Time Series Domain Adaptation via Sparse Associative Structure Alignment

Ruichu Cai[1], Jiawei Chen[1], Zijian Li[1], Wei Chen[1], Keli Zhang[2], Junjian Ye[2], Zhuozhang Li[1], Xiaoyan Yang[1], Zhenjie Zhang[1]

[1]Guangdong University of Technology
[2]Huawei Noah's Ark Lab

cairuichu@gdut.edu.cn, chenjiawei952@gmail.com, leizigin@gmail.com, chenweidelight@gmail.com,
zhangkeli1@huawei.com, yejunjian@huawei.com, chokjohnlee@gmail.com, yangxiaoyan@gmail.com,
zhenjie.zhang@pvoice.io

Abstract

Domain adaptation on time series data is an important but challenging task. Most of the existing works in this area are based on the learning of the domain-invariant representation of the data with the help of restrictions like MMD. However, such extraction of the domain-invariant representation is a non-trivial task for time series data, due to the complex dependence among the timestamps. In detail, in the fully dependent time series, a small change of the time lags or the offsets may lead to difficulty in the domain invariant extraction. Fortunately, the stability of the causality inspired us to explore the domain invariant structure of the data. To reduce the difficulty in the discovery of causal structure, we relax it to the sparse associative structure and propose a novel sparse associative structure alignment model for domain adaptation. First, we generate the segment set to exclude the obstacle of offsets. Second, the intra-variables and inter-variables sparse attention mechanisms are devised to extract associative structure time-series data with considering time lags. Finally, the associative structure alignment is used to guide the transfer of knowledge from the source domain to the target one. Experimental studies not only verify the good performance of our methods on three real-world datasets but also provide some insightful discoveries on the transferred knowledge.

Introduction

Domain adaptation (Pan and Yang 2009; Long et al. 2015; Cai et al. 2019), utilizing both the labeled source domain data and the unlabeled target domain data, has a wide range of applications (Ganin and Lempitsky 2015; Lin and Lu 2018). To address the well-known phenomenon named *domain shift*, a large number of methods have been proposed by exploring various assumptions between the source and target domains (Tzeng et al. 2014; Cai et al. 2019; Zhang et al. 2019).

One of the most widely used assumptions in domain adaptation is the existence of domain-invariant representation in both source and target domains. Since such an assumption has achieved great performance in non-time series data (Cai et al. 2019; Wang et al. 2019), researchers have extended it to the time series data, by employing models like Recurrent Neural Network(RNN) (Mikolov et al. 2010) and variational
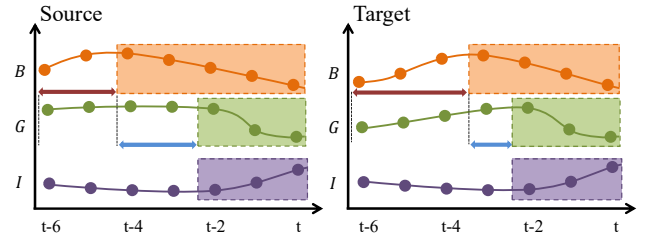
Figure 1: The illustration of the physiological mechanism in the human body among "Blood glucose (B) ↓", "Glucagon (G) ↓" and "Insulin (I) ↑". The decrease of "Blood glucose" leads to the decrease of "Glucagon" and the increase of "Insulin". The colored blocks denote the segments of change of variables. The different lengths of red double-head arrows denote different offsets. And the different lengths of blue double-head arrows denote different response times between "Blood glucose" and "Glucagon". Different response time means different time lags. (*Best view in color.*)

RNN (Chung et al. 2015), to learn representation from time series and using the gradient reversal layer (GRL) to align the representations learned from source and target time series data.

However, extracting domain-invariant information from time series data is a challenging task. Existing methods (da Costa et al. 2020; Purushotham et al. 2017), which simply employ the RNN based feature extractor, essentially assume that the conditional distributions are equal (Pan et al. 2010), i.e., $P_S(y|\varphi(x_1, x_2, \cdots, x_t) = P_T(y|\varphi(x_1, x_2, \cdots, x_t)$, in which $\varphi(\cdot)$ is the feature transformation mapping. This assumption works well in the static data but is difficult to satisfy in the time series data. Take Figure 1 as an example, due to the complex dependency structure among the timestamps, even small discrepancies from different domains (e.g. offsets and varying of time lags) may result in the difficulty to learn the domain invariant representation. Furthermore, as for the multivariable time series data, variables are not always i.i.d. Existing methods for time series domain adaptation, which ignore the associative structure among variables, might suffer from overfitting. Motivated by the toy example of Figure 2(a), data from the source and the target domains share the same stable
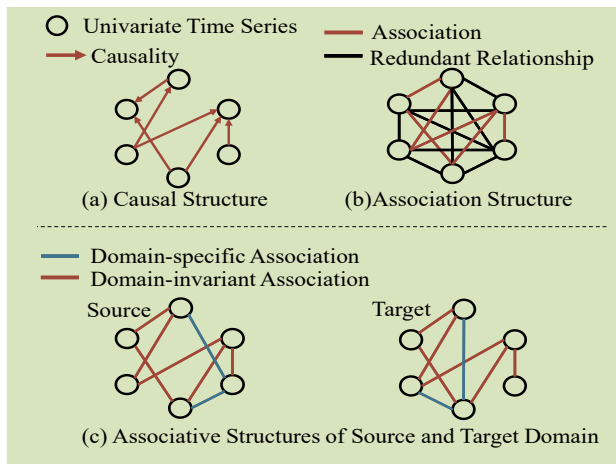
Figure 2: The illustration of various structures among six time series. (a) The causal structure of variables. (b) The existing methods take the conditional independence relationships into account and lead to overfitting. (c) Inspired by the stability of the causal mechanism, our method considers the stable and sparse associative structure among variables.

causal structure (e.g., the physiological mechanism among "Blood glucose↓"(B), "Glucagon↓"(G) and "Insulin↑" (I) shown in Figure 1), which is domain-invariant. However, as shown in Figure 2(b), the existing methods consider not only the ground truth associative structure but also the redundant relationships, which leads to overfitting. Since the causal structure from different domains is the same, time series data from the source and the target domains also share a similar associative structure. Figure 2(c) gives another insightful example, showing that considering domain-invariant associative structure and excluding domain-specific associations is important and can make the model robust and generalizable.

