Noname manuscript No.
(will be inserted by the editor)

Unsupervised Group Matching with Application to Cross-lingual Topic Matching without Alignment Information

Tomoharu Iwata · Motonobu Kanagawa · Tsutomu Hirao · Kenji Fukumizu

Received: date / Accepted: date

Abstract We propose a method for unsupervised group matching, which is the task of finding correspondence between groups across different domains without cross-domain similarity measurements or paired data. For example, the proposed method can find matching of topic categories in different languages without alignment information. The proposed method interprets a group as a probability distribution, which enables us to handle uncertainty in a limited amount of data, and to incorporate the high order information on groups. Groups are matched by maximizing the dependence between distributions, in which we use the Hilbert Schmidt independence criterion for measuring the dependence. By using kernel embedding which maps distributions into a reproducing kernel Hilbert space, we can calculate the dependence between distributions without density estimation. In the experiments, we demonstrate the effectiveness of the proposed method using synthetic and real data sets including an application to cross-lingual topic matching.

Keywords Unsupervised object matching, Kernel embedding of distributions, Multilingual corpus analysis

T. Iwata
NTT Communication Science Laboratories
2-4 Hikaridai, Seikacho, Sorakugun, Kyoto, 619-0237, Japan
Tel.: +81-774-93-5161
Fax: +81-774-93-5155
E-mail: iwata.tomoharu@lab.ntt.co.jp

M. Kanagawa
The Institute of Statistical Mathematics, Japan

T. Hirao
NTT Communication Science Laboratories, Japan

K. Fukumizu
The Institute of Statistical Mathematics, Japan

1 Introduction

Object matching is an important task in natural language processing, machine learning, data mining, image processing, bioinformatics, and so on. Examples of object matching include matching an image with a caption (Socher and Fei-Fei, 2010), an English word with a German word (Tripathi et al, 2010), concepts in different ontologies (Shvaiko and Euzenat, 2013), and user identification in different databases (Li et al, 2009). Most object matching methods require similarity measurements between objects in different domains, or paired data across domains that contain correspondence information. However, similarity measurements and paired data are unavailable in some applications because obtaining those information would incur a cost and require time, or invade privacy. For such situations, a number of methods for unsupervised object matching have been proposed, such as kernelized sorting (Quadrianto et al, 2010) and matching canonical correlation analysis (Haghighi et al, 2008).

In this paper, we consider a related but different task, which we call *unsupervised group matching*. Here, a group consists of a set of objects. In many data sets, objects form a group; documents are categorized according to their topics, users form communities, movies are associated with genres, and images are grouped by contents. Unsupervised group matching is the task of finding correspondence between groups given two sets of groups in different domains without cross-domain similarity measurements or paired data. The task appears in a wide variety of applications for data with groups. For instance, group matching can be used for matching topic categories in different languages given a multilingual corpora without dictionaries and parallel corpora, which is important especially for resource poor languages. Other examples include matching groups of images and documents according to their topics (Barnard et al, 2003), matching user communities and movie genres by their preferences (Kamahara et al, 2005), discovering companies in the same position across different countries by representing the companies with their products, and finding correspondence between groups of genes in protein-protein interaction networks of different species (Terada and Sese, 2012).

This paper proposes a kernel-based sorting method for unsupervised group matching. The main idea of the proposed method is to interpret a group as a probability distribution, and consider each object in a group as a sample from the underlying distribution. Then, our problem is transformed to finding a matching between sets of probability distributions in different domains. Our method consists of two-steps (Figure 1). First, we map each distribution (group) into a reproducing kernel Hilbert space (RKHS) defined on a distinct domain, based on the framework of *kernel embedding of distributions* (Smola et al, 2007). The kernel representation of distributions enables us to maintain and handle necessary information of groups such as their covariance or higher-order moments. Then we find a matching between the sets of elements in the different RKHSs that maximizes the dependency between them. As a dependence measure, we employ the Hilbert-Schmidt independence criterion (HSIC) (Gretton et al, 2005), which can be used for unsupervised matching tasks once appropriate kernels are defined for each domain. To apply HSIC, we define second-level kernels on each RKHS, for which we conduct necessary theoretical analysis in the paper. Our

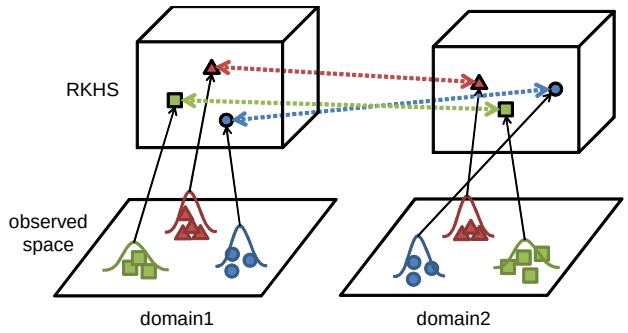


Fig. 1 Framework of the proposed method. The input is two sets of grouped objects, where an object is represented by a point and the group is represented by its color in the observation space. The distribution for each group is mapped into a reproducing kernel Hilbert space (RKHS) for each domain, where the mapped distributions are represented by points in the RKHS. The groups in different domains are matched based on the Hilbert Schmidt independence criterion of the mapped distributions.

approach only requires similarity between objects within a distinct domain, and thus does not need a similarity measure between objects across domains.

To our knowledge, this is the first study for unsupervised group matching. One might think that a naive extension of unsupervised object matching methods could be used for matching groups, where each group is represented by its typical value, such as mean and mode. However, in the process of transforming a group to a feature, information loss is inevitable in general. For example, when a group is represented by its mean, variance and skewness of the group cannot be considered in matching. On the other hand, because the proposed method represents a group by its distribution using kernel embedding, it can preserve necessary information of distributions. Another naive method for group matching is that objects are first aligned using an object matching method, and then groups are matched using the object alignment information. This method, however, requires much more computational time than the proposed one: the time complexity of the naive method is cubic to the number of objects, while that of the proposed method is cubic to the number of groups. Since the number of groups is much smaller than the number of objects, the proposed method can efficiently find group matching. By handling a set of objects as a group instead of as individual objects, we can reduce the computational time.

The remainder of this paper is organized as follows. In Section 2, we briefly review related work on unsupervised group matching. In Section 3, we formulate the proposed method, which calculates the dependency between distributions using the HSIC and kernel embedding. In Section 4, we give theoretical results on the proposed method. In Section 5, we demonstrate the effectiveness of the proposed method by using synthetic and real data, which include multilingual categorized documents for an application of category matching in different languages. Finally, we present concluding remarks and a discussion of future work in Section 6.

2 Related Work

There have been proposed a number of unsupervised object matching methods, such as kernelized sorting (Quadrianto et al, 2010), convex kernelized sorting (Djuric et al, 2012), matching canonical correlation analysis (Haghighi et al, 2008), least-squares object matching (Yamada and Sugiyama, 2011), variational Bayesian matching (Klami, 2012), and many-to-many matching latent variable models (Iwata et al, 2013). Their applications include aligning multilingual documents, visualizing data in a particular structure, and matching images. These methods find correspondence by maximizing the dependency between matched pairs. Intuitively, they match objects that have similar neighborhood relationships. For example, kernelized sorting uses the HSIC for measuring the dependency that is calculated based on kernel matrices for objects within each domain. These methods, however, are designed for object matching, and thus cannot used for the group matching problem straightforwardly.

