Improving Predictors via Combination Across Diverse Task Categories

Kwang In Kim [1]

Abstract

Predictor combination is the problem of improving a task predictor using predictors of other tasks when the forms of individual predictors are unknown. Previous work approached this problem by nonparametrically assessing predictor relationships based on their joint evaluations on a shared sample. This limits their application to cases where all predictors are defined on the same task category, e.g. all predictors estimate attributes of shoes. We present a new predictor combination algorithm that overcomes this limitation. Our algorithm aligns the heterogeneous domains of different predictors in a shared latent space to facilitate comparisons of predictors independently of the domains on which they are originally defined. We facilitate this by a new data alignment scheme that matches data distributions across task categories. Based on visual attribute ranking experiments on datasets that span diverse task categories (e.g. shoes and animals), we demonstrate that our approach often significantly improves the performances of the initial predictors.

1. Introduction

Can a predictor of a shoe attribute be improved by leveraging knowledge gained from learning animal attributes?

Predictor combination (PC) aims to improve a *target predictor* of some task based on *reference predictors* of other tasks when the forms of individual predictors are not known. This scenario occurs when reference predictors are precompiled software libraries or Web services. Further, in PC the unknown reference forms are different from each other and from the target predictor. For example, a support vector machine (SVM) ranker might be determined as the best model for a target task, with Gaussian process (GP) regressors and deep neural network (DNN) classifiers presented as references.

[1]UNIST, Ulsan, Korea. Correspondence to: Kwang In Kim <kimki@unist.ac.kr>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

Predictor combination is related to multi-task learning (MTL) and transfer learning (TL). However, since no access to the forms of the references are provided, existing MTL and TL approaches, e.g. hard or soft sharing of predictor parameters (Vandenhende et al., 2020; Gao et al., 2019; Misra et al., 2016), enforcing parameter similarities (Agarwal et al., 2010; Wang et al., 2009; Argyriou et al., 2008), or embedding these parameters into a shared latent space (Sanh et al., 2019; Gong et al., 2012; Titsias & Lázaro-Gredilla, 2011; Luo et al., 2013; Lee et al., 2016; Zhang et al., 2019) are not directly applicable. Furthermore, it is not known in advance whether a given reference is *relevant* (i.e. useful in improving the target) and therefore naïvely exploiting all references could even degrade the performance similarly to the case of *negative transfer* commonly observed in MTL (Lee et al., 2016; Maninis et al., 2019; Nguyen et al., 2020).

(Kim et al., 2017a) approached this problem by nonparametrically assessing the task relevance: Evaluating all predictors on a sample dataset, the relevance of a reference is measured based on the similarity of its sample evaluation to the corresponding target evaluation. Under this setting, the target is improved by selectively enhancing these reference similarities. This instantiates a nonparametric extension of the classical MTL where the similarities of predictor parameters are enforced (Agarwal et al., 2010; Argyriou et al., 2008). (Kim et al., 2020) improved upon this first PC approach by adopting a Bayesian framework, casting the determination of relevant references into Bayesian *automatic relevance determination* (Rasmussen & Williams, 2006).

Existing PC approaches meet the challenging requirements of automatically identifying relevant references and improving the target without requiring known reference forms. They demonstrated that significant performance improvements can be achieved when the target and references predictors are defined on the same task category (Kim et al., 2017a; 2020): For example, when the target estimates how *formal* shoes are while the references are specialized for *sporty* and *pointy-at-the-front* attributes, all predictors belong to the same *Shoes* task category (see Sec. 3).

However, these approaches cannot be directly applied when the target and references lie in different categories, e.g. the target is constructed for shoe attributes while the presented references correspond to animal or human face attributes.

This limitation stems from the nonparametric nature of their relevance determination process: To calculate the similarities among the predictors, all predictors need to be *jointly* evaluated on a *single* sample set X, i.e. each entry in the evaluation $f|_X$ of the target should match the corresponding reference evaluations $\{g^r|_X\}_{r=1}^R$. When the predictors are defined across multiple task categories such as shoes and animals, such a single dataset might not exist.

(Kim & Chang, 2019) partially addressed this by establishing instance-level correspondences across datasets. Once constructed, such correspondences enable to jointly evaluate all predictors. However, this approach cannot be applied to heterogeneous task categories since it requires example *ground-truth correspondences*, but such ground-truths might not exist e.g. for shoes and animals.