However, how to construct the associative structure among variables in time series data is another challenge, which is caused by the well-known discrepancy like time lags and offsets. According to the physiological mechanism of the human body, the decrease of "Blood glucose" leads to the decrease of "Glucagon" and the increase of "Insulin", and the response time of the physiological mechanism varies with ages and races, resulting in different time lags (i.e., different length of blue double-head arrows from the source and the target domains in Figure 1). Giving another example, let source domain data and target domain data be sampled from the elder and the younger patients respectively, the response time of the elder patients is longer than the younger ones. At the same time, the same mechanism often happens with varying start points as indicated by different offsets from a different domain (i.e., the different length of red double-head arrows in Figure 1). Existing work, which simply adopting RNNs as feature extractors to extract the domain-invariant representation, can not exclude the negative influence of time lags and offsets and further fails to extract the associative structure.

Based on the above intuition, we propose the sparse associative structure alignment (**SASA**) approach for time series

domain adaptation. The main challenges of SASA can be summarized into two folds. (1) How to get rid of the obstruction of time lags and offsets to extract the sparse associative structure? (2) How to extract the common associative structure and further extract the domain-invariant representation? To address these problems, first, we propose the adaptive segment summarization to ease the obstacle of offsets. Second, the proposed model extracts the sparse associative structure of the time series data via intra-variables and the inter-variables attention mechanism. Third, our model transfers the sparse associative structure from the source domain to the target domain by simply aligning the structure. Extensive experimental studies demonstrate that our **SASA** model outperforms the state-of-the-art time series unsupervised domain adaptation methods on three real world datasets.

Related Works

In this section, we mainly focus on the existing techniques on unsupervised domain adaptation as well as time series domain adaptation.

Domain Adaptation on Non-Time Series Data. The mainstream methods of unsupervised domain adaptation (Pan et al. 2010; Wei, Ke, and Goh 2016, 2018; Wen et al. 2019) aim to extract the domain invariant representation between the source and the target domains. Maximum Mean Discrepancy (**MMD**) is one of the most popular methods by using kernel-reproducing Hilbert space (Tzeng et al. 2014). Sun et al. (Sun, Feng, and Saenko 2016) propose second-order statistics for unsupervised domain adaptation. And Long et al. (Long et al. 2015) reduce the domain discrepancy by using an optimal multi-kernel selection method.

Another essential approach to unsupervised domain adaptation is to extract the domain-invariant representation by borrowing the idea of generative adversarial networks (Goodfellow et al. 2014). Ganin et al. (Ganin and Lempitsky 2015) introduce a gradient reversal layer to fool the domain classifier and further extract the domain-invariant representation. Tzeng et al. (Tzeng et al. 2017) propose a novel unified framework for adversarial domain adaptation. Recently, considering the fine-grained alignment and aiming to prevent the false alignment. Xie et al. (Xie et al. 2018) address the unsupervised domain adaptation problem by aligning the centroid for each class in source and target domains with the help of pseudo labels.

Under the causality view over the variables, the domain adaptation scenario can be determined by the causal mechanism. Three scenarios including target shift, condition shift, and generalized target shift, are discussed by Zhang et al. (Zhang et al. 2013). Based on the former, Germain et al.(Germain et al. 2016) and Zhang et al. (Zhang, Gong, and Schölkopf 2015) investigate more on the generalized target shift in the context of domain adaptation. Recently, following the causal model of the data generation process, Cai et al. (Cai et al. 2019) address this problem by extracting the disentangled semantic representation on the recovered latent space.

In this paper, we study the problem of unsupervised domain adaptation for time series data. Our **SASA** method is
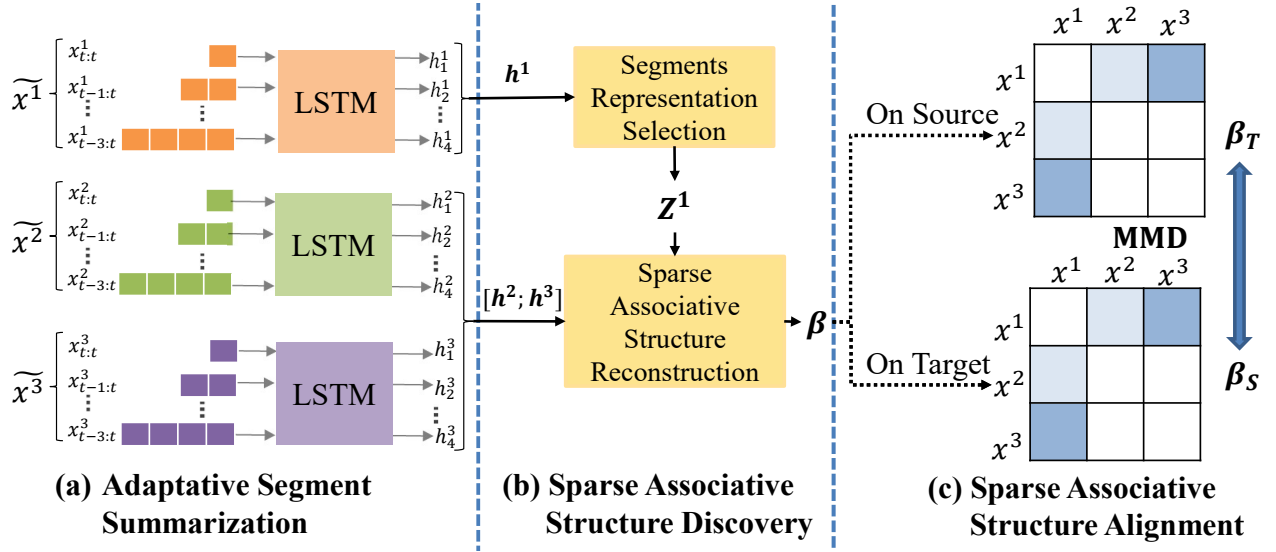
Figure 3: The framework of the sparse associative structure alignment model. (a) Adaptative segment summarization process with variable-specific LSTM. (b) Sparse associative structure discovery via intra-variables and inter-variables attention mechanism. (c) Sparse associative structure alignment between the source and the target domain. (*Best view in color.*)

