Supplementary Material: Unsupervised Hierarchical and Heterogeneous Coupling Learning on Dynamic Categorical Data

Introduction

In this paper, we propose an unsupervised hierarchical and heterogeneous coupling learning (UNICORN) method for dynamic categorical data representation. UNICORN tackles the two critical challenges, i.e. *heterogeneous couplings* and *coupling dynamics*, that is rarely or insufficiently addressed by current coupling learning methods. Let us illustrate the challenges by Figure 1.

First, as shown in the left part of the figure, there may exist different grouping (which can be captured by similarity) effects in objects (e.g. c_1, c_2 and c_3) w.r.t. different categorical attributes; and such grouping may change from one categorical attribute to another. For example, student researchers working in a research lab may consist of undergraduate students (e.g., c_1), master students (c_2) and doctoral students (c_3), who may share certain similarities (e.g., w.r.t. attributes research area and cultural background) while also keep their heterogeneous characteristics (e.g., w.r.t. research method), corresponding to different couplings between students w.r.t. their qualifications. Further, when we discuss the student similarities w.r.t. another categorical attribute, e.g., religion background, these students form another grouping effect, e.g., Christian students, Muslim students, Buddhist students, and nonbelievers etc., and the couplings within and between these groups are different from that in the above qualification-based groups. Therefore, there are heterogeneous relations between objects of the same or different groups coupled w.r.t. different categorical attributes.

Second, we are particularly concerned with dynamic categorical data, i.e., as shown in the right part of Figure 1, new objects at time t are added to each coupled group of objects at time $t-1$. Accordingly, the couplings within and between categorical attributes change from time $t-1$ to t. These issues are widely seen in the real world, e.g., different types of investors trading in a capital market and online shopping business.

In the rest of this supplementary material, we first give a comprehensive review of related work. Then, we discuss the details of the notation system used in our paper. We further provide more technical details and supplementary experiments.

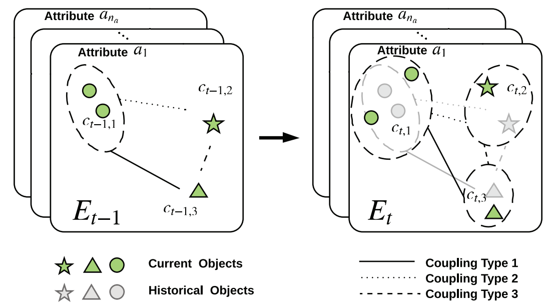


Figure 1: Hierarchical and heterogeneous couplings in dynamic categorical data.

Related Work

A fundamental issue in deeply understanding intrinsic data characteristics is to capture and represent complex coupling relationships (coupling for short) embedded in low-level attributes, attribute values, and objects. This is particularly important yet challenging for categorical data, as fewer methods are available and more complex scenarios exist in categorical data than in numerical data.

Traditionally, to model categorical data, related work focuses on capturing individual relations w.r.t. aspects of value frequency, co-occurrence, correlation, dependence or latent relations either within or between categorical attributes for similarity, metric, and representation learning (Ng et al. 2007; Ahmad and Dey 2007; Boriah, Chandola, and Kumar 2008; Cao et al. 2012; Cheung and Jia 2013). More recent focus is on integrating two to more types of above relations, leading to improved performance (Zhang et al. 2018), while also misses a comprehensive picture of complex interactions in data.

Increasing recent efforts have been made on designing different coupling learning methods to capture and represent value-to-attribute-to-object hierarchical and heterogeneous couplings (Cao 2015; Ienco, Pensa, and Meo 2012; Wang et al. 2015; Jian et al. 2017), which are fundamental yet challenging for an in-depth understanding of data complexities. These methods are created and applied for various areas and applications, demonstrating significant potential for mitigating gaps in existing categorical data similarity, metric and representation learning and for improving categorical data analytical applications, such as behavior analysis (Cao, Ou, and Yu 2012), recommendation (Zhang et al. 2018), graphic modeling (Do and Cao 2018), outlier detection (Pang et al. 2017), and video analysis (Shi et al. 2017).

