetic data set is generated from auto-regression model with 3 different coefficients $C1$, $C2$ and $C3$. It contains 95 normal temporal sequences and 5 abnormal sequences, and each temporal sequence consists of 1,000 observations. In normal sequences, all data points fit the model with coefficients $C1$. In abnormal sequences, there are 980 normal data points that fit the model with coefficients $C2$, and 20 abnormal data points that fit the model with coefficients $C3$.



(a) Synthetic Data (b) Yahoo-web1 (c) Yahoo-web2

(d) Yahoo-web3 (e) Yahoo-web4 (f) ECG (g) UIWA

Figure 2: Performance Evaluation on Real Data

In our experiments, we include 4 temporal data sets from Yahoo! Webscope program [1]. Each data set contains around 80 temporal sequences, and each sequence contains around 1,500 observations. The first data set contains regular anomaly points. The second and third data sets contain periodic outliers. The third and fourth data sets include anomaly points as well as changing points. To match the scenario of our studying problem, each data set is modified as containing 95% synthetic normal sequences and 5% abnormal sequences.

ECG data set is a collection of 100 ECG signal records, which is extracted from a public ECG database [2]. Each record consists of $\sim$ 300 segments, where each segment corresponds to one certain heart beat pulse. In this data set, 10% signal records are abnormal temporal sequences. Meanwhile, there are around 2% abnormal segments in these abnormal sequences. Our goal is to detect noisy and unstable heart beat pulses, which may be produced due to movements or changes of the environment conditions.

At last, ADL data set [18] comprises information regarding the sensor logs of users' daily activities during a 35-day interval. The data set is labeled with 10 different daily behaviors, i.e., "$Leaving''$, "$Toileting''$, "$Showering''$, "$Sleeping''$, "$Breakfast''$, "$Lunch''$, "$Dinner''$, "$Snack''$, "$Spare-Time''$, "$Grooming''$. In this experiment, we consider "$Snack''$ as the abnormal behavior, which only comprises around 5% of data, and the rest as the normal behaviors. In the end, we aim to identify all the time intervals of "Snack" for each user.

B. Effectiveness Analysis

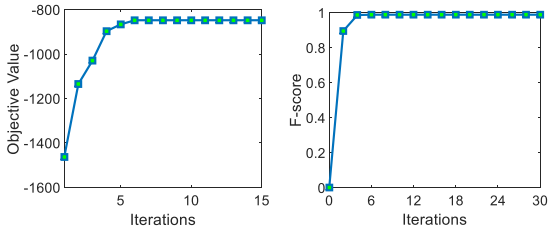
In this subsection, we evaluate the effectiveness of *BIRAD* and *BIRAD-K* over 1 synthetic data set and 6 real data sets based on precision, recall and F-score (defined as F-score = $2 \cdot$ Recall $\cdot$ Precision / (Recall + Precision)). Notice that, in these experiments, we are able to identify all the abnormal temporal sequences, and the following results are respect to $\boldsymbol{y}^{(m)}$, $m = 1, \ldots, M$.

First, the proposed algorithms are evaluated on the synthetic data set and 4 Yahoo-web data sets, all of which are temporal data sets with anomalies. From Fig. 2(a) to Fig. 2(e), we can discover the significant advantages of our proposed methods. The PCA methods always produce very low recall rate, which indicates that the PCA methods may not be suitable for capturing anomalies in the subspace with maximized variance. For NNDB and GRADE, they are very stable for both precision and recall rates, but perform unsatisfied when facing more complex conditions, such as changing points. In Fig. 2(d), both NNDB and GRADE achieve very low precision and recall rates. This is because they are built upon static methods, thus not effective in handling the temporal variations. Compared with *BIRAD* and *BIRAD-K*, we find that Semi-DTW-D always achieves good precision scores, while the recall rates are lower. This is because Semi-DTW-D is designed for time series classification, which only meaures the distance between temporal segments, but has not considered the hidden state transition between the adjacent temporal segments. It can be seen that our proposed methods always outperform the other methods, especially in the sense of recall rate and F-score rate. Comparing *BIRAD* and *BIRAD-K*, it is shown that *BIRAD-K* performs slightly better than *BIRAD*, especially in Fig. 2(d) and Fig. 2(e). This implies that *BIRAD-K* algorithm may be more suited for applications with outliers or changing points.