first inspired by the causal mechanism from observed data. And we further consider a more relaxed sparse associative structure since any two variables contain causal structure also have associative structure.

Domain Adaptation on Time Series data. Recently, unsupervised domain adaptation on time series data has received more and more attention. Da Costa et al. (da Costa et al. 2020) employ the most straightforward method and simply replace with feature extractor with RNN based feature extractors to extract the representation of time series data. Purushotham et al. (Purushotham et al. 2017) use variational RNN (Chung et al. 2015) to extract the latent representations of time series. There are limited works on time series domain adaptation. One possible solution is the direct extension of the unsupervised domain adaptation methods on non-time series data to the time series data. However, this straightforward method might not work in time series data, since it's difficult to align the conditional distribution of the observed data in all timestamps.

Since the existing methods (da Costa et al. 2020; Purushotham et al. 2017) for time series domain adaptation cannot well align the condition distribution of the time series data, we propose a novel domain adaptation method for time series data, which aims to distill the sparse associative structure and filter the domain-specific information.

Sparse Associative Structure Alignment

In this section, we elaborate our Sparse Associative Structure Alignment (**SASA**) Model that distills the sparse associative structure and extracts the domain-invariant information from time series data. In this section, we first formulate the problem. Then, we provide the details of our model.

Problem Formulation and Overview

In this work, we focus on the problem of unsupervised domain adaptation for time series data. We let $\boldsymbol{x} = \{\boldsymbol{x}_{t-N+1}, \cdots, \boldsymbol{x}_{t-1}, \boldsymbol{x}_t\}$ denote a multivariate time series sample with N time steps, where $\boldsymbol{x}_t \in \mathbb{R}^M$, and $y \in \mathbb{R}$ is the corresponding label. We assume that $P_S(\boldsymbol{x}, y)$ and $P_T(\boldsymbol{x}, y)$ are different distributions from the source and the target domains but are generated from a shared causal mechanism. Since the two variable sets generated by a same causal structure should share the same associative structure, $P_S(\boldsymbol{x}, y)$ and $P_T(\boldsymbol{x}, y)$ share the same associative structure. $(\mathcal{X}_S, \mathcal{Y}_S)$ and $(\mathcal{X}_T, \mathcal{Y}_T)$, which are sampled from $P_S(\boldsymbol{x}, y)$ and $P_T(\boldsymbol{x}, y)$ respectively, denote the source and target domain dataset. Then we further assume that each source domain sample $\boldsymbol{x}_S$ comes with y_S, while the target domain has no labeled sample. Our goal is to devise a predictive model that can predict y_T given time series sample $\boldsymbol{x}_T$ from the target domain.

To achieve this goal, we aim to extract the domain-invariant representation in the form of associative structure. This solution is inspired by intuition that the causal mechanism is invariant across the domains. Due to the complexity of learning causal structure, we relax the causal structure to the sparse associative structure. Considering that the offsets vary with different domains and hinder the model from extracting the domain-invariant associative structure, we first elaborate on how to obtain the fine-grain segments of time series data to ease the obstacle of the offsets. Sequentially, we reconstruct the associative structures via the intra-variables attention mechanism and the inter-variables attention mechanisms with considering time lags from different domains. Different from the existing works that align the feature from different domains, our **SASA** model aligns the common associative structures from different domains to

indirectly extract the domain-invariant representation.

Adaptive Segment Summarization

In this subsection, we will elaborate on how to obtain the candidate segments to remove the obstacle of offsets. As shown in Figure 1, the orange blocks, whose duration varies with different domains, denote the segment of the change of variable 'B'. Existing methods, which take the whole time series data as input, can not accurately capture when a segment starts and when a variable affects the others, i.e., the sphere of influence of any variables. Therefore, these methods can not address the noise of offsets (i.e., the duration between the start point of time series and the start point of a segment).