However, hierarchical and heterogeneous couplings are very complex and highly difficult and costly to capture and integrate, showing consistencies, inconsistencies, redundancies, and complementarities etc. characteristics between each other. While tremendous progress has been shown in modeling various complex couplings, e.g., for recommendation (Zhang et al. 2018), graphic modeling (Do and Cao 2018) and video analysis (Shi et al. 2017), existing methods either ignore the inconsistencies and redundancies etc. complexities in heterogeneous couplings (Wang et al. 2015), only reduces redundancies in heterogeneous couplings (Ienco, Pensa, and Meo 2012; Wang et al. 2015; Jian et al. 2017), or simply eliminate the coupling inconsistencies by involving class labels (Zhu et al. 2018). Our proposed UNICORN makes a step forward to represent couplings between values, attributes, objects and object categories in unlabeled data, while avoids coupling redundancies and inconsistencies and maximizes information captured.

Further, no methods are available to learn hierarchical and heterogeneous couplings in dynamic categorical applications (DEnza and Markos 2015), which cannot be captured by existing work on learning data streams (Cao et al. 2010), change detection (Chen, Chen, and Lin 2009), and concept drift (Bai et al. 2016) etc. Typical methods focus on clustering change detection in categorical data, including information entropy-based clustering (Barbará, Li, and Couto 2002; Li et al. 2014), importance-based clustering (Chen, Chen, and Lin 2009; Bai et al. 2016), and fuzzy theory-based clustering (Cao et al. 2010; Saha and Maulik 2014). They ignore the hierarchical and heterogeneous couplings and object/coupling changes at different time points as illustrated in Figure 1. In addition, deep recurrent models (Bengio, Courville, and Vincent 2013) are used to abstract latent relations in sequential data, while ignore observable relations and their incremental changes along the timeframe. Instead, UNICORN captures the hierarchical and heterogeneous couplings and object/coupling changes in data evolution in an unsupervised way.

Notation System

Let us denote a dynamic categorical data set with n_t time points as a sequential matrix of three-element tuples $E = \{E_1, \cdots, E_t, \cdots, E_{n_t}\}$, where $E_t =< O_t, A, V_t >$ refers to a three-element tuple of categorical data observed at time t, $O_t = \{o_{t,i} | i = 1, \cdots, n_{o_t}\}$ is an object set with n_{o_t} objects, $A = \{a_j | j = 1, \cdots, n_a\}$ is an attribute set with n_a attributes, and $V_t = \bigcup_{j=1}^{n_a} V_t^{(j)}$ is a collection of attribute values with n_{v_t} values, in which $V_t^{(j)} = \{v_{t,i}^{(j)} | i = 1, \cdots, n_{v_t}^{(j)}\}$ is a set with $n_{v_t}^{(j)}$ values of attribute a_j.

In the notation systems, a subscript denotes a time point. For example, O_t refers to the observed object set at time t. The index of an element is also denoted by a subscript followed by the time point notation with a comma to split them. For example, $o_{t,i}$ refers to the i-th object in O_t at time t. The index of an attribute is denoted as a superscript with a bracket. For example, $V_t^{(j)}$ refers to the set of values of j-th attribute at time t. When there is no ambiguity, we omit the time point and index for a more concise representation. For example, at time t, we use $v^{(j)}$ to represent a categorical value in the j-th attribute. We use n with a subscript to denote the number of an element. For example, n_t refers to the number of time points, n_{o_t} refers to the number of objects at time t, and n_a refers to the number of attributes

In the case to represent the output of a model, we use a subscript to denote the time point of the input data, and use a superscript to denote the time point at which the model is learned or needs to be learned. For example, $\mathbf{x}_t^{t-1}$ refers to the numerical representation of an object observed at time t generated by the model learned at time $t-1$, and $\mathcal{M}_t^t$ refers to the coupling spaces of observations at time t learned by coupling learning functions at time t.

The specifications for symbol styles in this paper are as follows. Value: lowercase; vector: lowercase with bold font; matrix: uppercase with bold font; set: uppercase; function: lowercase with parentheses; space: uppercase with Calligraphic font.