In this paper, we present a new algorithm that combines predictors across task categories. Adopting ideas from domain adaptation studies, we align datasets of the target and a reference by mapping them to a shared latent space where the respective data-generating distributions match. This facilitates applying the references, originally tailored for their own data distributions, to the (aligned) target data.

As the target and reference data spaces and the corresponding probability distributions can differ significantly, naïvely matching these distributions in the latent space can generate maps (to the latent space) that fail to preserve the *structure* of the original data. Therefore, we regularize the construction of such maps by enforcing the preservation of local distance structures. The resulting algorithm aligns datasets in an *unsupervised manner* and therefore enables us to combine predictors across heterogeneous categories without having to require any example correspondences.

In visual attribute ranking experiments with datasets that represent diverse task categories (human faces, birds, animals, shoes, and outdoor scenes), we demonstrate that predictor combination across heterogeneous task categories can indeed significantly improve the performance of target predictors, providing a positive answer to the question posed at the beginning of this section.

Related work. Our approach was inspired by the success of MTL and TL approaches: While there are only limited existing studies on explicitly combining predictors across significantly different data categories (e.g. shoes and animals), in principle, most existing MTL approaches can be applied to such cases (Meyerson & Miikkulainen, 2021; 2019; Rebuffi et al., 2017; Ammar et al., 2015; Mahmud & Ray, 2008). Indeed, it is common to use deep neural networks (DNNs) pre-trained on ImageNet (Deng et al., 2009; Russakovsky et al., 2015) as initialization for task-specific refinement, e.g. (Xian et al., 2019). Multi-task learning has also been applied to train predictors of multiple data modalities (e.g. surface normal, segmentation, and salience maps (Kokkinos, 2017)). Our approach can be considered as an adaptation of these MTL approaches to PC problems where access to the internals of references is restricted.

The crux of our approach is to *match* the target and reference distributions such that the target data can be properly evaluated by the reference predictors. This problem has been previously studied in the context of TL. For example, (Long et al., 2015) updated the reference DNNs (called the *source* in TL problems) in the way that the resulting activations match those of the target. For matching the distributions, the maximum mean discrepancy (MMD) (Gretton et al., 2012) was used similarly to our approach (Sec. 2). Later, (Long et al., 2015) improved this via the *joint MMD* that helps model activations of multiple domain-specific DNN layers (Long et al., 2017). (Wei et al., 2018) applied a similar strategy to determine the optimal source combinations avoiding negative transfer.

Unlike traditional TL problems, in predictor combination, the references cannot be updated as their forms are unknown. Therefore, instead of adapting the reference predictors to the target problem, we transfer the target data into a latent space where the references can be directly applied. In this respect, our latent space mapping approach makes an instance of *domain adaptation* where the MMD has been extensively used in aligning different data distributions: (Pan et al., 2009)'s *transfer component analysis* finds a subspace of the original data spaces where the source and target distributions match. (Saito et al., 2018) applied a similar idea to classification problems by modeling the discrepancy between the class-conditional distributions of the source and target domains. An extensive empirical study of different algorithms can be found in (Csurka et al., 2017).

Matching the probability distributions of the source (reference) and target data, e.g. using MMD, has been especially successful when the corresponding data domains are *inherently related* as in applying an object recognition system trained on a specific environment and dataset to new environments and data representations but with known object types (Yan et al., 2017; Long et al., 2015). However, in our PC scenario, the target and reference data might not have such inherent relationships (e.g. animals vs. shoes) and in such cases, directly altering the target to match the reference distributions can destroy the structure present in the original data. Our approach addresses this by explicitly retaining the local structure of the original data distributions (Sec. 2.2.1).

Recent advances in image translation offer a new approach to domain adaptation: Instead of matching the probability distributions of features, one could directly translate raw images to the target domain, as demonstrated by successful transfer of e.g. photographs to paintings (Zhu et al., 2017) and line drawings to photographs (Kim et al., 2017b).

(Murez et al., 2018) further demonstrated that these techniques can be used in adapting domains for classification problems by incorporating classification loss into the cycle-consistent translation framework. (Tzeng et al., 2017) exercised a similar approach using traditional (non-cyclic) GAN-type discriminators. In the experiments, we demonstrate that while such techniques can actually be used in realizing PC algorithms, our approach outperforms these alternatives as they often struggle to faithfully translate images when the source and target domains differ significantly.

