DISCRIMINATIVE RECURRING SIGNAL DETECTION AND LOCALIZATION

Zeyu You, Raviv Raich, Xiaoli Z. Fern and Jinsub Kim

School of EECS, Oregon State University, Corvallis, OR, 97331-5501, USA

youz@onid.orst.edu, {raich,xfern,kimjinsu}@eecs.oregonstate.edu

ABSTRACT

Recognizing and localizing a recurring pattern is a problem with a variety of applications such as classification and localization of home appliances from their activation signals and estimating the relative alignment between records of a natural repetitive electrocardiography (ECG) signals in Bio-medical data. Most common approaches for recognizing a recurring pattern are generative and focus on discovering and capturing the characteristics of the recurring pattern. One limitation of such generative approaches is that they are more sensitive to variations of the recurring signal. In this paper, we present a discriminative approach for finding a recurring pattern and localizing it within a collection of signals. We evaluate and compare our method to a generative approach [1] on both synthetic data and real-world home appliance data.

Index Terms— Discriminative pattern recognition, recurring signal detection, signal localization.

1. INTRODUCTION

Due to rapid data growth we are facing nowadays, the capability to recognize recurring patterns in data becomes increasingly important because it helps to find regularities in data and can be used for downstream data analysis tasks such as feature extraction and classification. A common goal in this context is to discover recurrent patterns from data without any prior knowledge of what the patterns might look like. Toward this goal, several approaches have been proposed recently, most of which focused on finding the fundamental characteristics of the signal pattern [2, 3, 4, 5] and are generative in nature. By contrast, limited work has considered a discriminative approach for this task. One important issue with generative approaches of discovering recurring pattern is that the detection performance significantly degrades with increased noise and variations of the recurring signal. To address this issue, we focus on the problem of finding a discriminative convolutional kernel of the unknown recurring pattern, such that the resulting signal will directly indicate the location of the pattern. The problem of discovering convolutional kernel of recurring unknown pattern has been less studied.

This work is partially supported by the National Science Foundation grants CCF-1254218, DBI-1356792 and IIS-1055113.

In this paper, we propose a probabilistic model and provide a systematic solution for detecting the recurring signal pattern. Our contributions are as follows: (i) We introduce a novel formulation of auto-detecting recurring signal patterns; (ii) We provide a maximum likelihood estimation (MLE) solution for finding the discriminative convolutional kernel; and (iii) We show an increased detection performance on a real-world home appliance data.

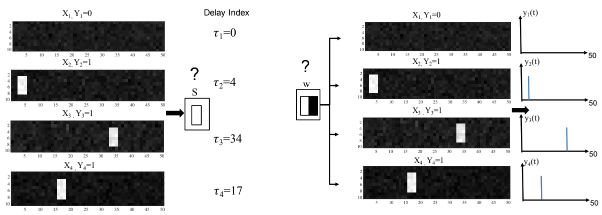
The rest of the paper is organized as follows. Section 2 provides the formulation of the discriminative recurring pattern recognition. Section 3 proposes a probabilistic model to the problem. Section 4 provides a maximum likelihood estimation approach for solving the aforementioned probabilistic model. Section 5 and 6 empirically examine the proposed approach and compare with the generative approach using both synthetic and real-world dataset. Section 7 concludes the paper.

2. PROBLEM FORMULATION

We are given a collection of signals and their labels $\{(\mathbf{x}_1, Y_1), (\mathbf{x}_2, Y_2) \ldots, (\mathbf{x}_M, Y_M)\}$, where $\mathbf{x}_m$ denotes the mth signal $\mathbf{x}_m(t)$ for $1 \leq t \leq T_m$ and $Y_m \in \{0, 1\}$ denotes the presence or absence of an arbitrarily time-delayed pattern within signal $\mathbf{x}_m$. Our goal is to develop a discriminative framework for training a detector based on the given training data that can detect presence or absence of an unknown recurring pattern in a test signal. In contrast to the classical time delay estimation, we do not assume that the patterns within different signals are identical or identical up to a scaling factor. A generative model for detecting a recurring pattern [1] aims at finding the pattern and its corresponding delay as shown in Fig. 1(a). A discriminative approach uses a convolution kernel to predict the presence and absence of that pattern as shown in Fig. 1(b). Unlike the generative approach, the discriminative kernel does not resemble the original shape of that recurring pattern, but transforms the original signal data into a new signal that matches up with the signal label. Here, we focus on the latter.