To address the aforementioned problem, we first propose the adaptive segment summarization. To obtain the candidate segments of i-th time series $\boldsymbol{x^i} = \{\boldsymbol{x}_{t-N+1}^i, \cdots, \boldsymbol{x}_{t-1}^i, \boldsymbol{x}_t^i\}$, we construct multiple segments with different length for each variable $\boldsymbol{x}^i$. We have:

$$\widetilde{\boldsymbol{x^i}} = \{\boldsymbol{x}_{t:t}^i, \boldsymbol{x}_{t-1:t}^i, \cdots, \boldsymbol{x}_{t-\tau+1:t}^i, \cdots, \boldsymbol{x}_{t-N+1:t}^i\}. \quad (1)$$

Motivated by RIM (Goyal et al. 2019), we allocate an independent LSTM for each variable. In detail, given a segment of i-th variable with τ time steps, we have:

$$h_\tau^i = f(\boldsymbol{x}_{t-\tau+1:t}^i; \boldsymbol{\theta}^i), \quad (2)$$

in which $\boldsymbol{\theta^i}$ denote the parameters of i-th LSTM. Note that the segments of any univariate time series $\boldsymbol{x^i}$ share the same LSTM, and finally we can obtain the segments representation set shown as follow:

$$\boldsymbol{h}^i = \{h_1^i, h_2^i, \cdots, h_\tau^i, \cdots, h_N^i\}. \quad (3)$$

Since it's almost impossible to manually extract all the exact segments from the multivariable time series data, we first obtain the representation of all candidate segments via the aforementioned processing. The most suitable segment representations are selected and used to reconstruct the associative structure.

Sparse Associative Structure Discovery

In this section, we will introduce how to generalize the most exact segment representations and how to reconstruct the associative structure with the help of intra-variable attention mechanism and inter-variable attention mechanism.

Segments Representation Selection via Intra-Variables Attention Mechanism. In order to get rid of the obstacle brought from the offsets, we need to pay more attention to the exact segment representation among all the candidate segment representations with the help of the self-attention mechanism (Vaswani et al. 2017). Formally, we calculate the weights of each segment of $\boldsymbol{x^i}$ as follow:

$$\begin{aligned} u_\tau^i &= \frac{1}{N}\sum_{k=1}^N \frac{(h_\tau^i \boldsymbol{W}^Q)(h_k^i \boldsymbol{W}^K)^\mathsf{T}}{\sqrt{d_h}}, \\ \boldsymbol{\alpha^i} &= \{\alpha_1^i, \alpha_2^i, \cdots, \alpha_\tau^i, \cdots, \alpha_N^i\} \\ &= \mathrm{sparsemax}(u_1^i, u_2^i, \cdots, u_\tau^i, \cdots, u_N^i), \end{aligned} \quad (4)$$

in which $\boldsymbol{W}^Q, \boldsymbol{W}^K$ are trainable projection parameters and $\sqrt{d_h}$ is the scaling factor. In order to obtain the

sparse weights that represent specific segment representation clearly, we choose sparsemax (Martins and Astudillo 2016) to calculate the weights. The sparsemax is defined as $\mathrm{sparsemax}(\boldsymbol{z}) = \arg\min_{\boldsymbol{p}\in\Delta^{K-1}} ||\boldsymbol{p} - \boldsymbol{z}||^2$, which returns the Euclidean projection of vector $\boldsymbol{z} \in \mathbb{R}^K$ onto probability simplex Δ^{K-1}. As a result, we obtain the weighted segment representation of variable $\boldsymbol{x^i}$ as follow:

$$Z^i = \sum_{\tau=1}^N \alpha_\tau^i \cdot (h_\tau^i \boldsymbol{W}^V), \quad (5)$$

in which $\boldsymbol{W}^V$ is trainable projection parameter. Note that $\boldsymbol{\alpha}$ also denotes the probability of the length of a segment. For generalization, we also consider the case that the duration of a segment of a given variable varies with different domains. In this case, in order to reconstruct the associative structure more precisely, we minimize the maximum mean discrepancy (**MMD**) between $\boldsymbol{\alpha}$ from the source and the target domain to remove the obstacle of offsets. It restricts the duration of the segment from different domains to be similar, which contributes to extracting structure for transfer. Formally, we have:

$$\mathcal{L}_\alpha = \sum_{m=1}^M ||\frac{1}{|\mathcal{X}_S|}\sum_{\boldsymbol{x}_S\in\mathcal{X}_S} \boldsymbol{\alpha}_S^m - \frac{1}{|\mathcal{X}_T|}\sum_{\boldsymbol{x}_T\in\mathcal{X}_T} \boldsymbol{\alpha}_T^m||, \quad (6)$$

in which $\boldsymbol{\alpha}_S^m$ and $\boldsymbol{\alpha}_T^m$ denote the weights of segments of the m-th variable from the source and the target domains calculated by Equation (4).

Sparse Associative Structure Reconstruction via Inter-variables Attention Mechanism. With the help of the intra-variables attention mechanism, we extract the weighted segment representations despite the obstacle of offsets. Then we utilize these weighted segment representations to reconstruct the sparse associative structure among different variables. So we propose the inter-variables attention mechanism to mine the associative structure among variables.

In this part, our goal is to reconstruct the associative structure among variables. Instead of using the self-attention mechanism in the intra-variables attention mechanism, we employ the "referenced" attention mechanism (Bahdanau, Cho, and Bengio 2015). One of the most straightforward methods to calculate the degree of association of variable i and variable j is shown as follow:

$$\boldsymbol{e}^{ij} = \frac{Z^i \cdot Z^j}{||Z^i|| \cdot ||Z^j||} \quad (7)$$

However, the associative structure calculated by Equation (7) ignores the time lags from different domains between i and j. Since Equation (7) does not refer to time lags among variables, the associative structure might be falsely estimated. To take the time lags into account, we calculate the degrees of association between variable i and variable j by:

$$\begin{aligned} e_\tau^{ij} &= \frac{Z^i \cdot h_\tau^j}{||Z^i|| \cdot ||h_\tau^j||}, \\ \boldsymbol{e}^{ij} &= \{e_1^{ij}, e_2^{ij}, \cdots, e_\tau^{ij}, \cdots, e_N^{ij}\}. \end{aligned} \quad (8)$$

Then we normalize these degrees of association with sparsemax (Martins and Astudillo 2016) as intra-variable atten-

tion. Formally, we have:

$$\boldsymbol{\beta}^i = \{\boldsymbol{\beta}^{i1}, \boldsymbol{\beta}^{i2}, \cdots, \boldsymbol{\beta}^{ij}, \cdots, \boldsymbol{\beta}^{iM}\}$$
$$= \text{sparsemax}(\{\boldsymbol{e}^{i1}, \boldsymbol{e}^{i2}, \cdots, \boldsymbol{e}^{ij}, \cdots, \boldsymbol{e}^{iM}\})(j \neq i). \quad (9)$$

Note that $\beta_\tau^{ij} \in \boldsymbol{\beta}^i$ denotes the associative strength between variable i and variable j with regard to segment duration of τ.