Kernel embedding has been used for extending kernel methods to grouped data. For example, the support measure machine (Muandet et al, 2012) is a method for kernel based discriminative learning on distributions, which generalizes the support vector machine by kernel embedding. The one-class support measure machine (Muandet and Schölkopf, 2013) is a group anomaly detection method that finds anomalous aggregated behaviors of objects. The support measure machine and the one-class support measure machine achieved high performance with classification and anomaly detection tasks, respectively, by assuming that a group is modeled as a distribution. The proposed method is a generalization of kernelized sorting for grouped data.

Group matching is related to ontology matching (Doan et al, 2004). Given two taxonomies, ontology matching methods find the most similar concept node in the other taxonomy. Although there have been proposed a number of methods for ontology matching (Shvaiko and Euzenat, 2013), they are not unsupervised methods. For example, they require that there are objects that appear in both domains, features are shared across different domains which enables us to calculate similarities, or graph structures of the given ontologies are used for matching.

3 Proposed Method

3.1 Group matching

Suppose that we are given two grouped data sets from different domains $\{\boldsymbol{X}_1, \cdots, \boldsymbol{X}_N\}$ and $\{\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_N\}$ with equal size N of groups. Here, $\boldsymbol{X}_n = \{x_{n1}, \cdots, x_{nI_n}\}$ is a set of objects in the nth group in the first domain, $x_{ni} \in \mathcal{X}$ is the ith object in the nth group in the first domain, $\boldsymbol{Y}_n = \{y_{n1}, \cdots, y_{nJ_n}\}$ is a set of objects in the nth group in the second domain, and $y_{nj} \in \mathcal{Y}$ is the jth object in the nth group in the second domain. The task is to find correspondence between groups across the first and second domains. We assume that similarity, or kernel, can be calculated within a domain, but similarity across different domains is not given. The correspondence between groups, or between objects, over different domains are unavailable. The number of objects in a group can be different over groups.

We represent each group, or set of objects, by a probability distribution assuming that observed objects are generated from an unknown distribution. Let $\mathbb{P}_n$ is a distribution of the nth group in the first domain, $\boldsymbol{X}_n \sim \mathbb{P}_n$, and $\mathbb{Q}_n$ is a distribution of the nth group in the second domain, $\boldsymbol{Y}_n \sim \mathbb{Q}_n$. Then, the task of group matching is considered as the task of matching two sets of distributions, $\{\mathbb{P}_1, \cdots, \mathbb{P}_N\}$ and $\{\mathbb{Q}_1, \cdots, \mathbb{Q}_N\}$. The cross-domain matching is encoded in a permutation matrix $\pi \in \mathit{\Pi}_N$, where

$$\mathit{\Pi}_N := \left\{\pi | \pi \in \{0,1\}^{N \times N}, \pi 1_N = 1_N, \pi^\top 1_N = 1_N\right\}, \tag{1}$$

and 1_N is the N dimensional vector of all ones. We find correspondence by maximizing the pairwise dependency as follows,

$$\arg\max_{\pi \in \mathit{\Pi}_N} D\left(\{(\mathbb{P}_n, \mathbb{Q}_{\pi(n)})\}_{n=1}^N\right), \tag{2}$$

where $D(\cdot)$ is a measurement for the dependency.

3.2 Dependency between distributions

For the dependency measure between distributions, we use the Hilbert-Schmidt information criterion (HSIC) (Gretton et al, 2005). The HSIC is successfully used for unsupervised object matching methods, such as kernelized sorting (Quadrianto et al, 2010; Djuric et al, 2012). However, these methods use HSIC for dependency between objects. In our task, the HSIC for dependency between distributions is required. For obtaining the HSIC for distributions, we map distributions into the reproducing kernel Hilbert space (RKHS) using the *kernel embedding* (Smola et al, 2007). The kernel embedding allows us to calculate the HSIC without density estimation while preserving necessary information of distributions.

3.2.1 Representation of distributions using kernel embeddings

Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a reproducing kernel on $\mathcal{X}$. Then it is known that there exists a *reproducing kernel Hilbert space (RKHS)* $\mathcal{H}_k$ uniquely associated with k. The RKHS $\mathcal{H}_k$ consists of functions on $\mathcal{X}$, e.g. the function $k(\cdot, x)$, which is called the *feature vector* of $x \in \mathcal{X}$, is included in $\mathcal{H}_\mathcal{X}$.

Then we represent any distribution P on $\mathcal{X}$ by its expectation of the feature vector (Smola et al, 2007):

$$\mu_\mathbb{P} := \mathbf{E}_{x \sim \mathbb{P}}[k(\cdot, x)] = \int_\mathcal{X} k(\cdot, x) d\mathbb{P}(x) \in \mathcal{H}_k, \tag{3}$$

which is an element in the RKHS, and called the *kernel embedding* of P.

Then by interpreting each group $\boldsymbol{X}_n = \{x_{ni}\}_{i=1}^{I_n}$ as an empirical distribution

$$\mathbb{P}_n := \frac{1}{I_n} \sum_{i=1}^{I_n} \delta_{x_{ni}}(\cdot), \tag{4}$$

where δ_x is the Dirac delta function at point $x \in \mathcal{X}$, the representation of the nth group is given by

$$\mu_{\mathbb{P}_n} = \frac{1}{I_n} \sum_{i=1}^{I_n} k(\cdot, x_{ni}) \in \mathcal{H}_k. \tag{5}$$

When samples $\boldsymbol{X}_n$ are generated i.i.d. from a distribution $\mathbb{P}_n^*$, the convergence rate of the empirical mean $\mu_{\mathbb{P}_n}$ to the expectation $\mu_{\mathbb{P}_n^*}$ (Smola et al, 2007) is

$$\|\mu_{\mathbb{P}_n^*} - \mu_{\mathbb{P}_n}\|_{\mathcal{H}_k} = O_p(I_n^{-\frac{1}{2}}). \tag{6}$$

Let us consider two groups of samples generated from distributions $\mathbb{P}_n^*$ and $\mathbb{P}_m^*$. The distance between the empirical mean of the two groups $\|\mu_{\mathbb{P}_n} - \mu_{\mathbb{P}_m}\|_{\mathcal{H}_k}$ converges to that of the true distributions $\|\mu_{\mathbb{P}_n^*} - \mu_{\mathbb{P}_m^*}\|_{\mathcal{H}_k}$ with the rate of $O_p(\min(I_n, I_m)^{-\frac{1}{2}})$ (Gretton et al, 2012b). In other words, we can detect the difference between groups with this rate.

Similarly, we represent the groups on $\mathcal{Y}$ using the kernel embedding. Let $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be a reproducing kernel on $\mathcal{Y}$, and $\mathcal{H}_\ell$ be the associated RKHS. Then the nth group $\boldsymbol{Y}_n = \{y_{nj}\}_{j=1}^{J_n}$ is represented as the embedding of the empirical distribution $\mathbb{Q}_n := \frac{1}{J_n} \sum_{i=1}^{J_n} \delta_{y_{nj}}(\cdot)$:

$$\mu_{\mathbb{Q}_n} = \frac{1}{J_n} \sum_{j=1}^{J_n} \ell(\cdot, y_{nj}) \in \mathcal{H}_\ell. \tag{7}$$

If the samples are i.i.d., its convergence rate is $O_p(J_n^{-\frac{1}{2}})$ as in (6).

Advantages of using the kernel embedding for the representation of groups are that 1) we do not need to prespecify the number of members in each group, 2) we can measure the distance, or similarity in terms of level-2 kernels introduced below, between groups, 3) we can deal with any domain of objects once we define a kernel, e.g. documents or images, and 4) we can capture the properties of each group described as a probability distribution, such as covariance structure of the group.