To predict the presence or absence of the common pattern, we consider a sliding window of size T_0 and treat the signal segment within each window as an instance. Specifically, we associate $\mathbf{x}_m(t)$, the mth signal at location t, with a corresponding sequence $y_{mt} \in \{0, 1\}$. The instance label

y_{mt} being equal to 1 indicates the presence of a pattern at location t in $\mathbf{x}_m$. The sequence of instance labels for $\mathbf{x}_m$, which we denote by $\mathbf{y}_m \triangleq [y_{m1}, \ldots, y_{mT_m}]$, directly determines the bag-level label Y_m. Specifically, if $\mathbf{y}_m$ contains any entry with value 1, then Y_m is 1; otherwise, Y_m is zero.



(a) Generative (b) Discriminative

Fig. 1: Problem formulation of generative and discriminative recurring pattern recognition

3. THE PROBABILISTIC MODEL

In developing our model, we focus on a special case of the problem in which a single observance of the pattern of interest is made in each signal. Consequently, we assume that although the signal the instance label sequence $\mathbf{y}_m$ are not observed, we have the information that $\mathbf{y}_m$ is either a vector with all zero entries or a vector with all zero entries except a single nonzero entry taking value 1. For completeness, we express the mth signal label Y_m in terms of the corresponding instance label sequence $\mathbf{y}_m$ as

$$Y_m = \begin{cases} 0, & \mathbf{y}_m = \mathbf{0} \\ 1, & \mathbf{y}_m \in \{\mathbf{e}_{m1}, \ldots, \mathbf{e}_{mT_m}\} \\ 2, & \text{otherwise}, \end{cases}$$

where T_m is the number of sliding window segments in $\mathbf{x}_m$, and $\mathbf{e}_{ml} = [0, \ldots, 0, 1, 0, \ldots, 0]^T \in \mathbb{R}^{T_m}$, $\forall l = 1, \ldots, T_m$ is the lth standard basis vector of R^{T_m}, *i.e.*, its lth entry is one, and all other entries are zeros. Note that $Y_m = 2$ is only used for ensuring a complete probabilistic characterization of the model. However, in our setting, it is never observed.

We employ a probabilistic model (shown in Fig. 2) with a logistic function to model the conditional distribution of an instance label y_{mt} given a realization of the corresponding sliding window segment $\mathbf{x}_{mt} = [\mathbf{x}_m(t), \mathbf{x}_m(t-1), \ldots, \mathbf{x}_m(t - T_0 + 1)]^T \in \mathbb{R}^{T_0}$ as the tth windowed instance and $\mathbf{w} = [\mathbf{w}(1), \ldots, \mathbf{w}(T_0)]^T \in \mathbb{R}^{T_0}$ as the kernel signal. Therefore, the probabilistic model for y_{mt} is given by:

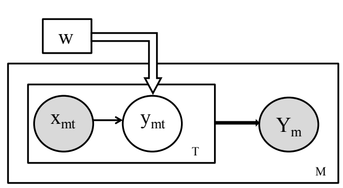


Fig. 2: The probabilistic graphical model

$$P(y_{mt}|\mathbf{x}_{mt}; \mathbf{w}) = \frac{e^{\mathbf{w}^T \mathbf{x}_{mt} y_{mt}}}{1 + e^{\mathbf{w}^T \mathbf{x}_{mt}}}. \quad (1)$$

Note that $\mathbf{w}^T \mathbf{x}_{mt}$ for all $t = 1, \ldots, T_m$ and $m = 1, \ldots, M$ are implemented as a convolution such that $\mathbf{w}^T \mathbf{x}_{mt} = \sum_{\tau=0}^{T_0} \mathbf{x}_m(t-\tau)\mathbf{w}(\tau)$.

To model the mth signal label Y_m given the instance labels $\mathbf{y}_m$, we consider two cases. When the signal label is positive $Y_m = 1$, only one out of T_m instance label can be one and the others are zeros. When the signal label is negative $Y_m = 0$, all of the T_m instances must be zeros. Therefore, the probabilistic model for the signal label Y_m given the instance labels $\mathbf{y}_m$ is:

$$P(Y_m|\mathbf{y}_m) = [\sum_{l=1}^{T_m} \mathbb{I}(\mathbf{y}_m = \mathbf{e}_{ml})]^{Y_m} [\mathbb{I}(\mathbf{y}_m = \mathbf{0})]^{1-Y_m}, \quad (2)$$

The probabilistic graphical model in Fig. 2 describes the conditional dependence structure of our model.