Sparse Associative Structure Alignment

In this subsection, we aim to extract the domain-invariant information for time series data with the help of the extracted associative structure from the source and the target domains.

We reconstruct the associative structure by Equation (8) and (9) taking the time lags into account. In order to extract the domain-invariant associative structure, we need to restrict the distance of the structure between the source and the target domains. Since $\boldsymbol{\beta}^{ij}$ can be seen as the associative strength distribution between i and j, we turn the problem of structure distance measure to the distribution distance measure. In this paper, we borrow the idea of domain confuse network (Tzeng et al. 2014) and employ maximum mean discrepancy (**MMD**) for associative structure alignment. Formally, we have:

$$\mathcal{L}_\beta = \sum_{m=1}^{M} \|\frac{1}{|\mathcal{X}_S|} \sum_{\boldsymbol{x}_S \in \mathcal{X}_S} \boldsymbol{\beta}_S^m - \frac{1}{|\mathcal{X}_T|} \sum_{\boldsymbol{x}_T \in \mathcal{X}_T} \boldsymbol{\beta}_T^m\|. \quad (10)$$

Note that we minimize the associative structure adjacent matrix instead of aligning the features like what (Tzeng et al. 2014) does.

Model Summary

Task based Label Predictor. We aim to obtain the domain-invariant representations which are combined with the sparse associative structure $\boldsymbol{\beta}$. In detail, we first calculate the associative structure representations of variable j, which is shown as follow:

$$\boldsymbol{U}^{ij} = \sum_{\tau=1}^{N} \beta_\tau^{ij} \cdot h_\tau^j,$$
$$U^i = \sum_{m=1, m \neq i}^{M} \boldsymbol{U}^{im}. \quad (11)$$

As a result, we can obtain the final representations by concatenating weighted segment representations and associative structure representations as follow:

$$H^i = \left[Z^i; U^i \right]. \quad (12)$$

For convenience, we describe the above process as:

$$\boldsymbol{H} = G_H(f(\boldsymbol{x}; \boldsymbol{\Theta}); \boldsymbol{W}^Q, \boldsymbol{W}^K, \boldsymbol{W}^V), \quad (13)$$

in which G_H actually denotes the feature extractor containing the aforementioned two kinds of attention mechanisms. $\boldsymbol{H} = [H^1; H^2; \cdots; H^M]$ denotes the final representation, we further let $\boldsymbol{\Theta}$ be the parameters of variable-specific LSTM.

After obtaining the final representation, we take $\boldsymbol{H}$ as the input of label classifier $G_y(\cdot; \boldsymbol{\phi})$ whose loss function is $\mathcal{L}_y$.

For the classification problems, we employ cross-entropy as the label loss. For the regression problems, we employ RMSE as the label loss.

The label classifier with the trained optimal parameters is adapted to the target domains.

$$y_{pre} = G_y(G_H \left(f(\boldsymbol{x}; \boldsymbol{\Theta}); \boldsymbol{W}^Q, \boldsymbol{W}^K, \boldsymbol{W}^V \right), \boldsymbol{\phi}). \quad (14)$$

Objective Function. The total loss of the proposed structure alignment model for time series domain adaptation is formulated as:

$$\mathcal{L} \left(\boldsymbol{\Theta}, \boldsymbol{W}^Q, \boldsymbol{W}^K, \boldsymbol{W}^V, \boldsymbol{\phi} \right) = \mathcal{L}_y + \omega(\mathcal{L}_\alpha + \mathcal{L}_\beta), \quad (15)$$

in which ω is hyper-parameter.

Under the above objective function, our model is trained on the source and target domain using the following procedure:

$$\left(\boldsymbol{\Theta}, \boldsymbol{W}^Q, \boldsymbol{W}^K, \boldsymbol{W}^V, \boldsymbol{\phi} \right) =$$
$$\arg \min_{\boldsymbol{\Theta}, \boldsymbol{W}^Q, \boldsymbol{W}^K, \boldsymbol{W}^V, \boldsymbol{\phi}} \mathcal{L} \left(\boldsymbol{\Theta}, \boldsymbol{W}^Q, \boldsymbol{W}^K, \boldsymbol{W}^V, \boldsymbol{\phi} \right). \quad (16)$$

Experiments and Result

Setup

Boiler Fault Detection Dataset. The boiler data consists of sensor data from three boilers from 2014/3/24 to 2016/11/30. There are 3 boilers in this dataset and each boiler is considered as one domain. The learning task is to predict the *faulty blowdown valve* of each boiler. Since the fault data is very rare. It's hard to obtain the fault samples in the mechanical system. So it's important to utilize the labeled source data and unlabeled target data to improve the model generalization.

Air Quality Forecast Dataset. The air quality forecast dataset(Zheng et al. 2015) is collected in the Urban Air project[1] from 2014/05/01 to 2015/04/30, which contains air quality data, meteorological data, and weather forecast data, etc. The dataset covers 4 major Chinese cities: Beijing (B), Tianjin (T), Guangzhou(G), and Shenzhen(S). We employ air quality data as well as meteorological data to predict the PM2.5. We choose the air quality station with the least missing value and take each city as a domain. We use this dataset because the air quality data is common and the sensors in the smart city systems usually contain complex causality. The association among sensors are often sparse, which is suitable for our model.