The property 4) is of special importance: for example, consider the case where there are different groups $\boldsymbol{X}_n$ and $\boldsymbol{X}_m$ which share the same point as a mean, but have different covariance structures. Then we cannot distinguish these groups if we naively represent them with their shared mean value. Thus it is important that we can distinguish the groups by their higher order moments.

It is known that the order of moments to which the kernel embedding representation can distinguish is determined by the kernel k (Fukumizu et al, 2004; Smola et al, 2007; Sriperumbudur et al, 2010). For example, if we use a polynomial kernel

$$k(x, x') = (\langle x, x' \rangle + c)^d, \tag{8}$$

of degree $d \in \mathbb{N}$, we can represent distributions up to their d-th moments using the kernel embedding. On the other hand, there exists a family of kernels that can distinguish any pair of distributions, such as the Gaussian kernel

$$k(x, x') = \exp\left(-\frac{\gamma}{2}\|x - x'\|^2\right), \tag{9}$$

with $\gamma > 0$. Such kernels are called *characteristic* [1] (Fukumizu et al, 2008; Sriperumbudur et al, 2010).

3.2.2 HSIC on embedded distributions

Given the representation of the groups (5, 7), the maximization problem (2) now turns to be

$$\arg\max_{\pi \in \Pi_N} D\left(\{(\mu_{\mathbb{P}_n}, \mu_{\mathbb{Q}_{\pi(n)}})\}_{n=1}^N\right). \tag{10}$$

Let $\mathcal{P}$ and $\mathcal{Q}$ be the sets of all probability distributions on $\mathcal{X}$ and $\mathcal{Y}$, respectively. Let $\mathcal{F}_\mathcal{P} := \{\mu_\mathbb{P} : \mathbb{P} \in \mathcal{P}\} \subset \mathcal{H}_k$ and $\mathcal{F}_\mathcal{Q} := \{\mu_\mathbb{Q} : \mathbb{Q} \in \mathcal{Q}\} \subset \mathcal{H}_\ell$ be the sets of all the embedded distributions. Then what we next do is to define the dependency measure D on $(\mathcal{F}_\mathcal{P}, \mathcal{F}_\mathcal{Q})$. To this end, we apply HSIC to the sets of kernel embeddings.

Without loss of generality, assume that we would like to quantify the amount of dependency of the matching $\{(\mu_{\mathbb{P}_n}, \mu_{\mathbb{Q}_n})\}_{n=1}^N$. HSIC measures the dependency of the set of pairs $\{(\mu_{\mathbb{P}_n}, \mu_{\mathbb{Q}_n})\}_{n=1}^N$ by 1) first assuming that they are samples from some joint distribution $\Pr(\mu_\mathbb{P}, \mu_\mathbb{Q})$ on $\mathcal{F}_\mathcal{P} \times \mathcal{F}_\mathcal{Q}$ and 2) then estimating the distance between the joint distribution $\Pr(\mu_\mathbb{P}, \mu_\mathbb{Q})$ and the product of their marginal distributions $\Pr(\mu_\mathbb{P})$ and $\Pr(\mu_\mathbb{Q})$ using the samples $\{(\mu_{\mathbb{P}_n}, \mu_{\mathbb{Q}_n})\}_{n=1}^N$.

HSIC also represents the distributions on $\mathcal{F}_\mathcal{P} \times \mathcal{F}_\mathcal{Q}$ using the kernel embedding. Thus, we need to additionally define kernels on $\mathcal{F}_\mathcal{P}$ and $\mathcal{F}_\mathcal{Q}$. We denote these kernels in capitals; let K and L be reproducing kernels on $\mathcal{F}_\mathcal{P}$ and $\mathcal{F}_\mathcal{Q}$, respectively. We will call these kernels *level-2*, in contrast to the kernels k and ℓ used for the representation of groups in the first level. We will show concrete examples of the level-2 kernels later in Section 3.4.

Let $\mathcal{H}_K$ and $\mathcal{H}_L$ be the RKHSs associated with the kernels K and L, respectively. Then we can define an RKHS on $\mathcal{F}_\mathcal{P} \times \mathcal{F}_\mathcal{Q}$ by their tensor product $\mathcal{H}_K \otimes \mathcal{H}_L$, which is equivalent to the RKHS associated with the joint kernel $K \otimes L$ on $\mathcal{F}_\mathcal{P} \times \mathcal{F}_\mathcal{Q}$ defined by

$$K \otimes L((\mu_\mathbb{P}, \mu'_\mathbb{P}), (\mu_\mathbb{Q}, \mu'_\mathbb{Q})) := K(\mu_\mathbb{P}, \mu'_\mathbb{P}) L(\mu_\mathbb{Q}, \mu'_\mathbb{Q}). \tag{11}$$

Using the joint kernel, we represent the joint distribution $\Pr(\mu_\mathbb{P}, \mu_\mathbb{Q})$ as the kernel embedding $\mathbf{E}_{\mu_\mathbb{P}, \mu_\mathbb{Q}}[K(\cdot, \mu_\mathbb{P}) \otimes L(\cdot, \mu_\mathbb{Q})]$ into $\mathcal{H}_K \otimes \mathcal{H}_L$. Likewise, the product of the marginal distributions $\Pr(\mu_\mathbb{P})$ and $\Pr(\mu_\mathbb{Q})$ is embedded as $\mathbf{E}_{\mu_\mathbb{P}}[K(\cdot, \mu_\mathbb{P})] \otimes \mathbf{E}_{\mu_\mathbb{Q}}[L(\cdot, \mu_\mathbb{Q})]$. Then HSIC (in population) is defined as the distance between these embeddings (Smola et al, 2007):

$$\|\mathbf{E}_{\mu_\mathbb{P}}[K(\cdot, \mu_\mathbb{P})] \otimes \mathbf{E}_{\mu_\mathbb{Q}}[L(\cdot, \mu_\mathbb{Q})] - \mathbf{E}_{\mu_\mathbb{P}, \mu_\mathbb{Q}}[K(\cdot, \mu_\mathbb{P}) \otimes L(\cdot, \mu_\mathbb{Q})]\|^2_{\mathcal{H}_K \otimes \mathcal{H}_L} \geq 0, \tag{12}$$

where $\|\cdot\|_{\mathcal{H}_K \otimes \mathcal{H}_L}$ denotes the norm of $\mathcal{H}_K \otimes H_L$.

[1] More precisely, kernel k is called characteristic if the map $\mathcal{P} \to \mathcal{H}_k : \mathbb{P} \to \mu_\mathbb{P} := \int k(\cdot, x) d\mathbb{P}(x)$ is injective. Thus, if we use a characteristic kernel, then the embedding $\mu_\mathbb{P}$ uniquely identifies the underling distribution $\mathbb{P}$.

It immediately follows that when the joint kernel $K \otimes L$ is characteristic, i.e. the embedding of distributions into $\mathcal{H}_K \otimes \mathcal{H}_L$ is injective, then HSIC (12) is equal to zero if and only if $\mu_{\mathbb{P}}$ and $\mu_{\mathbb{Q}}$ are independent. In other words, a large value of HSIC indicates that there is a strong dependency between $\mu_{\mathbb{P}}$ and $\mu_{\mathbb{Q}}$. In Section 4.2, we will show that the joint kernel is characteristic when the level-2 kernels K and L are given as Gaussian kernels.