3.1. Extension to 2-D signals

When the data signal is 2-D such as spectrogram *i.e.*, $\mathbf{x}_m \in \mathbb{R}^{F \times T}$ for some frequency F, the probabilistic model in (1) can be smoothly adopted by setting the convolutive kernel to be 2-D as well, *i.e.*, $\mathbf{w} \in \mathbb{R}^{F \times T_0}$. In this case, $\mathbf{w}^T \mathbf{x}_{mt}$ is replaced with trace$(\mathbf{w}^T \mathbf{x}_{mt}) = \sum_{f=1}^{F} \sum_{\tau=0}^{T_0} \mathbf{x}_m(f, t-\tau)\mathbf{w}(f, \tau)$.

4. MAXIMUM LIKELIHOOD ESTIMATION

Given our proposed model, we consider estimating the model parameter $\mathbf{w}$ using maximum likelihood estimation (MLE).

4.1. Data Likelihood

Denote $\mathbf{D} = \{(\mathbf{x}_1, Y_1), (\mathbf{x}_2, Y_2) \ldots, (\mathbf{x}_M, Y_M)\}$ as the observed data and assume that $Y_m \in \{0, 1\}$, the data likelihood $L(\mathbf{w}) = P(\mathbf{D}; \mathbf{w})$, is obtained as

$$L(\mathbf{w}) = \prod_{m=1}^{M} \frac{(\sum_{l=1}^{T_m} e^{\mathbf{w}^T \mathbf{x}_{ml}})^{Y_m}}{\prod_{t=1}^{T_m}(1 + e^{\mathbf{w}^T \mathbf{x}_{mt}})} P(\mathbf{x}_m). \quad (3)$$

Therefore, the negative log-likelihood function is:

$$f(\mathbf{w}) = \sum_{m=1}^{M} [\sum_{t=1}^{T_m} \log(1 + e^{\mathbf{w}^T \mathbf{x}_{mt}}) - Y_m \log(\sum_{t=1}^{T_m} e^{\mathbf{w}^T \mathbf{x}_{mt}})] + C,$$

where $C = \sum_{m=1}^{M} \log(P(\mathbf{x}_m))$ is a constant. The challenge is this function is a combination of convex and concave function such that the non-convexity of the problem makes it harder to minimize.

4.2. Solution with CCCP

Since the objective is a convex-concave function, we apply the convex-concave procedure (CCCP) [6] to update $\mathbf{w}$. The general idea is to construct a majorizing function

$g(\mathbf{w}, \mathbf{w}^i)$ such that (i) $g(\mathbf{w}, \mathbf{w}^i) \geq f(\mathbf{w})$ for any $\mathbf{w}, \mathbf{w}^i$; and (ii) $g(\mathbf{w}, \mathbf{w}^i) = f(\mathbf{w})$ for $\mathbf{w} = \mathbf{w}^i$. Minimizing $g(\mathbf{w}, \mathbf{w}^i)$ function instead of $f(\mathbf{w})$ results in the following update rule $\mathbf{w}^{(i+1)} = \arg\min_{\mathbf{w}} g(\mathbf{w}, \mathbf{w}^i)$, which yields non increasing sequence of the objective, i.e., $f(\mathbf{w}^{(i+1)}) \leq f(\mathbf{w}^i)$.

A simple upper bound function $g(\mathbf{w}, \mathbf{w}^i)$ can be obtained by linearizing the convex function $v(\mathbf{w}) = \log(\sum_{t=1}^{T} e^{\mathbf{w}^T \mathbf{x}_{mt}})$. Since $v(\mathbf{w}) \geq v(\mathbf{w}^i) + (\mathbf{w} - \mathbf{w}^i)^T \Delta v(\mathbf{w}^i)$, then $f(\mathbf{w}) \leq g(\mathbf{w}, \mathbf{w}^i)$ [7]. Therefore, the upper bound $g(\mathbf{w}, \mathbf{w}^i)$ is:

$$g(\mathbf{w}, \mathbf{w}^i) = \sum_{m=1}^{M} [\sum_{t=1}^{T_m} \log(1 + e^{\mathbf{w}^T \mathbf{x}_{mt}}) - Y_m[\log(\sum_{t=1}^{T_m} e^{\mathbf{w}^{iT} \mathbf{x}_{mt}}) + \Big(\frac{\sum_{t=1}^{T_m} e^{\mathbf{w}^{iT} \mathbf{x}_{mt}} \mathbf{x}_{mt}}{\sum_{t=1}^{T_m} e^{\mathbf{w}^{iT} \mathbf{x}_{mt}}}\Big)^T (\mathbf{w} - \mathbf{w}^i)].$$