In-hospital Mortality Prediction Dataset. MIMIC-III(Johnson et al. 2016; Che et al. 2018)[2] is another published dataset with de-identified health-related data associated with more than forty thousand patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. It's the benchmark of time series domain adaptation in VRADA(Purushotham

---
[1]https://www.microsoft.com/en-us/research/project/urban-air/
[2]https://mimic.physionet.org/gettingstarted/demo/

Method	B→T	G→T	S→T	T→B	G→B	S→B	B→G	T→G	S→G	B→S	T→S	G→S	Avg
LSTM_S2T	40.20	41.67	48.91	52.81	56.44	68.14	19.00	19.76	17.56	13.82	13.82	13.86	33.83
RDC	39.72	40.80	47.75	51.98	55.83	67.67	18.18	19.10	15.43	13.70	13.75	13.76	33.14
R-DANN	39.93	40.98	46.16	52.72	55.65	66.47	18.00	18.47	15.18	13.82	13.78	13.79	32.91
VRADA	38.12	38.69	45.29	52.14	54.51	64.41	17.30	17.95	14.63	13.80	13.90	13.80	32.04
SASA-α	36.60	34.42	41.31	48.34	54.20	59.09	16.42	16.48	14.30	13.68	13.53	13.47	30.15
SASA-β	35.54	35.10	42.16	48.40	54.42	60.45	16.66	16.58	14.62	13.62	13.49	13.68	30.39
SASA	**34.26**	**33.84**	**40.91**	**48.15**	**54.14**	**56.80**	**16.40**	**15.41**	**14.23**	**13.49**	**13.46**	**13.38**	**29.54**

Table 1: RMSE on air quality prediction.

Method	2→1	3→1	4→1	1→2	3→2	4→2	1→3	2→3	4→3	1→4	2→4	3→4	Avg
LSTM_S2T	80.52	78.79	76.85	80.24	81.43	77.24	75.77	79.30	75.56	65.79	68.93	69.41	75.82
RDC	81.36	78.94	77.11	80.66	82.40	78.47	75.96	79.39	75.63	66.20	69.59	70.21	76.33
R-DANN	81.38	79.30	77.57	80.70	82.71	78.38	76.00	79.18	76.18	66.64	69.83	69.62	76.46
VRADA	82.12	80.68	77.71	82.24	83.09	78.82	76.27	80.00	76.28	68.20	70.01	71.34	77.23
SASA-α	84.62	81.02	79.89	83.36	84.12	80.78	76.78	80.72	78.37	68.65	70.62	72.23	78.47
SASA-β	83.68	81.47	78.36	82.70	84.36	81.20	77.14	80.52	77.86	68.23	70.35	72.57	78.20
SASA	**85.03**	**82.91**	**80.32**	**83.82**	**85.20**	**82.03**	**77.83**	**81.10**	**78.93**	**69.02**	**70.96**	**72.76**	**79.16**

Table 2: AUC score(%) on in-hospital mortality prediction.

et al. 2017). Similar to Purushotham et al.(Purushotham et al. 2018), we choose 12 time series (such as Heart Rate, Temperature, Systolic blood pressure, etc) from 35637 records. In order to prepare the in-hospital mortality prediction dataset for time series domain adaptation, we split the patients into 4 groups according to their age (Group1: 20-45, Group2: 46-65, Group3: 66-85, Group4: >85).

Baseline

LSTM_S2T. LSTM_S2T uses the source domain data to train a vanilla LSTM model and applies it to the target domain without any adaptation (S2T stands for source to target). It's expected to provide the lower bound performance.
R-DANN. R-DANN (da Costa et al. 2020) is an unsupervised domain adaptation architecture proposed in (Ganin and Lempitsky 2015) with GRL (Gradient Reversal Layer) on LSTM, which is a straightforward solution for time series domain adaptation.
RDC. Deep domain confusion is an unsupervised domain adaptation method proposed in (Tzeng et al. 2014) which minimizes the distance between the source and target distributions by employing Maximum Mean Discrepancy (MMD). Similar to the aforementioned R-DANN, we use LSTM as the feature extractor for time series data.
VRADA. VRADA (Purushotham et al. 2017) is a time series unsupervised domain adaptation method which combines the GRL and VRNN (Chung et al. 2015).

For a fair comparison, the total numbers of parameters of all the baselines and our method are about equal, which is shown in Table 3. We use the same parameter combination on each dataset and also apply three different random seeds to each experiment.

Method	Boiler	Air	MIMIC-III
LSTM_S2T	82924	46191	72106
RDC	82924	46191	72106
R-DANN	82652	46183	71479
VRADA	83532	45898	72784
Ours	82322	45636	71402

Table 3: Total numbers of parameters of all the methods in different datasets.

Model Variants

In order to verify the effectiveness of each component of our model, we further devise the following model variants.

- SASA-α: We remove $\mathcal{L}_\alpha$ to verify the usefulness of the segment length restriction loss.

- SASA-β: We remove $\mathcal{L}_\beta$ to verify the usefulness of the sparse associative structure alignment loss.