3.3 Group kernelized sorting

Given the matching $\{(\mu_{\mathbb{P}_n}, \mu_{\mathbb{Q}_n})\}_{n=1}^N$, which can be seen as samples from $\Pr(\mu_{\mathbb{P}}, \mu_{\mathbb{Q}})$, the empirical estimate of HSIC (12) is then given by (Gretton et al, 2005)

$$D(\{(\mu_{\mathbb{P}_n}, \mu_{\mathbb{Q}_n})\}_{n=1}^N) := \frac{1}{N^2}\mathrm{tr}\mathbf{HKHL} = \frac{1}{N^2}\mathrm{tr}\bar{\mathbf{K}}\bar{\mathbf{L}}, \tag{13}$$

where $\mathbf{H} = \mathbf{I} - 1_N 1_N^\top / N$ is a centering matrix, and $\mathbf{K}, \mathbf{L} \in \mathbb{R}^{N \times N}$ are the kernel matrices for the kernel embeddings, i.e. $\mathbf{K}_{n,m} = K(\mu_{\mathbb{P}_n}, \mu_{\mathbb{P}_m})$ and $\mathbf{L}_{n,m} = L(\mu_{\mathbb{Q}_n}, \mu_{\mathbb{Q}_m})$, and $\bar{\mathbf{K}} = \mathbf{HKH}$ and $\bar{\mathbf{L}} = \mathbf{HLH}$ denote the centered versions of $\mathbf{K}$ and $\mathbf{L}$, respectively.

Then, we solve the task of unsupervised group matching based on the framework of (2) using the HSIC (13) for the dependency between distributions as follows,

$$\arg\max_{\pi \in \mathbf{\Pi}_N} \mathrm{tr}(\bar{\mathbf{K}}\pi^\top \bar{\mathbf{L}}\pi). \tag{14}$$

We refer to the proposed method as *group kernelized sorting*.

We find group matching by solving a convex relaxation of (14) as described in (Djuric et al, 2012). The convex version achieved better matching performance than the original kernelized sorting. Equation (14) is rewritten by the following equivalent problem,

$$\arg\min_{\pi \in \mathbf{\Pi}_N} \parallel \bar{\mathbf{K}}\pi^\top - (\bar{\mathbf{L}}\pi)^\top \parallel^2, \tag{15}$$

where $\parallel \cdot \parallel$ denotes the Frobenius norm. We relax the constraint that π is a permutation matrix, and then obtain the following convex problem,

$$\begin{aligned} \arg\min_{\pi^*} \quad & \parallel \bar{\mathbf{K}}\pi^{*\top} - (\bar{\mathbf{L}}\pi^*)^\top \parallel^2 \\ \text{subject to} \quad & \pi^*_{ij} \geq 0, \quad \pi^*\mathbf{1} = \mathbf{1}, \quad \pi^{*\top}\mathbf{1} = \mathbf{1}, \end{aligned} \tag{16}$$

where the permutation matrix binary constraint $\pi_{ij} \in \{0, 1\}$ is replaced by the interval constraint $\pi^*_{ij} \in [0, 1]$, and π^* is a doubly-stochastic matrix. This convex problem can be solved by a numerical optimization method, such as the trust-region-reflective algorithm (Coleman and Li, 1996), which is used for our experiments. The time complexity for each iteration is $O(N^2)$.

After doubly-stochastic matrix π^* is obtained by (16), we find hard assignments by solving the following linear assignment problem defined by π^*,

$$\arg\min_{\pi \in \mathbf{\Pi}_N} \sum_{i,j} \pi_{ij}\pi^*_{ij}, \tag{17}$$

Algorithm 1 Procedures of the proposed method.

Require: two grouped data sets $\{\mathbf{X}_n\}, \{\mathbf{Y}_n\}$, level-1 and level-2 kernels
Ensure: assignments π
1: calculate kernel values between groups $\mathbf{K}$, $\mathbf{L}$
2: centerize the kernel values $\bar{\mathbf{K}}$, $\bar{\mathbf{L}}$
3: obtain a doubly-stochastic matrix π^* by (16)
4: obtain hard assignments π by (17)

where the Hungarian algorithm (Kuhn, 1955) is used. Algorithm 1 shows the procedures of the proposed method. The time complexity for solving (17) by the Hungarian algorithm is cubic to the number of groups, $O(N^3)$.

The doubly-stochastic matrix π^* contains soft assignments. Therefore, we can rank matched groups with the ith group using probabilities $\pi^*_{i1}, \cdots, \pi^*_{iN}$.

3.4 Level-2 kernels

We show here examples of level-2 kernel K on the set of embedded distributions $\mathcal{F}_\mathcal{P}$, as well as their empirical estimates (Muandet et al, 2012). We will omit those for L on $\mathcal{F}_\mathcal{Q}$, as they are similar to the case of K. Recall that we represent each group $\boldsymbol{X}_n$ as an empirical embedding $\mu_{\mathbb{P}_n} = \frac{1}{I_n}\sum_{i=1}^{I_n} k(\cdot, x_{ni})$ using kernel k on the original space $\mathcal{X}$.

Linear kernel. Linear kernel K_{LIN} on $\mathcal{F}_\mathcal{P} \subset \mathcal{H}_k$ is defined as $K_{\mathrm{LIN}}(\mu_\mathbb{P}, \mu_{\mathbb{P}'}) := \langle \mu_\mathbb{P}, \mu_{\mathbb{P}'} \rangle_{\mathcal{H}_k}$, $\forall \mu_\mathbb{P}, \mu'_\mathbb{P} \in \mathcal{F}_\mathcal{P}$. Thus, the kernel value for group $\boldsymbol{X}_n$ and group $\boldsymbol{X}_m$ is given by

$$\begin{aligned} K_{\mathrm{LIN}}(\mu_{\mathbb{P}_n}, \mu_{\mathbb{P}_m}) &= \left\langle \frac{1}{I_n}\sum_{i=1}^{I_n} k(\cdot, x_{ni}), \frac{1}{I_m}\sum_{j=1}^{I_m} k(\cdot, x_{mj}) \right\rangle_{\mathcal{H}_k} \\ &= \frac{1}{I_n I_m}\sum_{i=1}^{I_n}\sum_{j=1}^{I_m} k(x_{ni}, x_{mj}). \end{aligned} \tag{18}$$

Polynomial kernels. Nonlinear kernels can also be defined on $\mathcal{F}_\mathcal{P}$. For example, a polynomial kernel on $\mathcal{F}_\mathcal{P}$ is given by

$$K_{\mathrm{POLY}}(\mu_{\mathbb{P}_n}, \mu_{\mathbb{P}_m}) = (\langle \mu_{\mathbb{P}_n}, \mu_{\mathbb{P}_m} \rangle_{\mathcal{H}_k} + c)^d, \tag{19}$$

where $d \in \mathbb{N}$ is the order of polynomial and $c > 0$ is a constant.

Gaussian kernels. We can define a Gaussian kernel K_γ on $\mathcal{F}_\mathcal{P}$ with parameter $\gamma > 0$ by

$$\begin{aligned} K_\gamma(\mu_{\mathbb{P}_n}, \mu_{\mathbb{P}_m}) &= \exp\left(-\frac{\gamma}{2}||\mu_{\mathbb{P}_n} - \mu_{\mathbb{P}_m}||^2_{\mathcal{H}_k}\right) \\ &= \exp\left(-\frac{\gamma}{2}(\langle \mu_{\mathbb{P}_n}, \mu_{\mathbb{P}_n} \rangle_{\mathcal{H}_k} - 2\langle \mu_{\mathbb{P}_n}, \mu_{\mathbb{P}_m} \rangle_{\mathcal{H}_k} + \langle \mu_{\mathbb{P}_m}, \mu_{\mathbb{P}_m} \rangle_{\mathcal{H}_k})\right). \end{aligned} \tag{20}$$

Note that we can calculate the inner-product $\langle\cdot\rangle_{\mathcal{H}_k}$ in the above nonlinear kernels as for linear kernel K_{LIN} (18).