Using the gradient descent method, we obtain the update rule as follows:

$$\mathbf{w}^{i+1} = \mathbf{w}^i + \gamma \frac{\partial g(\mathbf{w}, \mathbf{w}^i)}{\partial \mathbf{w}} |_{\mathbf{w}=\mathbf{w}^i}, \text{where,} \quad (4)$$

$$\frac{\partial g(\mathbf{w}, \mathbf{w}^i)}{\partial \mathbf{w}} |_{\mathbf{w}=\mathbf{w}^i} = \sum_{m=1}^{M} \sum_{t=1}^{T_m} [P(y_{mt}) - Y_m P(y_{mt}|Y_m)] \mathbf{x}_{mt},$$

and γ is a learning rate. We refer to $P(y_{mt}) = P(y_{mt} = 1|\mathbf{x}_{mt}; \mathbf{w}^i)$ in (1) as a prior probability and $P(y_{mt}|Y) = P(y_{mt} = 1|Y, \mathbf{x}; \mathbf{w}^i) = \frac{e^{\mathbf{w}^{iT} \mathbf{x}_{mt}}}{\sum_{t=1}^{T_m} e^{\mathbf{w}^{iT} \mathbf{x}_{mt}}}$ as a posterior probability, which can also be directly computed using Bayes rule.

Prediction: Given a test signal $\mathbf{x}^{test}$, the localization signal or instance label signal $\hat{y}_t^{test}$ is obtained by

$$\hat{y}_t^{\text{test}} = \arg\max_{a \in \{0,1\}} P(y_t = a|\mathbf{x}^{\text{test}}, \mathbf{w}) \; \forall \, t = 1, \ldots T.$$

A signal level label is obtained by

$$\hat{Y}^{\text{test}} = \cup_{t=1}^{T} \hat{y}_t^{\text{test}}.$$

4.3. Computational complexity

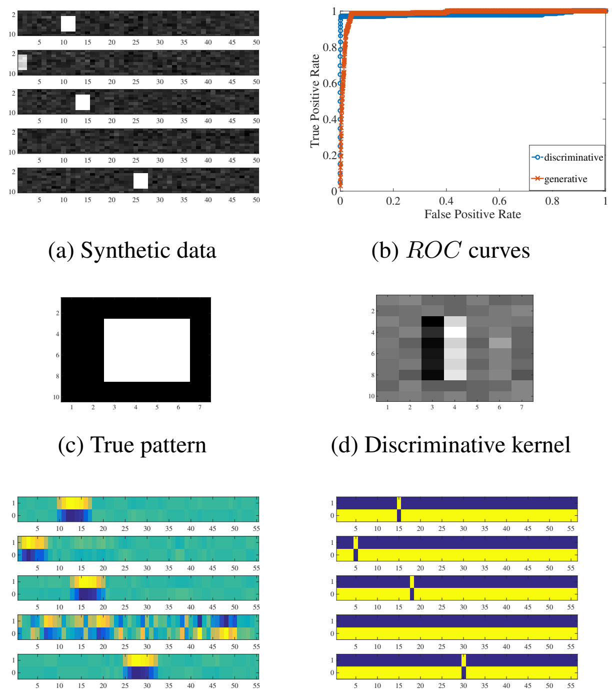
To simplify the computational complexity analysis, assume that the number of instance per signal T_m are all the same and equal to T. The overall computational complexity is $\mathcal{O}(NMTT_0)$, where N is the total number of iteration needed for updating the kernel $\mathbf{w}$. If T_0 is set to be large ($T_0 \approx T$), we can apply Fast Fourier Transform (FFT) and Inverse of FFT to speed up the convolution [8] such that the computational complexity will become $\mathcal{O}(NMT \log T)$.

5. NUMERICAL EVALUATION

In order to evaluate our discriminative pattern recognition approach, we perform a numerical synthetic experiment.