The main notations in this paper are defined in Table 1 at the end of this material.

Supplementary Technical Detail

Illustrating the UNICORN Learning Process Below, we illustrate the UNICORN learning process, which is described in the Section *UNICORN Working Mechanism*, in terms of two time points t and $t-1$. At time t, as shown at the right column of Figure 2, UNICORN learns a model $UNICORN_t$ in terms of multiple steps. First, as shown in Section *Heterogeneous Coupling Learning*, heterogeneous couplings hidden in the data E_t are learned w.r.t. different coupling learning functions, e.g., intra-attribute couplings $m_{Ia,t}^{(j)}(v^{(j)})$ as shown in Eq. (1) (Eq. (1) in main manuscript) and the inter-attribute couplings $m_{Ie,t}^{(j)}(v^{(j)})$ as shown in Eq. (2) (Eq. (7) in main manuscript) for the categorical value $v^{(j)}$ of attribute a_j. As a result, various coupling spaces are generated, e.g., the intra-attribute coupling spaces $\mathcal{M}_{Ia,t}$ (see Eq. (4), i.e. Eq. (3) in main manuscript), inter-attribute coupling spaces $\mathcal{M}_{Ie,t}$ (Eq. (5), i.e. Eq. (9) in main manuscript), and the aggregated entire attribute coupling spaces $\mathcal{M}_t$ (Eq. (6), i.e. Eq. (10) in main manuscript), Second, following the kernelization procedure described in Section *Transforming Heterogeneous Couplings to Kernelized Spaces*, we transform the learned coupling spaces to kernelized spaces $\mathcal{K}_t$. For example, Eq. (7) (Eq. (11) in main manuscript) illustrates how to construct a kernel space from a coupling space. Third, we further unify all learned kernel spaces to a global kernel space in terms of the approach described in Section *Building a Unified Global Representation*, as illustrated in Eq. (8) (Eq. (16) in main manuscript).

Consequently, as the result of undertaking the above UNICORN learning process, we obtain two UNICORN models, i.e, $UNICORN_{t-1}$ for $t-1$ and $UNICORN_t$ for t respectively, which consist of respective coupling spaces $\mathcal{M}_t^{t-1}$ and $\mathcal{M}_t^t$. These spaces further correspond to the unified kernel representations $\mathbf{S}_t^{t-1}$ and $\mathbf{S}_t^t$. To capture the data dynamics between times $t-1$ and t, we further involve the parameters learned for $UNICORN_{t-1}$ as a prior of learning $UNICORN_t$. The representation $\mathbf{S}_t^t$ at t on data E_t needs to be aligned with the representation $\mathbf{S}_t^{t-1}$ from $UNICORN_{t-1}$ on data E_t in terms of the degree ξ_t. ξ_t is calculated by comparing the difference between $\mathcal{M}_t^{t-1}$ and $\mathcal{M}_t^t$. Lastly, UNICORN maximizes the informativeness of $\mathbf{S}_t^t$ for time t, as shown in Eq. (9) (Eq. (17) in main manuscript).

$$m_{Ia,t}^{(j)}(v^{(j)}) = \begin{cases} [\frac{|g_t^{(j)}(v^{(j)})|}{n_{o_t}}] & t=1 \\ \frac{m_{Ia,t-1}^{(j)} \sum_{l=1}^{t-1} n_{o_l} + [g_t^{(j)}(v^{(j)})]}{\sum_{l=1}^{t-1} n_{o_l} + n_{o_t}} & t>1 \end{cases}, \quad (1)$$

$$m_{Ie,t}^{(j)}(v^{(j)}) = \left[\ p_t(v^{(j)}|v_{*1}), \quad \cdots, \quad p_t(v^{(j)}|v_{*|V_{*,t}|})\ \right]^\top, \quad (2)$$

$$\mathcal{M}_{Ia,t}^{(j)} = \{m_{Ia,t}^{(j)}(v^{(j)}) | v^{(j)} \in V_t^{(j)}\}. \quad (3)$$