2. Combining predictors over task categories

Problem definition. Suppose that we are given a *predictor* function $f^I : \mathcal{X}^f \to \mathcal{Y}^f$ presented as an initial solution of a *target* task. The goal of predictor combination (PC) is to improve f^I based on the predictors $G = \{g^r : \mathcal{X}^r \to \mathcal{Y}^r\}_{r=1}^R$ of other tasks (referred to as *references*).

The internal structures of the target and reference predictors are unknown, and it is not known in advance, if there is any reference in G that is *relevant* (i.e. useful in improving f^I). Therefore, a PC algorithm has to 1) identify relevant references in G and 2) use such references to construct an improved version f^O of f^I without having to require access to the forms of the references.

Existing PC algorithms addressed this challenge by taking nonparametric approaches: All predictors are evaluated on a single dataset $X \subset \mathcal{X}^f$, and the corresponding outputs are used to assess the predictor relevance. This requires that all predictors are defined on the same input space, i.e. $\mathcal{X}^f = \mathcal{X}^r$ for all $1 \leq r \leq R$, and each instance in X should be *jointly* evaluated by all predictors. However, when the reference tasks are defined on different spaces or if their data distributions are significantly different from the target task, such a single dataset tying all tasks might not exist.

Algorithm overview. Our algorithm enables to combine predictors of diverse task categories by explicitly aligning the respective data domains and their distributions: We map the target and reference data into a shared latent space where the maximum mean discrepancy (MMD) (Gretton et al., 2012) between the mapped distributions is minimized. To ensure that the global structures of the original data are preserved under these maps, they are regularized by maximizing the Hilbert-Schmidt independence criterion (HSIC) of the original and the latent data distributions (Gretton et al., 2005). Once data are aligned, the subsequent combination process is carried out by employing (Kim et al., 2020)'s joint predictability enhancement framework.

In this section, we first present a model space of predictors that provides a unified view of previous PC approaches, and based on that, develop a new model space and the corresponding predictor combination algorithm.

2.1. Model space of predictors

Here, we denote a predictor (either a target f or a reference g) by h. For the i-th task, we assume that its input space $\mathcal{X}^i$ is equipped with a probability distribution $\mathbb{P}^i$. For simplicity of exposition, we will assume that the output space of each task is $\mathbb{R}$. However, extending our algorithm to multi-dimensional outputs is straightforward.

Baseline model space. Existing PC approaches assume that all predictors share the same input space: $(\mathcal{X}^i, \mathbb{P}^i) := (\mathcal{X}, \mathbb{P})$. Their model space is the Hilbert sphere $\mathcal{M}^1$, a submanifold of L^2 space provided with the inner product $\langle h^i, h^j \rangle_{\mathbb{P}} := \int h^i(\mathbf{x}) h^j(\mathbf{x}) \mathbb{P}(\mathbf{x})$: For $h \in \mathcal{M}^1$

$$\langle h, 1(\cdot) \rangle_{\mathbb{P}} = 0, \quad \langle h, h \rangle_{\mathbb{P}} = 1 \tag{1}$$

with $1(\cdot)$ being a constant function of ones. As the predictors in $\mathcal{M}^1$ are centered and normalized, their similarities can be measured independently of the scales via the inner product $\langle \cdot, \cdot \rangle_{\mathbb{P}}$. Under this setting, existing PC algorithms improve the initial predictor f^I by applying iterative averaging processes on $\mathcal{M}^1$: Given the predictor f^t at step t, the new solution f^{t+1} is obtained as the maximizer of

$$\mathcal{O}^1(f) = \langle f, f^t \rangle_{\mathbb{P}}^2 + \lambda \langle f, \mathcal{K}_G[f] \rangle_{\mathbb{P}}, \tag{2}$$

where $\lambda \geq 0$ is a hyperparameter and $\mathcal{K}_G[\cdot]$ is a linear positive definite operator on $\mathcal{M}^1$ responsible for capturing the reference relevance. Specific PC algorithms are instantiated based on how $\mathcal{K}_G[\cdot]$ is defined: For example, the first PC algorithms of (Kim et al., 2017a) and (Kim & Chang, 2019) are obtained by

$$\mathcal{K}_G[f] = \sum_{r=1}^{R} g^r w^r \langle f, g^r \rangle$$

with $w^r = \exp\left(-\frac{\|f^t - g^r\|^2}{\sigma_w^2}\right)$ for a parameter $\sigma_w^2 > 0$, rendering them into *diffusion processes* on $\mathcal{M}^1$ (Kim & Chang, 2019).[1] Here, the relevance (weight) w^r of g^r at time $t+1$ is determined based on its similarity to f^t. This helps ignore outlier references as they tend to get assigned smaller weights as diffusion progresses.