5.1. Synthetic data generation

The synthetic 2-D signals are generated with height (number of frequency bins) $F = 10$ and width (time frames) $T = 50$ by randomly placing a rectangular shape into one of the $T - 6$ maximally overlapped 10×7 windows of the signals. Each window is referred as an instance and is labeled as 1 if the rectangular shape is within that window, otherwise, it will be labeled as 0, the negative class. See Fig.3 (a) for an example.



(a) Synthetic data (b) ROC curves

(c) True pattern (d) Discriminative kernel

(e) Generative localization (f) Discriminative localization

Fig. 3: Synthetic data results

5.2. Numerical results

To verify our proposed approach, we use 10 independent random shuffles of 200 signals with balanced label that are split into 160 training signals and 40 test signals. The convolution kernel dimensions are set to $F = 10$ and $T_0 = 7$. Fig. 3(c) shows the original rectangular shape, while Fig. 3(d) shows the learned kernel, which appears to approximate the gradient of the rectangular shape. Fig. 3(f) verifies that the position where the rectangular signal lies is correctly predicted, however, using the original signal pattern in the generative framework yields some ambiguity about the shape location (see Fig. 3(e)). To show the resulting detection performance, we plot the receiver operating characteristic (ROC) curve [9]. The averaged test ROC curves based on the 10 different training sets are shown in Fig. 3(b). We can see that using a discriminative kernel produces higher true positive rate when the threshold is low.

6. REAL-WORLD EXPERIMENT

In this experiment, our goal is to learn a discriminative activation signature for each appliance using a set of training data and to test the detection performance on a separate test data set.

6.1. Data Set and preprocessing

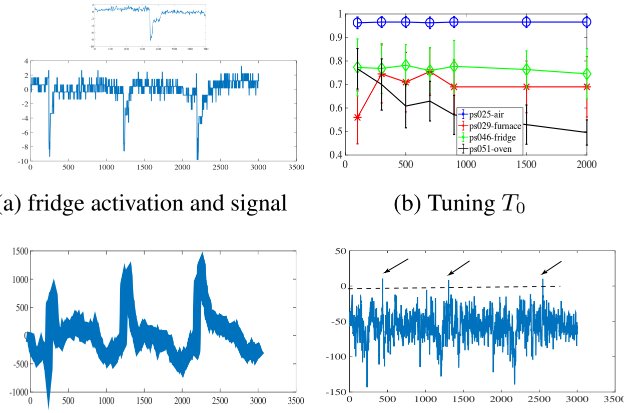
We use the Pecan Street dataset (*Source: Pecan Street Research Institute*), which contains four homes of disaggregated, time-sampled electricity usage data. The data set includes both voltage and apparent power readings in a period of 25 days. For the experimental setup, we split the four home data into training data with extracted activation signature with 1000 samples of a period of $11/17/2012$-$11/25/2012$ and test data is one hour readings, which contains around $500,000$ samples, with a period of $11/26/2012$-$12/11/2012$. For each home and each appliance, the training activation event of short sequence voltage responses are generated based on the ground truth of a power increase from 0 to 80 watt or more on the independent measurement from a commercial power meter. The negative labeled data of short sequence voltage responses are randomly extracted based on non-increase of the power meter.

Due to a time varying DC offset on voltage peak to peak (V_{pp}) value, we consider a moving window (each window contains 1000 samples) approach to calculate the average DC offset signal. For both training and test data, we remove that DC offset. We also apply a five-tap median filter to despike the voltage waveforms, since the voltage peak to peak (V_{pp}) waveform is corrupted by spike noise. For home ps-029, we use a fifteen-tap median filter.

6.2. Results and Analysis

In the training phase, we tune the window size T_0 using ten random shuffles of the data. On each of the ten, we first shuffle and then pick the first 80% of the data for training and the remaining 20% for validation. For each random shuffle, we compute the signal label accuracy $1/M \sum_{m=1}^{M} I(\hat{Y}_m^{val} = Y_m^{val})$ for $T_0 \in \{100, 300, 500, 700, 900, 1500, 2000\}$ and present mean and standard deviation (over the ten shuffles) as in Fig. 4(b). An iterative gradient descent method is used to find the discriminative activation signature. To compare with the results obtained from a generative approach [1], we show a detection example from the generative approach and the discriminative approach in the Fig. 4. Since [1] uses T_0=700, we show the resulting AUCs comparison with our approach by using both T_0=700 and best window size in the Table 1.