$$\mathcal{M}_{Ia,t} = \{\mathcal{M}_{Ia,t}^{(1)}, \cdots, \mathcal{M}_{Ia,t}^{(n_a)}\}. \quad (4)$$

$$\mathcal{M}_{Ie,t} = \{\mathcal{M}_{Ie,t}^{(1)}, \cdots, \mathcal{M}_{Ie,t}^{(n_a)}\}. \quad (5)$$

$$\mathcal{M}_t = \mathcal{M}_{Ia,t} \cup \mathcal{M}_{Ie,t}. \quad (6)$$

$$\mathbf{K}_{t,p} = \begin{bmatrix} k_p(\mathsf{m}_{t,1}, \mathsf{m}_{t,1}) & \cdots & k_p(\mathsf{m}_{t,1}, \mathsf{m}_{t,n_{v_t^*}}) \\ k_p(\mathsf{m}_{t,2}, \mathsf{m}_{t,1}) & \cdots & k_p(\mathsf{m}_{t,2}, \mathsf{m}_{t,n_{v_t^*}}) \\ \vdots & \ddots & \vdots \\ k_p(\mathsf{m}_{t,n_{v_t^*}}, \mathsf{m}_{t,1}) & \cdots & k_p(\mathsf{m}_{t,n_{v_t^*}}, \mathsf{m}_{t,n_{v_t^*}}) \end{bmatrix}, \quad (7)$$

$$\mathbf{S}_{t,ij} = \sum_{p=1}^{n_k} \mathbf{K}_{t,p,\mathrm{i}\cdot}^\top \boldsymbol{\omega}_{t,p} \mathbf{K}_{t,p,\mathrm{j}\cdot}. \quad (8)$$

$$\begin{aligned} \underset{\boldsymbol{\omega}_t}{\text{minimize}} \quad & \frac{n_{o_t}(1-\bar{s_t})}{\|\mathbf{H}\mathbf{S}_t^t\mathbf{H}\|_F \sqrt{n_{o_t}-1}} \\ & - \delta\xi_t \frac{\mathbf{H}\mathbf{S}_t^t\mathbf{H}\mathbf{H}\mathbf{S}_t^{t-1}\mathbf{H}}{\|\mathbf{H}\mathbf{S}_t^t\mathbf{H}\|_F \|\mathbf{H}\mathbf{S}_t^{t-1}\mathbf{H}\|_F} + \sigma\|\boldsymbol{\omega}_t\|_1, \end{aligned} \quad (9)$$

The Construction of Transformation Matrix The transformation matrix for the p-th kernel space $\mathcal{K}_{t,p}$ at time t is constructed as follows,

$$\mathbf{T}_{t,p} = \begin{bmatrix} \alpha_{t,p1} & 0 & \cdots & 0 \\ 0 & \alpha_{t,p2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \alpha_{t,pn_v} \end{bmatrix}, \quad (10)$$

where parameter $\alpha_{t,pi}$, which need to be learned, is the weight of the i-th categorical value (corresponding to the i-th row) in the p-th kernel space at time t.

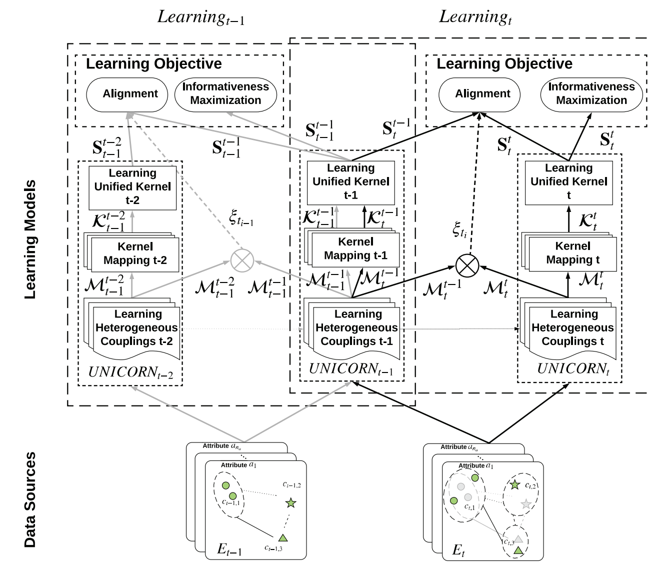


Figure 2: The UNICORN Architecture: A multi-step dynamic learning process to capture and convert local heterogeneous couplings to unified global representations on dynamic categorical data.