In the recent algorithm of (Kim et al., 2020), $\mathcal{K}_G[f]$ is constructed based on a Gaussian process (GP) estimator of the target f based on the references in G as inputs. Adopting the Bayesian framework, this method casts the identification of relevant references into mathematically rigorous, automatic

[1] The original algorithm of (Kim et al., 2017a) was designed for Bayesian predictors: $\{f, g^i\}$ are *predictive distributions* and their inner product in $\mathcal{O}^1$ is derived based on the Kullback Leibler-divergence. Our re-interpretation is obtained by taking only the predictive means. This ignores the potentially useful predictive uncertainties but offers a wider range of PC applications.

relevance determination problem (Rasmussen & Williams, 2006). Our algorithm builds upon this as detailed in Sec. 2.2.

Model space of predictors of heterogeneous task categories. Our model space $\mathcal{M}$ consists of predictors each associated with its own input space: $h^i : \mathcal{X}^i \to \mathbb{R}$ is individually normalized by the corresponding distribution $\mathbb{P}^i$:

$$\langle h^i(\mathbf{x}), 1(\cdot)\rangle_{\mathbb{P}^i} = 0, \quad \langle h^i, h^i\rangle_{\mathbb{P}^i} = 1 \tag{3}$$

As there is no natural inner product operation defined on $\mathcal{M}$, existing PC algorithms (Eq. 2) cannot be directly applied. Our next step is to add a structure that joins these *bundles* and enable comparisons of different predictors.

First, we connect each pair of input data spaces $\{(\mathcal{X}^i, \mathbb{P}^i), (\mathcal{X}^j, \mathbb{P}^j)\}$ based on a latent space $\mathcal{Z}^{ij}$ and the corresponding smooth *chart* maps $\{q^i : \mathcal{Z}^{ij} \to \mathcal{X}^i, q^j : \mathcal{Z}^{ij} \to \mathcal{X}^j\}$ each having a smooth inverse. Section 2.2 will detail the construction of such chart maps. Using these charts, the probability distributions $\mathbb{P}^i_{\mathcal{Z}}$ and $\mathbb{P}^j_{\mathcal{Z}}$ corresponding to $\mathbb{P}^i$ and $\mathbb{P}^j$, respectively are induced on $\mathcal{Z}^{ij}$: $\mathbb{P}^i_{\mathcal{Z}}(\mathbf{z}) = [\mathbb{P}^i \circ q^i](\mathbf{z}) := \mathbb{P}^i(q^i(\mathbf{z}))$.

This enables to measure the distance between $\mathbb{P}^i$ and $\mathbb{P}^j$, originally defined on different input spaces, indirectly using the distance of $\mathbb{P}^i_{\mathcal{Z}}$ and $\mathbb{P}^j_{\mathcal{Z}}$. Specifically, we use the *maximum mean discrepancy* (MMD), defined based on the *kernel mean embeddings* of distributions on $\mathcal{Z}^{ij}$, as such a distance measure (Gretton et al., 2012): The kernel mean embedding $\mu^i_{\mathcal{Z}}$ of $\mathbb{P}^i_{\mathcal{Z}}$ into the reproducing kernel Hilbert space (RKHS) $\mathcal{H}_{\mathcal{Z}}$, corresponding to a kernel $k_{\mathcal{Z}}(\cdot,\cdot) : \mathcal{Z}^{ij} \times \mathcal{Z}^{ij} \to \mathbb{R}$ is defined as

$$\mu^i_{\mathcal{Z}} := [\mathbb{E}^i_{\mathcal{Z}}]_{\mathbf{z}}[k_{\mathcal{Z}}(\mathbf{z}, \cdot)] = \int k_{\mathcal{Z}}(\mathbf{z}, \cdot) d\mathbb{P}^i_{\mathcal{Z}}(\mathbf{z}). \tag{4}$$