Fig. 4(c) shows an example for the detector output (before applying the threshold) for the generative detector. We observe that the peak level of the discriminative detector output in Fig. 4(d) appears more consistent than that of the generative detector. In general, we observe that the discriminative approach presents higher detection performance than the generative approach, especially for some of the appliances which contains more variation in their template. For example, for oven in home ps-025 and Fridge in ps-046, the detection AUCs of the discriminative approach are 0.86 and 0.87 respectively, which are significantly higher than the generative approach AUCs, 0.52 and 0.49 respectively.



(a) fridge activation and signal (b) Tuning T_0

(c) Generative detector (d) Discriminative detector

Fig. 4: Detection comparison between generative and discriminative fridge activation patterns.

House ID	App. Name	Gen.(T_0=700)	Disc.(T_0=700)	Disc.(best T_0)
PS-025	Air-Cond.	0.95	**0.97**	**0.97** T_0=700
PS-025	Oven	0.52	0.80	**0.86** T_0=100
PS-029	Air-Cond.	0.92	**0.99**	**0.99** T_0=700
PS-029	Dryer	**0.99**	0.93	0.96 T_0=100
PS-029	Fridge	0.72	0.85	**0.86** T_0=900
PS-029	Furnace	0.86	**0.89**	**0.89** T_0=700
PS-029	Microwave	0.88	**0.94**	**0.94** T_0=700
PS-029	Oven	**0.91**	0.78	0.88 T_0=100
PS-046	Air-Cond.	0.85	0.93	**0.97** T_0=500
PS-046	Fridge	0.49	0.85	**0.87** T_0=900
PS-046	Furnace	0.54	**0.56**	**0.56** T_0=700
PS-046	Oven	**0.92**	0.76	0.88 T_0=100
PS-051	Air-Cond.	0.91	**0.97**	**0.97** T_0=700
PS-051	Oven	**0.78**	0.61	0.72 T_0=100

Table 1: AUC for the generative method [1] and for our discriminative method.

Discussion on computational complexity: In a generative approach, [1] proposes an algorithm of computational complexity of $\mathcal{O}(T_0(MT)^2)$. In our discriminative approach, the computational complexity is $\mathcal{O}(NMTT_0)$. If the total number of iteration N is set to be less than MT, our discriminative approach is more efficient.

7. CONCLUSION

In this paper, we proposed a discriminative recurring pattern recognition model and provided an MLE solution approach. We first evaluated our proposed approach on a synthetic data and then we compared our discriminative approach with the generative approach on a real-world appliance voltage data. The results indicated that the discriminative approach are better at localizing the appliance activation patterns and more robust to their template variations. The resulting detection AUCs are higher than the AUCs of the generative approach on a large number of cases.

Acknowledgment: We would like to thank the Pecan Street Research Institute and Intel corporation for providing the dataset used in this paper.

8. REFERENCES

[1] Zeyu You, Raviv Raich, and Yonghong Huang, "An inference framework for detection of home appliance activation from voltage measurements," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 6033–6037.

[2] Alex S Park and James R Glass, "Unsupervised pattern discovery in speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 186–197, 2008.

[3] Aline Cabasson and Olivier Meste, "Time delay estimation: a new insight into the woody's method," *IEEE signal processing letters*, vol. 15, pp. 573–576, 2008.

[4] Yoshiki Tanaka, Kazuhisa Iwamoto, and Kuniaki Uehara, "Discovery of time-series motif from multi-dimensional data based on mdl principle," *Machine Learning*, vol. 58, no. 2-3, pp. 269–300, 2005.

[5] Jessica Lin, Eamonn Keogh, Stefano Lonardi, and Pranav Patel, "Finding motifs in time series," in *Proc. of the 2nd Workshop on Temporal Data Mining*, 2002, pp. 53–68.

[6] Alan L Yuille and Anand Rangarajan, "The concave-convex procedure (cccp)," *Advances in neural information processing systems*, vol. 2, pp. 1033–1040, 2002.

[7] Gert R Lanckriet and Bharath K Sriperumbudur, "On the convergence of the concave-convex procedure," in *Advances in neural information processing systems*, 2009, pp. 1759–1767.

[8] Henri J Nussbaumer, *Fast Fourier transform and convolution algorithms*, vol. 2, Springer Science & Business Media, 2012.

[9] Andrew P Bradley, "The use of the area under the roc curve in the evaluation of machine learning algorithms," *Pattern recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.