The time complexity for calculating the kernel matrix $\mathbf{K}$ of N kernel embeddings is $O(N^2I^2)$, where I is the average number of objects in each group.

3.5 Discussion

We assume that a group is modeled as a distribution. When samples in a group $\boldsymbol{X}_n = \{x_{n1}, \cdots, x_{nI_n}\}$ are generated i.i.d. according to a distribution $\mathbb{P}_n^*$, it is valid to represent this group by the empirical distribution $\mathbb{P}_n$. In general, the samples are assumed to have a stationary distribution $\mathbb{P}_n^*$, and the empirical distribution $\mathbb{P}_n$ should be a consistent estimator of that distribution.

The performance of unsupervised group matching depends on kernels. When correspondence information between groups in different domains is available, we would select the best kernel by evaluating the matching performance. In the case that correspondence information is unavailable, it is impossible to select the best kernel in terms of the matching performance. However, some heuristics could be used. For example, the median trick can be used for Gaussian kernels, where the median of pairwise distances between the kernel embeddings is used for its width. The median trick has been widely used for kernel embeddings (Quadrianto et al, 2010; Gretton et al, 2012b; Song et al, 2012; Muandet and Schölkopf, 2013). Another heuristic is to use kernels that maximize RKHS distance (Sriperumbudur et al, 2009), which is equivalent to minimizing the error in classifying groups under linear loss.

4 Theoretical Analysis

4.1 Relation between HSIC on the embeddings and the original spaces

Here, we see that HSIC on the kernel embeddings defined in the previous section can be reduced to HSIC on the original spaces for special situations. Proposition 1 below, which is similar to the one obtained for the support measure machines (Muandet et al, 2012), shows that if the level-2 kernels are linear and the distributions of the groups have a specific form, HSIC of the embeddings can be written in terms of HSIC on the original spaces using specific kernels. The proof is given in the appendix.

Proposition 1 *Let k and ℓ be bounded kernels. Let $p(x|x')$ and $q(y|y')$ be conditional densities on $\mathcal{X}$ and $\mathcal{Y}$, respectively. Assume that we are given distributions $\{(\mathbb{P}_n, \mathbb{Q}_n)\}_{n=1}^N$ whose densities are given by $p(x|x_n)$ and $q(y|y_n)$, where $x_n \in \mathcal{X}$, $y_n \in \mathcal{Y}$ are points in the original spaces. Then HSIC (13) on embeddings $\{(\mu_{\mathbb{P}_n}, \mu_{\mathbb{Q}_n})\}_{n=1}^N$ using linear kernels for $\mathcal{F}_\mathcal{P}$ and $\mathcal{F}_\mathcal{Q}$ is equivalent to HSIC on $\{(x_n, y_n)\}_{n=1}^N$ using the kernels*

$$k_p(x, x') := \int\int k(\tilde{x}, \tilde{x}')dp(\tilde{x}|x)dp(\tilde{x}'|x'), \tag{21}$$

and

$$\ell_q(y, y') := \int\int \ell(\tilde{y}, \tilde{y}') dq(\tilde{y}|y) dq(\tilde{y}'|y') \tag{22}$$

for $\mathcal{X}$ and $\mathcal{Y}$, respectively.

Let us explain the proposition by some instantiations. For example, if each group is represented by one point in the original space, i.e. $p(\cdot|x_n) := \delta_{x_n}(\cdot)$ and $p(\cdot|y_n) := \delta_{y_n}(\cdot)$, then we have $k_p(x, x') = k(x, x')$ and $\ell_q(y, y') = \ell(y, y')$. Thus, HSIC on groups subsumes HSIC on the original spaces as a special case.

Consider another case where each distribution is described by a Gaussian distribution with mean x_n and fixed variance $\sigma_{\mathcal{X}}^2 > 0$, i.e. $p(x|x_n) = \mathcal{N}(x_n, \sigma_{\mathcal{X}}^2 \mathbf{I})$. Assume also that the kernel is Gaussian with bandwidth $\sigma > 0$, i.e. $k(x, x') = k_{\sigma^2}(x, x') := \exp(-\|x - x'\|^2/2\sigma^2)$. Then the convolution theorem of Gaussian distributions shows that $k_p(x, x') = (\frac{\sigma^2}{\sigma^2 + 2\sigma_{\mathcal{X}}^2})^{d/2} k_{\sigma^2 + 2\sigma_{\mathcal{X}}^2}(x, x')$. Namely, in this case HSIC on the kernel embeddings is reduced to HSIC on the original spaces using Gaussian kernels with larger bandwidth.

Note that our approach does not assume any model for the distributions, and thus can be applied to situations where such prior knowledge on the group structures is not available.

4.2 Characteristic property of the Gaussian kernel on the set of kernel embeddings

As stated in Section 3.2.2, HSIC in population (12) is equal to zero if and only $\mu_{\mathbb{P}}$ and $\mu_{\mathbb{Q}}$ are independent, when the joint kernel $K \otimes L$ is characteristic. It can be easily shown that the joint kernel is characteristic if both of the level-2 kernels K and L are characteristic. Here, we show that the Gaussian kernel K_γ (20) defined on $\mathcal{F}_{\mathcal{P}}$ (and thus the Gaussian kernel on $\mathcal{F}_{\mathcal{Q}}$) is characteristic. The proofs are given in the appendix.

We first show in Lemma 1 below that the set of all the kernel embeddings $\mathcal{F}_{\mathcal{P}}$ is compact if the original space $\mathcal{X}$ is compact.

Lemma 1 *Let $(\mathcal{X}, d_{\mathcal{X}})$ be a compact metric space and $\mathcal{P}$ be the set of all Borel probability measures on $\mathcal{X}$. Let k be a continuous kernel on $\mathcal{X}$, $\mathcal{H}_k$ be its RKHS, and*

$$\mathcal{F}_{\mathcal{P}} := \{f \in \mathcal{H}_k : f = \int k(\cdot, x) d\mathbb{P}(x), \exists \mathbb{P} \in \mathcal{P}\}. \tag{23}$$

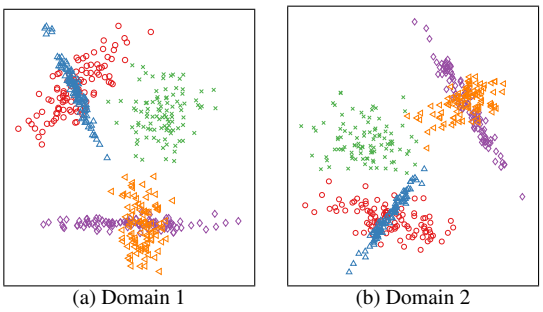
Let γ_k be a metric on $\mathcal{F}_{\mathcal{P}}$ defined by

$$\gamma_k(f, g) = \|f - g\|_{\mathcal{H}_k}, \ \forall f, g \in \mathcal{F}_{\mathcal{P}}. \tag{24}$$

Then $(\mathcal{F}_{\mathcal{P}}, \gamma_k)$ is a compact metric space.

It is known that a universal kernel[2] on a compact metric space is characteristic. By showing that K_γ is universal using the result of (Christmann and Steinwart, 2010), we have the following theorem.