When $k_{\mathcal{Z}}(\cdot,\cdot)$ is *characteristic* such as Gaussian kernels

$$k_{\mathcal{Z}}(\mathbf{z}, \mathbf{z}') = \exp\left(-\frac{\|\mathbf{z} - \mathbf{z}'\|^2}{\sigma^2_{\mathcal{Z}}}\right)$$

with $\sigma^2_{\mathcal{Z}} > 0$, $\mu^i_{\mathcal{Z}}$ uniquely determines $\mathbb{P}^i_{\mathcal{Z}}$ (Jitkrittum et al., 2017). This makes the MMD, the squared distance in $\mathcal{H}_{\mathcal{Z}}$, a proper distance measure of distributions: $\mathrm{MMD}(\mathbb{P}^i_{\mathcal{Z}}, \mathbb{P}^j_{\mathcal{Z}}) = \|\mu^i_{\mathcal{Z}} - \mu^j_{\mathcal{Z}}\|^2_{\mathcal{H}_{\mathcal{Z}}}$. With this structure, we can embed any pair of predictors h^i and h^j (defined on $\mathcal{X}^i$ and $\mathcal{X}^j$, respectively) to a space $\mathcal{M}^{ij}$ of predictors on $\mathcal{Z}^{ij}$ where direct comparisons can be conducted: Our PC algorithm will construct chart maps $\{q^{fr}, q^r\}$ that minimize $d^{\mathrm{MMD}}(\mathbb{P}^{fr}_{\mathcal{Z}}, \mathbb{P}^r_{\mathcal{Z}})$ such that the dataset X^f originally sampled from $(\mathcal{X}^f, \mathbb{P}^f)$ can be evaluated by the embedded version of g^r (Sec. 2.2).[2]

Next, we add a measure of *similarity* between the original target distribution $\mathbb{P}^f$ and the corresponding induced distribution $\mathbb{P}^{fr}_{\mathcal{Z}}$ to ensure that $(q^{fr})^{-1}$ preserves the *structure* of the target space $(\mathcal{X}^f, \mathbb{P}^f)$. Specifically, we measure the similarity between two random variables $\mathbf{x} \in \mathcal{X}^f$ and $\mathbf{z} = (q^{fr})^{-1}(\mathbf{x}) \in \mathcal{Z}^{fr}$ using the Hilbert-Schmidt independence criterion (HSIC), defined as the MMD between the joint distribution of $\mathbf{x}$ and $\mathbf{z}$ and the corresponding product of marginals (Gretton et al., 2005). To facilitate this, we define an RKHS on $\mathcal{X}^f$ corresponding to a kernel $k^f : \mathcal{X}^f \times \mathcal{X}^f \to \mathbb{R}$:

$$\begin{aligned}\mathrm{HSIC}(\mathbf{x}, \mathbf{z}) = {} & \mathbb{E}_{\mathbf{x}\mathbf{x}'\mathbf{z}\mathbf{z}'}[k^f(\mathbf{x}, \mathbf{x}')k_{\mathcal{Z}}(\mathbf{z}, \mathbf{z}')] \\ & + \mathbb{E}_{\mathbf{x}\mathbf{x}'}[k^f(\mathbf{x}, \mathbf{x}')]\mathbb{E}_{\mathbf{z}\mathbf{z}'}[k_{\mathcal{Z}}(\mathbf{z}, \mathbf{z}')] \\ & - 2\mathbb{E}_{\mathbf{x}\mathbf{z}}\left[\mathbb{E}_{\mathbf{x}'}[k^f(\mathbf{x}, \mathbf{x}')]\mathbb{E}_{\mathbf{z}'}[k_{\mathcal{Z}}(\mathbf{z}, \mathbf{z}')]\right].\end{aligned} \tag{5}$$

Similarly to $k_{\mathcal{Z}}$, we use a Gaussian kernel for k^f:

$$k^f(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{(\sigma^f)^2}\right)$$

with $(\sigma^f)^2 > 0$. Finally, our similarity s^{HSIC} is given by scale-normalizing HSIC into $[0, 1]$:

$$s^{\mathrm{HSIC}}(\mathbb{P}^f, \mathbb{P}^{fr}_{\mathcal{Z}}) = \frac{\mathrm{HSIC}(\mathbf{x}, \mathbf{z})}{\sqrt{\mathrm{HSIC}(\mathbf{x}, \mathbf{x})}\sqrt{\mathrm{HSIC}(\mathbf{z}, \mathbf{z})}}. \tag{6}$$

Discussion. We constructed a separate latent space for each predictor pair. An alternative is to construct a single latent space shared by all predictors. In this setting, the chart maps can be jointly constructed, potentially benefiting from capturing the *interdependence of all data spaces*. However, this will lead to significantly higher computational overhead and further, it will make adding new references challenging (as all chart maps need to be re-calculated).