Result

Results on Boiler Fault Detection. The AUC result in the boiler fault detection dataset is shown in Table 4. Our SASA model significantly outperforms other baselines on all the tasks. It's worth mentioning that our sparse associative structure alignment model promotes the AUC score substantially on harder transfer tasks, e.g. $1 \rightarrow 2$ and $3 \rightarrow 2$, which are respectively improved by 3.95 and 2.56 points compared with VRADA. On some easy tasks such as $1 \rightarrow 3$, $2 \rightarrow 3$, and $3 \rightarrow 1$, though the other baselines perform well, our method still achieves a comparable result. We also conduct the Wilcoxon signed-rank test (Wilcoxon 1945) on the reported score, the p-value is 0.027, which means that our method significantly outperforms the baselines with the p-value threshold 0.05.

Method	1→2	1→3	3→1	3→2	2→1	2→3	Avg
LSTM_S2T	67.09	94.54	93.14	56.09	84.99	91.31	81.19
RDC	68.29	94.65	93.38	57.32	85.31	92.57	81.92
R-DANN	67.71	94.69	93.92	58.53	85.67	91.66	82.03
VRADA	67.59	94.88	93.65	60.59	85.96	92.62	82.55
SASA-α	70.83	95.86	94.63	60.76	87.27	93.28	83.77
SASA-β	69.76	95.01	94.56	61.31	86.78	92.84	83.38
SASA	**71.54**	**96.39**	**94.77**	**63.15**	**87.76**	**93.59**	**84.53**

Table 4: AUC score(%) on boiler fault detection.

Results on Air Quality Forecast. Similar to the result in the boiler fault detection dataset, our model also achieves great results and outperforms all the other baselines on all tasks, which is reported in Table 1. According to the result, we can observe that: 1) The performance between the closer cities is better than that of farther cities. For example, since the distance between Beijing and Tianjin is smaller than the distance between Beijing and Guangzhou and between Beijing to Shenzhen, the promotion of $B \to T$ is better than that of $B \to G$ and $B \to S$. This is because the cities pair with closer distance share more common associative structure. 2) Our method still achieves the best result even the source city is far away from the target city, e.g. Beijing and Shenzhen, this phenomenon reflects that our sparse associative structure alignment model well extracts the associative structure among different variables. 3)The performance is not so notable compared with other tasks when Shenzhen is taken as the target domain. This is because the label value of this domain is much lower than other domains. We also conduct the Wilcoxon signed-rank test (Wilcoxon 1945) on the reported score, the p-value is 0.002, which means that our method significantly outperforms the baselines with the p-value threshold 0.05.

Results on In-hospital Mortality Prediction. We also testify our model on MIMIC-III dataset, which is chosen as the benchmark of time series domain adaptation in (Purushotham et al. 2017). We choose 12 variables described in (Purushotham et al. 2018) and reproduce a similar result of VRADA. As shown in Table 2, our model overpasses the other comparison methods on all the transfer tasks. Some domain adaptation tasks such as $2 \to 1$ and $3 \to 2$ are even improved by 2.91 and 2.11 points respectively. Furthermore, we also find that the performance between similar domains like 1 and 2, 2 and 3, as well as 3 and 4 are better than others. We also conduct the Wilcoxon signed-rank test (Wilcoxon 1945) on the reported score, the p-value is 0.0022, with the p-value threshold 0.05.

Ablation Study and Visualization

The study of the usefulness of the segment length restriction. In order to verify the effectiveness of sparse associative structure alignment, we remove $\mathcal{L}_\alpha$ and the model is named **SASA-α**. Compared the result of **SASA** and **SASA-α**, we can find that the performance of **SASA-α** drops. This is because of α represents the probability of the length of a segment. And the duration of segments varies with domains.

(a) Beijing   (b) Shenzhen
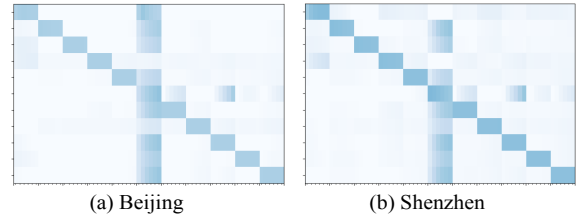
Figure 4: The illustration of visualization of correlation structure adjacent matrix of Beijing → Shenzhen. Deeper the color is, the stronger the relationship is. We can find that the structure is sparse.

With the restriction of α, we can exclude the influence of domain-specific segments duration.

The study of the effectiveness of sparse associative structure alignment. In order to verify the effectiveness of the segment length restriction, we remove the sparse associative structure alignment loss. According to the experiment result of **SASA-β**, we can find that the performance of **SASA-β** is worse than standard **SASA**. This is because the sparse associative structure has been extracted, which is also more robust than that of normal feature extractor. But the reserved domain-specific associative relationships lead to the suboptimal result. Note that **SASA-β** is still better than the other baselines. This is because $\mathcal{L}_\alpha$ aligns the offsets between different domains, which benefits to extracting sparse associative structure for adaptation.

Visualization of Aligned Structure. To further investigate our approach, we perform the visualization of aligned sparse associative structure over the air quality dataset and attempt to extract the common sparse associative structure, which is shown in Figure 4. The visualization shows the sparse associative structures of Beijing and Shenzhen respectively. The deeper the color is, the stronger the association between two variables. we can find that (1) the associative structures from different domains are very sparse. (2) the sparse associative structures from the source and the target domains have many shared associative relationships, which means that similar sparse associative structure is shared in different domains.

Conclusion

This paper presents a sparse associative structure alignment model for time series unsupervised domain adaptation. In our proposal, a sparse associative structure discovery method, equipped with an adaptive summarization of the series segments, is used to extract the structure of the time series, and an MMD based structure alignment method is used to transfer the knowledge from the source domain to the target domain. The success of the proposed approach not only provides an effective solution for the time-series domain adaptation task, but also provides some insightful results on what is transferable on the time-series data.