Supplementary Experiments

Critical Difference Comparison

We show the critical difference comparison of the overall performance of UNICORN-enabled K-Means Clustering against the state-of-the-art categorical data stream clustering methods by 3 per Bonferroni-Dunn Test (Demšar 2006). The results show the performance of UNICORN-enabled k-means clustering is significantly ($p < 0.05$) better than that of the state-of-the-art categorical data stream clustering methods.

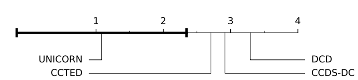
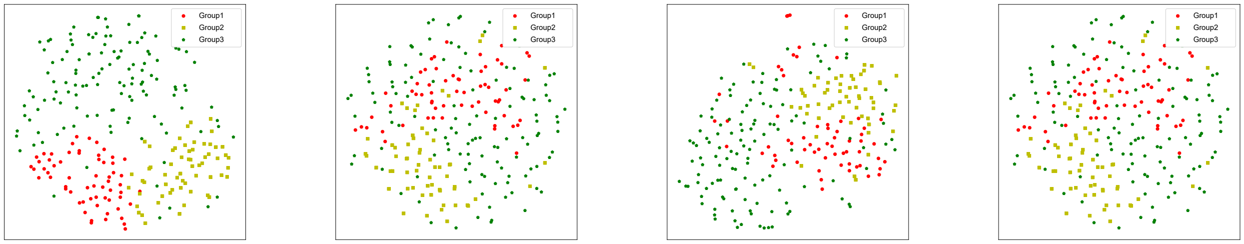


Figure 3: Comparison of UNICORN-enabled K-Means Clustering against Categorical Data Stream Clustering Methods per the Bonferroni-Dunn Test. All clustering methods with ranks outside the marked interval are significantly different ($p < 0.05$) from UNICORN-enabled k-mean.

Visualization

To further illustrate the value of representing hierarchical heterogeneous couplings in dynamic categorical data, we visualize the representation results generated by different methods. Figure 4 displays the outcomes w.r.t. a two-dimensional space by the t-SNE method (Maaten and Hinton 2008) converted from the high-dimensional representations made by each method on DNANominal data set. In comparison, objects belonging to the same group are more likely co-located with others. It also quantitatively demonstrates that the UNICORN-represented results are more suitable for general learning tasks.



(a) UNICORN (b) CDE (c) COS (d) DILCA

Figure 4: The Visualization of Different Representation Methods: A better representation yields closer grouping results.

Stability Test

We evaluate UNICORN stability with regard to its hyper-parameter δ, which is a hyper-parameter that controls the impact of kernel alignment (the second term in the objective function Eq. (9)). The larger δ is, the more historical information carried forward to current learning process. Figure 5 shows the UNICORN-enabled clustering F-score under different settings of δ on the Led24 data set. With the value of δ increasing, the F-score increases from 0.626 to 0.635. This small increasing range indicates UNICORN is stable w.r.t. δ. Meanwhile, the F-score increasing trend also demonstrates the importance of carrying forward historical heterogeneous couplings.

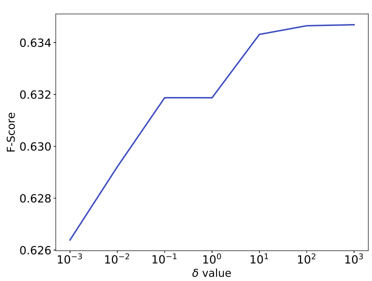


Figure 5: The UNICORN-enabled Clustering F-score w.r.t. hyper-parameter δ.