Next, two challenging real world problems are considered for anomaly detection. In Fig. 2(f), we study the problem of anomaly pattern detection on ECG signals. It reveals that

[1] https://webscope.sandbox.yahoo.com/catalog.php?datatype=s&did=70

[2] https://www.physionet.org/physiobank/database/ptbdb/



(a) Objective Value over Iterations (b) F-score over Iterations

Figure 3: Convergence Analysis

all the methods perform very well on this data set, however our *BIRAD* and *BIRAD-K* algorithms still outperform the others. In addition, in Fig. 2(g), we apply our algorithms on wireless sensor networks data so as to detect all the abnormal behaviors. Due to the unremovable randomness in human's daily behaviors, this problem is more challenging than the previous 5 data sets. In this experiment, lack of the ability to extract temporal information is the main reason why GRADE and NNDB get much lower precision than the others. Compared with *BIRAD* and *BIRAD-K*, the PCA methods and Semi-DTW-D get a lower recall rates because it may not be able to precisely catch the rules of state transition, especially in the occurrence of randomness. In general, we have the following observations about our proposed algorithms from these 6 experiments: (1) Both *BIRAD* and *BIRAD-K* outperform our 3 baseline algorithms in most cases; (2) *BIRAD* produces comparable results as *BIRAD-K* in most cases; (3) *BIRAD-K* performs modestly better than *BIRAD* especially in the presence of outliers and changing points.

C. Convergence and Efficiency Analysis

In this subsection, we first examine the convergence of *BIRAD* algorithm on the synthetic data set. Fig. 3(a) illustrates the non-decreasing and upper-bounded characteristics of the objective function when applying *BIRAD*. In Fig. 3(b), we present the changes of F-score among different iterations. It is shown that the F-score monotonically increases with objective values and then saturates, implying that the performance improves with increasing objective values.

Then, we examine the running time and parameter sensitivity of *BIRAD* and *BIRAD-K* algorithms. First, we perform our algorithms on a series of synthetic data sets with increasing number of temporal sequences. Let the prior be 5% and the length of each temporal sequence be 1,000, we generate a series of synthetic data sets with increasing number of temporal sequences, from 100 to 1,000. The results are shown in Fig. 4. After that, we test our algorithms on a series of data sets with increasing sequence length. Different from the experiments in Fig. 4, we let each data set contain 100 temporal sequences and the prior be 5%, and we generate the series of synthetic data sets with increasing sequence length, from 500 to 5,000. The results are shown in Fig. 5. From the preceding two experiments, we have the following observations: (1) *BIRAD* is slightly faster than *BIRAD-K*; (2) the running time of both algorithms increases linearly in general for both cases, i.e., increasing the sequence length and increasing the number of temporal sequences. we run the experiments with Matlab 2014a on a workstation with four 3.5 GHz CPUs, 256 GB memory and 2 TB disk space.

D. Parameter Analysis

In this subsection, we empirically study the parameter sensitivity of *BIRAD* and *BIRAD-K* algorithms on the synthetic data set. Fig. 6 shows our analysis results. Notice that the exact proportion of abnormal temporal sequences is 5% in the data set. For *BIRAD-K* algorithm, we can see the F-score increases sharply as the prior changes from 1% to 5%. This is because *BIRAD-K* discovers more abnormal sequences with the increase of input prior ($P < 5\%$). As the prior goes beyond 5%, the F-score of *BIRAD-K* slightly diminishes but stabilizes near 0.89. The reason is that several normal temporal sequences are included in the group of abnormal sequences, as the input prior exceeds the exact prior. Thus, the input prior would introduce a bias especially when we update the transition probability distribution A and emission probability distribution B. Different from the previous case, the experiments show that the precision rate reduces slightly and the recall rate is kept stable when the input prior ($P > 5\%$) increases. Compared with *BIRAD-K*, we can see the F-score rates of *BIRAD* are more stable. This implies that *BIRAD* is more reliable than *BIRAD-K* in the cases with unprecise priors.