[2] A kernel is called universal if its associated RKHS is dense in the space of bounded continuous functions (Steinwart, 2001).



(a) Domain 1 (b) Domain 2

Fig. 2 Example of two-domain synthetic data with five groups. Each point represents an object, and its color shows the assigned group.

Theorem 1 *Let $(\mathcal{X}, d_{\mathcal{X}})$ be a compact metric space. Then the Gaussian kernel K_γ (20) is characteristic to the set of all probability measures on $(\mathcal{F}_{\mathcal{P}}, \gamma_k)$.*

Let us intuitively explain the obtained result. Suppose that the joint distribution $\Pr(\mu_{\mathbb{P}}, \mu_{\mathbb{Q}})$ and the product of their marginals $\Pr(\mu_{\mathbb{P}})$ and $\Pr(\mu_{\mathbb{Q}})$ have the same "mean" and "covariance" over $\mathcal{F}_{\mathcal{P}} \times \mathcal{F}_{\mathcal{Q}}$, but differ in their higher order moments. If we use polynomial kernels of order 2 as K and L, then HSIC (12) gives the value zero. This is because in this case the kernel embeddings into $\mathcal{H}_K \otimes \mathcal{H}_L$ only distinguish the distributions on $\mathcal{F}_{\mathcal{P}} \times \mathcal{F}_{\mathcal{Q}}$ up to their covariance structures. As seen from this example, if the level-2 kernels are not characteristic, HSIC may not detect the dependency between $\mu_{\mathbb{P}}$ and $\mu_{\mathbb{Q}}$ even if there exists dependency in their higher order moments. On the other hand, we can measure any dependency between $\mu_{\mathbb{P}}$ and $\mu_{\mathbb{Q}}$ using HSIC if we use characteristic kernels for K and L.

Thus Theorem 1 ensures that if we use the Gaussian kernels for K and L, theoretically we can detect any dependency between $\mu_{\mathbb{P}}$ and $\mu_{\mathbb{Q}}$. Note, however, that we can also use kernels that are not characteristic in practice, as we use the linear kernel (18) in Section 5.1 and 5.2. In such cases, HSIC takes a large value if there exists strong linear dependency between $\mu_{\mathbb{P}}$ and $\mu_{\mathbb{Q}}$.

5 Experiments

5.1 Synthetic data

To demonstrate the performance of the proposed method, we used synthetic two-domain data. The data consist of five groups for each domain, and objects in a group are distributed according to a group-specific two-dimensional Gaussian distribution. An example of the synthetic data is shown in Figure 2. The five Gaussians in the second domain are the rotated versions of Gaussians in the first domain. Two pairs of Gaussians (red-blue and purple-orange) share their means but have different covariance matrices.

We compared the proposed method, group kernelized sorting (GKS), with two other kernelized sorting based methods: KS-mean and KS-object. With the KS-mean

Table 1 Average matching accuracies and their standard errors with different level-1 kernels using the synthetic data. Linear represents the linear kernel. Polynomial represents the polynomial kernel with degree two and constant one. Gaussian ($\gamma = 1$) represents the Gaussian kernel with width $\gamma = 1$. Gaussian (median) represents the Gaussian kernel with median trick.

Level-1 kernel	Linear	Polynomial	Gaussian ($\gamma = 1$)	Gaussian (median)
Accuracy	0.536 ± 0.029	0.524 ± 0.025	0.680 ± 0.037	$\mathbf{0.946} \pm 0.014$

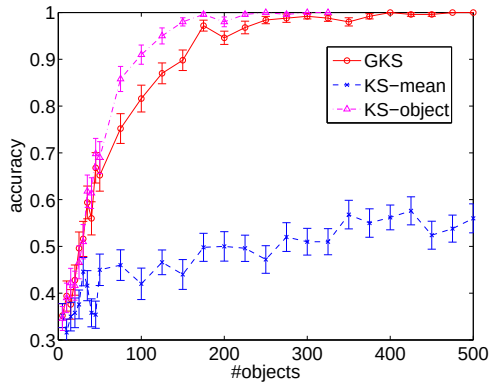
method, first the empirical mean for each group is calculated, and then convex kernelized sorting (Djuric et al, 2012) is performed using the empirical means as features for each group. With the KS-object method, objects are first matched based on kernelized sorting using the original features. Next, we calculate a doubly-stochastic matrix that represents correspondence probabilities between groups, which are estimated by the number of matched objects between groups. Then, groups are matched using the doubly-stochastic matrix by the Hungarian algorithm. For the proposed method, we used Gaussian kernels with median trick as level-1 kernels, and linear kernels as level-2 kernels.

Figure 3 shows the matching accuracy and computational time with different numbers of objects, which are averaged over 100 experiments. The accuracy with the KS-mean was low because it did not use covariance information for each group. The accuracy with the KS-object was slightly higher than that with the proposed method (GKS) with this simple example. However, the computational time of the KS-object was much higher than those of the proposed method and KS-mean. The computational time of the proposed method with 500 objects was 3.4 seconds. The time complexity of the proposed method and KS-mean is cubic to the number of groups, while that of the KS-object is cubic to the number of objects, which is costly since the number of groups is much smaller than the number of objects.

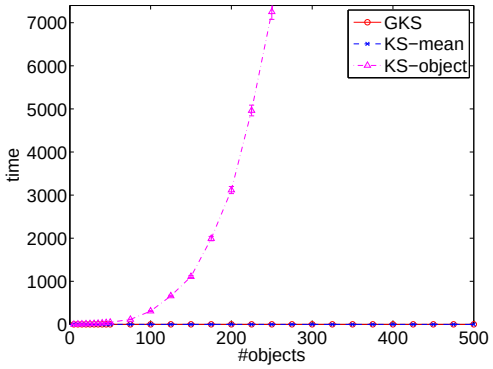
Table 1 shows the matching accuracy with different level-1 kernels using the synthetic data with 200 objects. For the level-2 kernel, we used the linear kernel. Since the linear and polynomial kernels are not characteristic, their accuracies were lower than the Gaussian kernel, which is characteristic. The performance of the Gaussian kernel with median trick was better than the Gaussian kernel with fixed width $\gamma = 1$. This result indicates that the kernel parameter setting is important, and the median trick works for the proposed method.

5.2 Splitted real data

We evaluated the proposed method by using four data sets with multiple class labels, which were obtained from the LIBSVM data sets (Chang and Lin, 2011). The statistics of the four data sets are summarized in Table 2. The data sets with two domains were generated by random splitting of the features as (Quadrianto et al, 2010; Djuric et al, 2012) did in their experiments. Because the two domains do not share their features, similarities between objects in different domains cannot be calculated. We assume that objects with the same class labels form a group. We used Gaussian level-1 kernel and linear level-2 kernel for the proposed method, and Gaussian kernel for the KS-mean method. The result is shown in Table 3, which are averaged over 100



(a) Matching accuracy



(b) Computational time in seconds

Fig. 3 Average matching accuracy and computational time with different numbers of objects using the synthetic data. The bars show their standard errors.

Table 2 Statistics of four splitted real data sets.

	Glass	Vowel	Satimage	Letter
#objects	214	528	4,435	15,000
#groups	6	11	6	26

experiments. The KS-object method was not applicable to Satimage and Letter data sets because they contain a large number of objects. The proposed method (GKS) achieved higher matching accuracy than the KS-mean and KS-object methods in all of the data sets. This result indicates that the proposed method successfully finds correspondence between groups by representing each group by its distribution using kernel embedding.