HSIC was originally conceived as a test of statistical dependence between random variables $\mathbf{x}$ and $\mathbf{z}$ (Gretton et al., 2005). However, in our case, q^{fr} is a function, and as such $\mathbf{x} = q^{fr}(\mathbf{z})$ and $\mathbf{z}$ have a deterministic dependence, making interpretation of HSIC as a statistical dependence test difficult. Instead, it can be interpreted as a similarity measure for the structures of the respective domains: s^{HSIC} has the maximum value of 1 when the pairwise similarities captured by $k^f(\mathbf{x}, \mathbf{x}')$ are exactly preserved by $(q^{fr})^{-1}$: $k^f(\mathbf{x}, \mathbf{x}') = k_{\mathcal{Z}}(q^{fr}(\mathbf{x}), q^{fr}(\mathbf{x}'))$ for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}^f$. As Gaussian kernels are localized (i.e. $k^f(\mathbf{x}, \mathbf{x}') \sim 0$ when $\|\mathbf{x} - \mathbf{x}'\|$ is large), our similarity measure s^{HSIC} evaluates the global similarity via combining *local structures*.

2.2. Predictor combination algorithm

With the latent spaces $\{\mathcal{Z}^{fr}\}$ and chart maps $\{q^{fr}, q^r\}_{r=1}^R$ aligning the probability distributions $\{\mathbb{P}^{fr}_{\mathcal{Z}}, \mathbb{P}^r_{\mathcal{Z}}\}_{r=1}^R$, we can evaluate the similarity of f and g^r based on their embeddings in $\{\mathcal{M}^{fr}\}$:

$$\begin{aligned}\langle f_{\mathcal{Z}}, g^r_{\mathcal{Z}}\rangle_{\mathbb{P}^{fr}_{\mathcal{Z}}} &= \langle f, (q^f)^{-1} \circ g^r_{\mathcal{Z}}\rangle_{\mathbb{P}^f} \\ &= \langle (q^r)^{-1} \circ f_{\mathcal{Z}}, g^r\rangle_{\mathbb{P}^r},\end{aligned} \tag{7}$$

where $f_{\mathcal{Z}} := f \circ q^{fr}$ and $g^r_{\mathcal{Z}} := g^r \circ q^r$. Now replacing $\langle\cdot,\cdot\rangle_{\mathbb{P}}$ in Eq. 2 with $\langle\cdot,\cdot\rangle_{\mathbb{P}^f}$, we obtain a new algorithm that

[2] We use the symbol q^{fr} for denoting a chart map from $\mathcal{Z}^{fr}$ to $\mathcal{X}^f$ to signify that it depends on both $(\mathcal{X}^r, \mathbb{P}^r)$ and $(\mathcal{X}^f, \mathbb{P}^f)$.

combines predictors across heterogeneous domains:

$$\mathcal{O}(f) = \langle f, f^t \rangle^2_{\mathbb{P}^f} + \lambda \langle f, \mathcal{K}_{(q^f)^{-1}[G_\mathcal{Z}]}[f] \rangle_{\mathbb{P}^f}, \quad (8)$$

where $(q^f)^{-1}[G_\mathcal{Z}] = \{(q^f)^{-1} \circ g^r_\mathcal{Z}\}_{r=1}^R$. Note that the first equality in Eq. 7 enables us to transfer the reference evaluations (each lying in the respective spaces) to the target space $\mathcal{X}^f$ to facilitate the evaluation of the second term in $\mathcal{O}$.