V. Conclusion and Future Work

In this paper, we introduce a novel data mining problem - bi-level rare temporal pattern detection, which aims to fill the gap in the literature by conducting rare category analysis on temporal data. Specifically, we address the challenging case where the labels of the temporal data are highly skewed on both the sequence-level and the segment-level. We formulate the problem as an optimization problem, which maximizes the likelihood of observing the data on both the sequence-level and the segment-level. To solve the optimization problem, we propose an unsupervised algorithm *BIRAD* and its semi-supervised version *BIRAD-K*, which iteratively update the model parameters based on the block coordinate update method and return the bi-level labels that are consistent on the sequence-level and the segment-level. The comparison experiments with state-of-the-art techniques demonstrate the effectiveness of our proposed algorithms. In our future work, we will extend the proposed framework to the cases when multiple types of rare temporal patterns exist such that the number of hidden states $N > 2$.

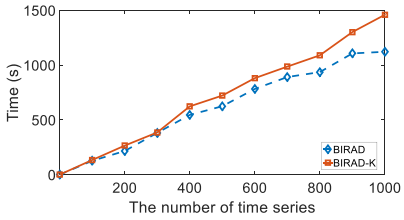


Figure 4: Efficiency Analysis on Increasing Number of Temporal Sequences

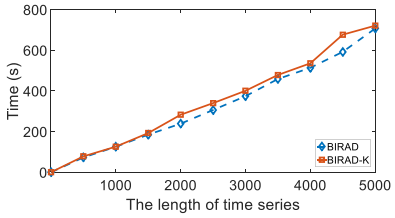


Figure 5: Efficiency Analysis on Increasing Length of Temporal Sequences

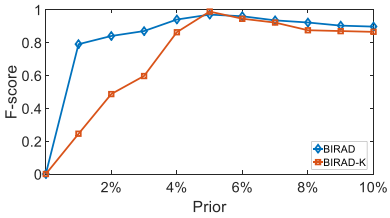


Figure 6: Parameter Analysis

ACKNOWLEDGMENT

This work is supported by NSF research grant IIS-1552654, an IBM Faculty Award, and a research grant by Samsung Advanced Institute of Technology. The views and conclusions are those of the authors and should not be interpreted as representing the official policies of the funding agencies or the U.S. Government.

REFERENCES

[1] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *Advances in neural information processing systems*.

[2] L. E. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state markov chains. *The annals of mathematical statistics*, 1966.

[3] N. Begum and E. Keogh. Rare time series motif discovery from unbounded streams. *VLDB*, 2014.

[4] V. Chandola, V. Mithal, and V. Kumar. Comparative evaluation of anomaly detection techniques for sequence data. In *ICDM*, 2008.

[5] L. Chen, J. Warner, P. L. Yung, D. Zhou, W. Heinzelman, I. Demirkol, U. Muncuk, K. Chowdhury, and S. Basagni. Reach 2-mote: A range-extending passive wake-up wireless sensor node. *ACM Transactions on Sensor Networks (TOSN)*, 2015.

[6] Y. Chen, B. Hu, E. Keogh, and G. E. Batista. Dtw-d: time series semi-supervised learning from a single example. In *SIGKDD*, 2013.

[7] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 1997.

[8] F. A. González and D. Dasgupta. Anomaly detection using real-valued negative selection. *Genetic Programming and Evolvable Machines*, 2003.

[9] M. Gupta, J. Gao, C. Aggarwal, and J. Han. Outlier detection for temporal data. *Synthesis Lectures on Data Mining and Knowledge Discovery*, 2014.