5.3 Cross-lingual topic matching

For an application of unsupervised group matching, we employed the proposed method for matching text document topic categories in different languages. We used 78 cat-

Table 3 Average matching accuracies and their standard errors with four splitted real data sets. Values in bold typeface are statistically better at the 5% level from those in normal typeface as indicated by a paired t-test.

Method	Glass	Vowel	Satimage	Letter
GKS	$\mathbf{0.442} \pm 0.030$	$\mathbf{0.249} \pm 0.017$	$\mathbf{0.993} \pm 0.005$	$\mathbf{0.058} \pm 0.005$
KS-mean	0.297 ± 0.024	0.167 ± 0.017	0.955 ± 0.015	$\mathbf{0.049} \pm 0.005$
KS-object	0.245 ± 0.020	0.168 ± 0.013	N/A	N/A
Random	0.167 ± 0.000	0.091 ± 0.000	0.167 ± 0.000	0.038 ± 0.000

Table 4 Matching accuracy with multilingual Wikipedia data. Values in bold typeface are better than the other.

	EN-DE	EN-FI	EN-FR	EN-IT	EN-JA	DE-FI	DE-FR	DE-IT
GKS	**0.603**	**0.321**	**0.564**	**0.782**	**0.410**	**0.385**	**0.538**	**0.615**
KS-mean	0.192	0.269	0.346	0.564	0.231	0.295	0.397	0.385
Vocabulary-size	0.051	0.051	0.051	0.064	0.051	0.090	0.051	0.051

	DE-JA	FI-FR	FI-IT	FI-JA	FR-IT	FR-JA	IT-JA	Average
GKS	**0.526**	**0.564**	**0.526**	**0.449**	0.577	**0.513**	**0.564**	**0.529**
KS-mean	0.167	0.449	0.308	0.064	**0.679**	0.487	0.256	0.339
Vocabulary-size	0.026	0.026	0.077	0.051	0.090	0.038	0.103	0.058

egories from Wikipedia in six languages: English (EN), German (DE), Finnish (FI), French (FR), Italian (IT) and Japanese (JA). Each category contains at least 100 documents, and a document can be associated with multiple categories. The number of documents is 7,447 for each language. For the feature vector of documents, we used tf-idf (term frequency - inverse document frequency). The feature vectors were then normalized to unit length in terms of ℓ_2 norm. For the level-1 kernel, we used a polynomial kernel with degree two and zero constant, which is a common kernel for the bag of words representation of text documents (Taira and Haruno, 1999). For the level-2 kernel, we used a Gaussian kernel with median trick. We compared with KS-mean and Vocabulary-size. The vocabulary-size method finds category matching based on the average vocabulary size per document, where it assumes that a group with large vocabulary in a language is likely to have large vocabulary in different languages. We did not compare with KS-object because it is inapplicable to data with a large number of objects as shown above.

The matching accuracy between pairs of six languages is shown in Table 4. The accuracy with random matching is 0.013. With 14 language pairs out of 15 pairs, the proposed method achieved the highest accuracy. In some language pairs, e.g. EN-DE and DE-JA, improvement of the proposed method from the KS-mean was significant. This implies that the empirical average of word counts is not enough for representing document categories, and it is important to model the distribution of documents in each category. Table 5 shows incorrectly matched categories estimated by the proposed method. Even though they were incorrect, the proposed method found correspondence between related categories across different languages, such as ‘Premier League player’ and ‘Football (soccer) strikers’, ‘English-language films’ and ‘American films’, and ‘American stage actors’ and ‘New York actors’ in matching between English and German.

Table 5 Examples of incorrectly matched categories in Wikipedia

English	German
Disambiguation	Greek mythology
Premier League players	Football (soccer) strikers
English-language films	American films
2006 FIFA World Cup players	Premier League players
American films	English-language films
American film directors	American novelists
American stage actors	New York actors
Greek mythology	20th century classical composers
2002 FIFA World Cup players	Football (soccer) midfielders
Emmy Award winners	Jewish actors
UEFA Euro 2004 players	UEFA Euro 2000 players
English	Finnish
American film actors	American television actors
Disambiguation	2000s automobiles
American television actors	California actors
Hollywood Walk of Fame	American film directors
Grammy Award winners	Disambiguation
English-language films	Rock and Roll Hall of Fame inductees
American films	20th century classical composers
American film directors	BAFTA winners (people)
La Liga footballers	Football (soccer) midfielders
American Jews	Hollywood Walk of Fame
Greek mythology	Medicinal plants
English	French
Disambiguation	2000s automobiles
American television actors	New York actors
Grammy Award winners	Disambiguation
Least Concern species	Grammy Award winners
English-language films	American films
American films	Companies listed on the New York Stock Exchange
Greek mythology	1990s music groups
BAFTA winners (people)	American voice actors
2002 FIFA World Cup players	Football (soccer) strikers
Emmy Award winners	BAFTA winners (people)
Jewish actors	American television actors
English	Italian
Disambiguation	Chemical elements
English-language films	American films
2006 FIFA World Cup players	Football (soccer) midfielders
American films	Greek mythology
American film directors	BAFTA winners (people)
Greek mythology	English-language films
BAFTA winners (people)	Emmy Award winners
Emmy Award winners	Jewish actors
Jewish actors	American film directors
2000s music groups	1990s music groups
1990s music groups	2000s music groups
English	Japanese
Disambiguation	Knights of the Golden Fleece
English-language films	American films
2006 FIFA World Cup players	Football (soccer) midfielders
Serie A players	Italian footballers
American films	American novelists
American film directors	English-language films
La Liga footballers	Football (soccer) strikers
American Jews	People from New York City
American stage actors	Jewish actors
Greek mythology	Chemical elements
National flags	Disambiguation

6 Conclusion

We proposed a method for unsupervised group matching, called group kernelized sorting, for finding correspondence between groups in different domains. With the proposed method, a group is represented by a distribution, and distributions are matched by maximizing the dependency. The dependency is measured by the Hilbert space information criterion, which can be calculated efficiently by mapping distributions into reproducing kernel Hilbert spaces. In experiments, we confirmed that the proposed method can perform much better than object matching based methods using synthetic and real data sets. The high performance of the proposed method indicates the validity of modeling groups as distributions.

Our approach can be further improved in a number of ways. We can extend the proposed method to semi-supervised setting, where a small number of correspondences across different domains are available. The correspondences can be either between groups or between objects. The proposed method can also be extended to data with different number of groups across different domains. This can be achieved by using the rectangular doubly-stochastic matrix in (16), and adding dummy groups with low connecting weights so as to discourage aligning with these dummy groups in (17) as proposed in (Jagarlamudi et al, 2010).

Since we assume that two domains have common groups, a method to distinguish whether the given data contain common groups or not would be beneficial. The doubly-stochastic matrix π^* obtained by (16) might be used for this. When there are common groups, the doubly-stochastic matrix are likely to be clear as with similar to a permutation matrix. On the other hand, when there is no common groups, the doubly-stochastic matrix would be unconcentrated, and the probability mass is scattered in a wide variety of possible matchings.