2.2.1. MMD-HSIC ALIGNMENT OF CHART MAPS

Crucial to the success of our PC approach is the construction of (inverse) chart maps $Q = \{(q^{fr})^{-1}, (q^r)^{-1}\}_{r=1}^R$: Our goal is to evaluate the target data X^f originally presented in $(\mathcal{X}^f, \mathbb{P}^f)$, based on the references $\{g^r\}$. As g^r is defined on its own domain $\mathcal{X}^r$ and is tailored for the corresponding distribution $\mathbb{P}^r$, the chart $(q^{fr})^{-1}$ needs to map X^f (and equivalently, its distribution $\mathbb{P}^f$) to match $\mathbb{P}^r_\mathcal{Z}$ such that it can be faithfully evaluated by $g^r_\mathcal{Z} = g^r \circ q^r$.

A simple approach to build such maps is to minimize the MMD between $\mathbb{P}^{fr}_\mathcal{Z}$ and $\mathbb{P}^r_\mathcal{Z}$:

$$\mathcal{E}^{\text{MMD}}((q^{fr})^{-1}, (q^r)^{-1}) = \text{MMD}(\mathbb{P}^{fr}_\mathcal{Z}, \mathbb{P}^r_\mathcal{Z}). \quad (9)$$

However, directly minimizing $\mathcal{E}^{\text{MMD}}$ without imposing any constraints can lead to arbitrary complex chart maps $(q^{fr})^{-1}$ and $(q^i)^{-1}$ that align the output distributions well in $\mathcal{Z}^{fr}$ but fail to transfer the *structure* of the respective input domains. In this case, the evaluation of the reference $g^r_\mathcal{Z}$ on the mapped input $(q^{fr})^{-1}(X^f)$ is unlikely to provide useful information for improving the target predictor.

We address this by first explicitly representing $(q^{fr})^{-1}$ as a kernel combination parameterized by $A^{fr} \in \mathbb{R}^{\dim(\mathcal{Z}^{fr}) \times B}$:

$$[(q^{fr})^{-1}(\cdot)]_n = \sum_{j=1}^B [A^{fr}]_{n,j} k^f(\mathbf{b}^f_j, \cdot), \quad (10)$$

where k^f is the Gaussian kernel used in defining the HSIC (Eq. 5): Using Gaussian kernels regularizes chart learning as in this case, q^f is inherently *smooth* (Schölkopf & Smola, 2002). The selection of the *basis set* $\{\mathbf{b}^f_j\}_{j=1}^B \subset \mathcal{X}^f$ will be discussed shortly. Further, we match the local structure of $(\mathcal{X}^f, \mathbb{P}^f)$ and $(\mathcal{Z}^{fr}, \mathbb{P}^{fr}_\mathcal{Z})$ by enhancing their HSIC

$$\mathcal{O}^{\text{HSIC}}((q^{fr})^{-1}) = s^{\text{HSIC}}(\mathbb{P}^f, \mathbb{P}^{fr}_\mathcal{Z}). \quad (11)$$

Figure 1 illustrates the effectiveness of regularizing the chart map learning process using the HSIC.

While the chart map $(q^i)^{-1}$ for the reference g^i can be similarly constructed, we set it as an identity map making $\mathcal{Z}^{fr}$ the same as $\mathcal{X}^r$. This reduces the overall computational complexity of the chart learning process. Our PC framework (Eq. 8) still applies thanks to the second equality of Eq. 7.

Finally, fixing q^r as identity, our chart $(q^{fr})^{-1}$ (equivalently, A^{fr}) is constructed as the minimizer of the energy

$$\mathcal{E}^{\text{chart}}(A^{fr}) = \mathcal{E}^{\text{MMD}}(q^{fr}, q^r) - \lambda^{\text{HSIC}} \mathcal{O}^{\text{HSIC}}(q^{fr}) \quad (12)$$

with a regularization parameter λ^{HSIC}.