[10] J. He and J. G. Carbonell. Nearest-neighbor-based active learning for rare category detection. In *NIPS*, 2007.

[11] J. He, Y. Liu, and R. Lawrence. Graph-based rare category detection. In *ICDM*, 2008.

[12] D. J. Hill and B. S. Minsker. Anomaly detection in streaming environmental sensor data: A data-driven modeling approach. *Environmental Modelling and Software*, 2010.

[13] D. L. Holloway and A. Anderson. Online payment system for merchants, June 16 2006. US Patent App. 11/922,346.

[14] W. Khreich, E. Granger, A. Miri, and R. Sabourin. A survey of techniques for incremental learning of hmm parameters. *Information Sciences*, 2012.

[15] J. Li, X. Hu, J. Tang, and H. Liu. Unsupervised streaming feature selection in social media. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1041–1050, 2015.

[16] Z.-Q. Luo and P. Tseng. On the convergence of the coordinate descent method for convex differentiable minimization. *Journal of Optimization Theory and Applications*, 1992.

[17] O. Maron and T. Lozano-Pérez. A framework for multiple-instance learning. *Advances in neural information processing systems*, 1998.

[18] F. J. Ordónez, P. de Toledo, and A. Sanchis. Activity recognition using hybrid generative/discriminative models on home environments using binary sensors. *Sensors*, 2013.

[19] J. J. Oresko, Z. Jin, J. Cheng, S. Huang, Y. Sun, H. Duschl, and A. C. Cheng. A wearable smartphone-based platform for real-time cardiovascular disease detection via electrocardiogram processing. *Information Technology in Biomedicine, IEEE Transactions on*, 2010.

[20] S. Pan, Q. Ye, S. Liu, and D. Zhou. Joint resource allocation for wlan&wcdma integrated networks based on spectral bandwidth mapping. *Journal of Electronics (China)*, 2011.

[21] X. Pan, J. Tan, S. Kavulya, R. Gandhi, and P. Narasimhan. Ganesha: blackbox diagnosis of mapreduce systems. *ACM SIGMETRICS*, 2010.

[22] D. Pelleg and A. W. Moore. Active learning for anomaly and rare-category detection. In *NIPS*, 2004.

[23] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 1989.

[24] E. L. Russell, L. H. Chiang, and R. D. Braatz. *Data-driven methods for fault detection and diagnosis in chemical processes*. Springer Science & Business Media, 2012.

[25] K. Shu, P. Luo, W. Li, P. Yin, and L. Tang. Deal or deceit: detecting cheating in distribution channels. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 1419–1428, 2014.

[26] P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications*, 2001.

[27] Y. Wang, Q. Zhang, and B. Li. Efficient unsupervised abnormal crowd activity detection based on a spatiotemporal saliency detector. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9, 2016.

[28] Y. Xu and W. Yin. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM Journal on imaging sciences*, 2013.

[29] Q. Zhang and S. A. Goldman. Em-dd: An improved multiple-instance learning technique. In *Advances in neural information processing systems*, 2001.

[30] X. Zhang, G. Xue, R. Yu, D. Yang, and J. Tang. Truthful incentive mechanisms for crowdsourcing. In *2015 IEEE Conference on Computer Communications (INFOCOM)*, pages 2830–2838, 2015.

[31] D. Zhou, J. He, K. Candan, and H. Davulcu. MUVIR: multi-view rare category detection. In *IJCAI*, 2015.

[32] D. Zhou, K. Wang, N. Cao, and J. He. Rare category detection on time-evolving graphs. In *ICDM*, 2015.

[33] Y. Zhou and J. He. Crowdsourcing via tensor augmentation and completion. In *IJCAI*, 2016.

[34] Z.-H. Zhou, Y.-Y. Sun, and Y.-F. Li. Multi-instance learning by treating instances as non-iid samples. In *ICML*, 2009.