Acknowledgments

Statement About The Potential Ethical Impact

The Time series domain adaptation model, which is combined with the correlative relationship, is more robust than the existing methods and yields significant performance in unlabeled test data, which can apply in mechanical systems, smart cities as well as healthcare. The positive implications of applying our method include:
(1) Significant improvement of unsupervised domain adaptation for time series data, which reduces the requirement for manually acquiring labeled data and make the machine learning model available for use the in low-resource settings.
(2) Unsupervised domain adaptation for time series data is beneficial to mitigate overfitting.
However, the negative implications of the increasingly powerful artificial intelligence technology should not be ignored. These technologies lack interpretability so it's hard to be trusted sometimes. Our method can figure out this circumstance to some extent, but it can be better if we take causality into account.

References

Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. URL http://arxiv.org/abs/1409.0473.

Cai, R.; Li, Z.; Wei, P.; Qiao, J.; Zhang, K.; and Hao, Z. 2019. Learning disentangled semantic representation for domain adaptation. In *IJCAI: proceedings of the conference*, volume 2019, 2060. NIH Public Access.

Che, Z.; Purushotham, S.; Cho, K.; Sontag, D.; and Liu, Y. 2018. Recurrent neural networks for multivariate time series with missing values. *Scientific reports* 8(1): 1–12.

Chung, J.; Kastner, K.; Dinh, L.; Goel, K.; Courville, A. C.; and Bengio, Y. 2015. A recurrent latent variable model for sequential data. In *Advances in neural information processing systems*, 2980–2988.

da Costa, P. R. d. O.; Akçay, A.; Zhang, Y.; and Kaymak, U. 2020. Remaining useful lifetime prediction via deep domain adaptation. *Reliability Engineering & System Safety* 195: 106682.

Ganin, Y.; and Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, 1180–1189. PMLR.

Germain, P.; Habrard, A.; Laviolette, F.; and Morvant, E. 2016. A new PAC-Bayesian perspective on domain adaptation. In *International conference on machine learning*, 859–868.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, 2672–2680.

Goyal, A.; Lamb, A.; Hoffmann, J.; Sodhani, S.; Levine, S.; Bengio, Y.; and Schölkopf, B. 2019. Recurrent independent mechanisms. *arXiv preprint arXiv:1909.10893* .

Johnson, A. E.; Pollard, T. J.; Shen, L.; Li-wei, H. L.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Celi, L. A.; and Mark, R. G. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data* 3: 160035.

Lin, B. Y.; and Lu, W. 2018. Neural Adaptation Layers for Cross-domain Named Entity Recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2012–2022.

Long, M.; Cao, Y.; Wang, J.; and Jordan, M. 2015. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, 97–105. PMLR.

Martins, A.; and Astudillo, R. 2016. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *International Conference on Machine Learning*, 1614–1623.

Mikolov, T.; Karafiát, M.; Burget, L.; Černockỳ, J.; and Khudanpur, S. 2010. Recurrent neural network based language model. In *Eleventh annual conference of the international speech communication association*.

Pan, S. J.; Tsang, I. W.; Kwok, J. T.; and Yang, Q. 2010. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks* 22(2): 199–210.

Pan, S. J.; and Yang, Q. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* 22(10): 1345–1359.

Purushotham, S.; Carvalho, W.; Nilanon, T.; and Liu, Y. 2017. Variational Recurrent Adversarial Deep Domain Adaptation. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net. URL https://openreview.net/forum?id=rk9eAFcxg.

Purushotham, S.; Meng, C.; Che, Z.; and Liu, Y. 2018. Benchmarking deep learning models on large healthcare datasets. *Journal of biomedical informatics* 83: 112–134.

Sun, B.; Feng, J.; and Saenko, K. 2016. Return of frustratingly easy domain adaptation. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Tzeng, E.; Hoffman, J.; Saenko, K.; and Darrell, T. 2017. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7167–7176.

Tzeng, E.; Hoffman, J.; Zhang, N.; Saenko, K.; and Darrell, T. 2014. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474* .

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Wang, X.; Li, L.; Ye, W.; Long, M.; and Wang, J. 2019. Transferable attention for domain adaptation. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Wei, P.; Ke, Y.; and Goh, C. K. 2016. Deep nonlinear feature coding for unsupervised domain adaptation. In *IJCAI*, 2189–2195.

Wei, P.; Ke, Y.; and Goh, C. K. 2018. Feature analysis of marginalized stacked denoising autoenconder for unsupervised domain adaptation. *IEEE transactions on neural networks and learning systems* 30(5): 1321–1334.

Wen, J.; Liu, R.; Zheng, N.; Zheng, Q.; Gong, Z.; and Yuan, J. 2019. Exploiting local feature patterns for unsupervised domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 5401–5408.

Wilcoxon, F. 1945. Individual Comparisons by Ranking Methods. *Biometrics* 1(6): 80–83.

Xie, S.; Zheng, Z.; Chen, L.; and Chen, C. 2018. Learning semantic representations for unsupervised domain adaptation. In *International Conference on Machine Learning*, 5419–5428.

Zhang, K.; Gong, M.; and Schölkopf, B. 2015. Multi-source domain adaptation: A causal view. In *Twenty-ninth AAAI conference on artificial intelligence*.

Zhang, K.; Schölkopf, B.; Muandet, K.; and Wang, Z. 2013. Domain adaptation under target and conditional shift. In *International Conference on Machine Learning*, 819–827.

Zhang, Y.; Liu, T.; Long, M.; and Jordan, M. 2019. Bridging Theory and Algorithm for Domain Adaptation. In *International Conference on Machine Learning*, 7404–7413.

Zheng, Y.; Yi, X.; Li, M.; Li, R.; Shan, Z.; Chang, E.; and Li, T. 2015. Forecasting fine-grained air quality based on big data. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2267–2276. ACM.