A Proofs

A.1 Proof of Proposition 1

Proof First note that k_p is well-defined since k is bounded. Then we have

$$\begin{aligned} K_{\mathrm{LIN}}(\mu_{\mathbb{P}_n}, \mu_{\mathbb{P}_m}) &:= \langle \mu_{\mathbb{P}_n}, \mu_{\mathbb{P}_m} \rangle_{\mathcal{H}_k} \\ &= \left\langle \int k(\cdot, x) dp(x|x_n), \int k(\cdot, \tilde{x}) dp(\tilde{x}|x_m) \right\rangle_{\mathcal{H}_k} \\ &= \int \int k(x, \tilde{x}) dp(x|x_n) dp(\tilde{x}|x_m) \\ &= k_p(x_n, x_m). \end{aligned} \tag{25}$$

This completes the proof for k_p. The proof for ℓ_q is obtained by a literal repetition of the arguments above.

A.2 Proof of Lemma 1

Proof It is known that the space of probabilities $\mathcal{P}$ with the weak topology is compact, provided that $(\mathcal{X}, d_{\mathcal{X}})$ is compact (Thm. 6.4., (Parthasarathy, 1967)). Let $(\mathbb{P}_n)_{n=1}^{\infty}$ be a sequence in $\mathcal{P}$ such that $\mathbb{P}_n$

converges weakly to $\mathbb{P}$. Then,

$$\gamma_k(\mu_{\mathbb{P}_n}, \mu_{\mathbb{P}})^2 = \int k(x, \tilde{x}) d\mathbb{P}_n(x) d\mathbb{P}_n(\tilde{x}) - 2 \int k(x, \tilde{x}) d\mathbb{P}_n(x) d\mathbb{P}(\tilde{x}) + \int k(x, \tilde{x}) d\mathbb{P}(x) d\mathbb{P}(\tilde{x})$$

converges to zero by the definition of weak convergence, since $k(x, \tilde{x})$ is a bounded continuous function. This implies that the mapping $P \mapsto \mu_{\mathbb{P}}$ is continuous, and thus its image $\mathcal{F}_{\mathcal{P}}$ is compact (Thm. 2.2.3., (Dudley, 2002)).

A.3 Proof of Theorem 1

Proof Thm. 2.2. of (Christmann and Steinwart, 2010) shows that K_γ is universal on $\mathcal{F}_{\mathcal{P}}$, since the identity map

$$\mathrm{id} : (\mathcal{F}_{\mathcal{P}}, \gamma_k) \to (\mathcal{H}_k, \|\cdot\|_{\mathcal{H}_k}), \tag{26}$$

is clearly continuous and injective. Therefore, K_γ is characteristic on $\mathcal{F}_{\mathcal{P}}$, since a universal kernel on a compact metric space is characteristic (Thm. 5, (Gretton et al, 2012a)).

References

Barnard K, Duygulu P, Forsyth D, De Freitas N, Blei DM, Jordan MI (2003) Matching words and pictures. Journal of Machine Learning Research 3:1107–1135

Chang C, Lin C (2011) LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2(3):27

Christmann A, Steinwart I (2010) Universal kernels on non-standard input spaces. In: Advances in Neural Information Processing Systems, pp 406–414

Coleman TF, Li Y (1996) An interior trust region approach for nonlinear minimization subject to bounds. SIAM Journal on Optimization 6(2):418–445

Djuric N, Grbovic M, Vucetic S (2012) Convex kernelized sorting. In: AAAI Conference on Artificial Intelligence

Doan A, Madhavan J, Domingos P, Halevy A (2004) Ontology matching: A machine learning approach. In: Handbook on Ontologies, pp 385–403

Dudley RM (2002) Real Analysis and Probability. Cambridge University Press

Fukumizu K, Bach FR, Jordan MI (2004) Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. Journal of Machine Learning Research 5:73–99

Fukumizu K, Gretton A, Sun X, Schölkopf B (2008) Kernel measures of conditional dependence. In: Advances in Neural Information Processing Systems, pp 489–496

Gretton A, Bousquet O, Smola A, Schölkopf B (2005) Measuring statistical dependence with Hilbert-Schmidt norms. In: Algorithmic Learning Theory, vol 3734, pp 63–77

Gretton A, Borgwardt K, Rasch M, Schölkopf B, Smola A (2012a) A kernel two-sample test. Journal of Machine Learning Research 13:723–773

Gretton A, Borgwardt KM, Rasch MJ, Schölkopf B, Smola A (2012b) A kernel two-sample test. Journal of Machine Learning Research 13(1):723–773

Haghighi A, Liang P, Berg-Kirkpatrick T, Klein D (2008) Learning bilingual lexicons from monolingual corpora. In: Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp 771–779

Iwata T, Hirao T, Ueda N (2013) Unsupervised cluster matching via probabilistic latent variable models. In: AAAI Conference on Artificial Intelligence, pp 445–451

Jagarlamudi J, Juarez S, Daumé III H (2010) Kernelized sorting for natural language processing. In: AAAI Conference on Artificial Intelligence, pp 1020–1025

Kamahara J, Asakawa T, Shimojo S, Miyahara H (2005) A community-based recommendation system to reveal unexpected interests. In: International Multimedia Modelling Conference, pp 433–438

Klami A (2012) Variational Bayesian matching. In: Asian Conference on Machine Learning, pp 205–220

Kuhn HW (1955) The hungarian method for the assignment problem. Naval Research Logistics Quarterly 2(1–2):83–97

Li B, Yang Q, Xue X (2009) Transfer learning for collaborative filtering via a rating-matrix generative model. In: International Conference on Machine Learning, pp 617–624

Muandet K, Schölkopf B (2013) One-class support measure machines for group anomaly detection. In: Conference on Uncertainty in Artificial Intelligence, pp 449–458

Muandet K, Fukumizu K, Dinuzzo F, Schölkopf B (2012) Learning from distributions via support measure machines. In: Advances in Neural Information Processing Systems, pp 10–18

Parthasarathy KR (1967) Probability Measures on Metric Spaces. Academic Press

Quadrianto N, Smola AJ, Song L, Tuytelaars T (2010) Kernelized sorting. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(10):1809–1821

Shvaiko P, Euzenat J (2013) Ontology matching: state of the art and future challenges. IEEE Transactions on Knowledge and Data Engineering 25(1):158–176

Smola A, Gretton A, Song L, Schölkopf B (2007) A Hilbert space embedding for distributions. In: Algorithmic Learning Theory, pp 13–31

Socher R, Fei-Fei L (2010) Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 966–973

Song L, Smola A, Gretton A, Bedo J, Borgwardt K (2012) Feature selection via dependence maximization. Journal of Machine Learning Research 13(1):1393–1434

Sriperumbudur BK, Fukumizu K, Gretton A, Lanckriet GR, Schölkopf B (2009) Kernel choice and classifiability for RKHS embeddings of probability distributions. In: Advances in Neural Information Processing Systems, pp 1750–1758

Sriperumbudur BK, Gretton A, Fukumizu K, Schölkopf B, Lanckriet GR (2010) Hilbert space embeddings and metrics on probability measures. Journal of Machine Learning Research 11:1517–1561

Steinwart I (2001) On the influence of the kernel on the consistency of support vector machines. Journal of Machine Learning Research 2:67–93

Taira H, Haruno M (1999) Feature selection in SVM text categorization. In: National Conference on Artificial Intelligence, pp 480–486

Terada A, Sese J (2012) Global alignment of protein-protein interaction networks for analyzing evolutionary changes of network frameworks. In: Proceedings of 4th International Conference on Bioinformatics and Computational Biology, pp 196–201

Tripathi A, Klami A, Virpioja S (2010) Bilingual sentence matching using kernel CCA. In: IEEE International Workshop on Machine Learning for Signal Processing, pp 130–135

Yamada M, Sugiyama M (2011) Cross-domain object matching with model selection. In: International Conference on Artificial Intelligence and Statistics, pp 807–815