References

Ahmad, A., and Dey, L. 2007. A method to compute distance between two categorical values of same attribute in unsupervised learning for categorical data set. *Pattern Recognition Letters* 28(1):110–118.

Bai, L.; Cheng, X.; Liang, J.; and Shen, H. 2016. An optimization model for clustering categorical data streams with drifting concepts. *IEEE Transactions on Knowledge and Data Engineering* 28(11):2871–2883.

Barbará, D.; Li, Y.; and Couto, J. 2002. Coolcat: an entropy-based algorithm for categorical clustering. In *Proceedings of the eleventh international conference on Information and knowledge management*, 582–589. ACM.

Bengio, Y.; Courville, A.; and Vincent, P. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence* 35(8):1798–1828.

Boriah, S.; Chandola, V.; and Kumar, V. 2008. Similarity measures for categorical data: A comparative evaluation. In *SDM*, 243–254. SIAM.

Cao, F.; Liang, J.; Bai, L.; Zhao, X.; and Dang, C. 2010. A framework for clustering categorical time-evolving data. *IEEE Transactions on Fuzzy Systems* 18(5):872–882.

Cao, F.; Liang, J.; Li, D.; Bai, L.; and Dang, C. 2012. A dissimilarity measure for the k-modes clustering algorithm. *Knowledge-Based Systems* 26:120–127.

Cao, L.; Ou, Y.; and Yu, P. S. 2012. Coupled behavior analysis with applications. *IEEE Transactions on Knowledge and Data Engineering* 24(8):1378–1392.

Cao, L. 2015. Coupling learning of complex interactions. *Information Processing & Management* 51(2):167–186.

Chen, H.-L.; Chen, M.-S.; and Lin, S.-C. 2009. Catching the trend: A framework for clustering concept-drifting categorical data. *IEEE transactions on knowledge and data engineering* 21(5):652–665.

Cheung, Y.-M., and Jia, H. 2013. Categorical-and-numerical-attribute data clustering based on a unified similarity metric without knowing cluster number. *Pattern Recognition* 46(8):2228–2238.

Demšar, J. 2006. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research* 7(Jan):1–30.

Do, T. D. T., and Cao, L. 2018. Coupled poisson factorization integrated with user/item metadata for modeling popular and sparse ratings in scalable recommendation. *AAAI2018* 1–7.

DEnza, A. I., and Markos, A. 2015. Incremental visualization of categorical data. In *Advances in Statistical Models for Data Analysis*. Springer. 137–148.

Ienco, D.; Pensa, R. G.; and Meo, R. 2012. From context to distance: Learning dissimilarity for categorical data clustering. *ACM TKDD* 6(1):1–25.

Jian, S.; Cao, L.; Pang, G.; Lu, K.; and Gao, H. 2017. Embedding-based representation of categorical data by hierarchical value coupling learning. In *IJCAI*, 1937–1943.

Li, Y.; Li, D.; Wang, S.; and Zhai, Y. 2014. Incremental entropy-based clustering on categorical data streams with concept drift. *Knowledge-Based Systems* 59:33–47.

Maaten, L. v. d., and Hinton, G. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research* 9(Nov):2579–2605.

Ng, M. K.; Li, M. J.; Huang, J. Z.; and He, Z. 2007. On the impact of dissimilarity measure in k-modes clustering algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29(3):503–507.

Pang, G.; Cao, L.; Chen, L.; and Liu, H. 2017. Learning homophily couplings from non-iid data for joint feature selection and noise-resilient outlier detection. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 2585–2591. AAAI Press.

Saha, I., and Maulik, U. 2014. Incremental learning based multiobjective fuzzy clustering for categorical data. *Information Sciences* 267:35–57.

Shi, Y.; Li, W.; Gao, Y.; Cao, L.; and Shen, D. 2017. Beyond iid: Learning to combine non-iid metrics for vision tasks. In *AAAI*, 1524–1531.

Wang, C.; Dong, X.; Zhou, F.; Cao, L.; and Chi, C.-H. 2015. Coupled attribute similarity learning on categorical data. *IEEE TNNLS* 26(4):781–797.

Zhang, Q.; Cao, L.; Zhu, C.; Li, Z.; and Sun, J. 2018. Coupledcf: Learning explicit and implicit user-item couplings in recommendation for deep collaborative filtering. In *IJCAI*, 3662–3668.

Zhu, C.; Cao, L.; Liu, Q.; Yin, J.; and Kumar, V. 2018. Heterogeneous metric learning of categorical data with hierarchical couplings. *IEEE Transactions on Knowledge and Data Engineering* 30(7):1254–1267.

Table 1: List of Notations

Symbol	Meaning
$1, \cdots, t, \cdots, n_t$	The time points from 1 to n_t
E_t	The observed data set at time t
O_t	The set of objects observed at time t
A	The set of attributes
V_t	The set of categorical values observed at time t
$o_{t,i}$	The i-th object in O_t
a_j	The j-th attribute in A
$V_t^{(j)}$	The set of values of j-th attribute at time t
$v_{t,i}^{(j)}$	The i-th value in the j-th attribute at time t
n_t	The number of time points
n_{o_t}	The number of objects at time t
n_a	The number of attributes
n_{v_t}	The number of values at time t
$n_{v_t^{(j)}}$	The number of values in the j-th attribute at time t
$n_{v_t^*}$	The number of categorical values represented in the coupling space $\mathcal{M}_{t,j}$ at time t
n_k	The number of kernel spaces
$\mathbf{x}_{t,i}$	The vector representation of the i-th object at time t
$\mathbf{S}_i^j$	The similarity representation of E_i generated by the models learned at time j
$\mathcal{M}_i^j$	The coupling spaces of observations at time i generated by the model learned at time j
$\mathcal{M}_{Ia,t}^{(j)}$	The intra-attribute coupling space of observations at time t
$\mathcal{M}_{Ie,t}^{(j)}$	The inter-attribute coupling space of observations at time t
$m_{Ia,t}^{(j)}(\cdot)$	The intra-attribute coupling function for the j-th attribute at time t
$m_{Ie,t}^{(j)}(\cdot)$	The inter-attribute coupling function for the j-th attribute at time t
$\mathsf{m}_{t,i}$	A vector of i-th categorical value in a coupling space at time t
$p_t(v^{(j)}\|v^{(k)})$	The information conditional probability function for values $v^{(j)}$ and $v^{(k)}$ at time t
$g_t^{(j)}(v^{(j)})$	A function that returns objects which have value $v^{(j)}$ in j-th attribute at time t
$k_p(\cdot,\cdot)$	The kernel function of the p-th kernel space
$\mathcal{K}_{t,p}$	The p-th kernel space at time t
$\mathcal{K}'_{t,p}$	The p-th heterogeneous kernel space at time t
$\mathbf{K}_{t,p}$	The kernel matrix of the p-th kernel space at time t
$\mathbf{K}'_{t,p}$	The kernel matrix of the p-th heterogeneous kernel space at time t
$\mathbf{T}_{t,p}$	The transform matrix for $\mathbf{K}_{t,p}$ at time t
$\mathbf{S}_t$	The similarity representation of observed categorical data at time t
$\mathbf{S}_{t,p,ij}$	The similarity of the i-th and j-th objects in the p-th heterogeneous coupling space
$\mathbf{S}_{t,ij}$	The similarity of the i-th and j-th objects in the heterogeneous coupling space
ξ_t	The coupling difference degree between t and $t-1$
$\alpha_{t,pi}$	The weight of the i-th value in the p-th kernel space at time t
$\beta_{t,p}$	The weight of the p-th base similarity at time t
$\boldsymbol{\omega}_t$	The heterogeneity parameter at time t
$\boldsymbol{\omega}_{t,p}$	The diagonal heterogeneity parameter matrix for the p-th base similarity
$\boldsymbol{\omega}_{t,p,ij}$	The (i,j)-th entry of $\boldsymbol{\omega}_{t,p}$
δ	The trade-off parameter for centered kernel alignment
σ	The trade-off parameter for sparse regularization for heterogeneity parameter
$\mathbf{I}$	The identity matrix