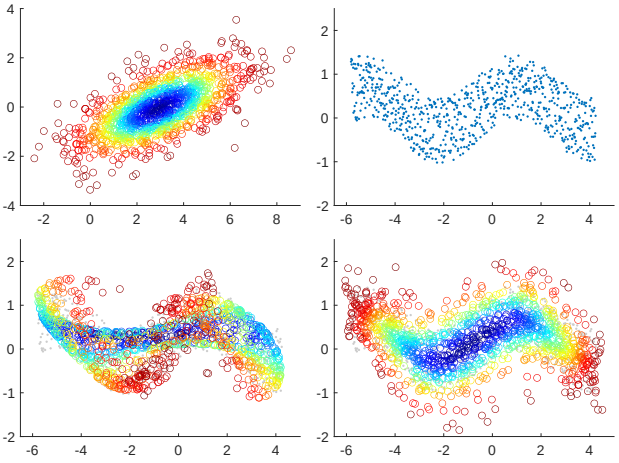

Figure 1. Chart construction examples: (top left) Target data X^f sampled from a Gaussian distribution $\mathbb{P}^f$. Data points are color-coded based on their Mahalanobis distances to the mean of $\mathbb{P}^f$. (top right) Reference data $X^1 \subset \mathcal{X}^1 = \mathcal{Z}^{f1}$ sampled from $\mathbb{P}^1$. (bottom) X^f transferred to $\mathcal{X}^1$ (overlaid on X^1; gray dots) by minimizing only the MMD (left; Eq. 9) and minimizing a combination of MMD and HSIC (right; Eq. 12): Using the MMD led to a slightly better alignment with $\mathbb{P}^1$. However, it failed to preserve the local distance structure of X^f as indicated by the mixed colors of the transferred data. Combining the MMD and HSIC provided a good trade-off between the reference alignment and the preservation of the original data structure.

2.2.2. SAMPLE-BASED APPROXIMATIONS

In practice, MMD and HSIC in Eq. 12 cannot be directly evaluated as they require integration with respect to unknown probability distributions $\{\mathbb{P}^f, \mathbb{P}^r\}$. Also, it is infeasible to directly optimize the target function f which is an infinite-dimensional object. As such, we take sample based approximations (Gretton et al., 2005; Jitkrittum et al., 2017).

For given target and reference datasets $X^f = \{\mathbf{x}^f_1, \dots, \mathbf{x}^f_N\}$ and $X^r = \{\mathbf{x}^r_1, \dots, \mathbf{x}^r_M\}$ sampled from $\mathbb{P}^f$ and $\mathbb{P}^r$, respectively, the sample X^f-based HSIC estimate is

$$\begin{aligned} \widehat{\mathcal{O}}^{\text{HSIC}} &= \text{trace}[\mathbf{K}^f \mathbf{C} \mathbf{K}_\mathcal{Z} \mathbf{C}], \quad (13) \\ [\mathbf{K}^f]_{m,n} &= k^f(\mathbf{x}^f_m, \mathbf{x}^f_n), \\ [\mathbf{K}_\mathcal{Z}]_{n,m} &= k_\mathcal{Z}((q^{fr})^{-1}(\mathbf{x}^f_n), (q^{fr})^{-1}(\mathbf{x}^f_m)), \end{aligned}$$

where $\mathbf{C} = I - \frac{1}{N}\mathbf{1}\mathbf{1}^\top$ and $\mathbf{1} = [1, \dots, 1]^\top$. As q^r is identity, $k_\mathcal{Z}$ is defined on $\mathcal{X}^r \times \mathcal{X}^r$. The corresponding approximate MMD is given as

$$\begin{aligned} \widehat{\mathcal{E}}^{\text{MMD}}(q^{fr}, q^i) = &-\frac{2}{NM} \sum_{n=1}^N \sum_{m=1}^M [\mathbf{K}''_\mathcal{Z}]_{n,m} \quad (14) \\ &+ \frac{1}{M^2} \sum_{m,n=1}^M [\mathbf{K}_\mathcal{Z}]_{n,m} + \frac{1}{N^2} \sum_{m,n=1}^N [\mathbf{K}'_\mathcal{Z}]_{n,m}, \end{aligned}$$

where $[\mathbf{K}''_\mathcal{Z}]_{n,m} = k_\mathcal{Z}((q^{fr})^{-1}(\mathbf{x}^f_n), \mathbf{x}^r_m)$ and $[\mathbf{K}'_\mathcal{Z}]_{n,m} = k_\mathcal{Z}(\mathbf{x}^r_n, \mathbf{x}^r_m)$.

Efficient approximations. The computational complexities of $\widehat{\mathcal{E}}^{\text{MMD}}$ and $\widehat{\mathcal{O}}^{\text{HSIC}}$ evaluations are $O(MN+M^2+N^2)$ and $O(N^2)$, respectively which are prohibitive for large-scale problems. We obtain computationally affordable approximations by adopting finite-rank approximations of the kernel functions $k_{\mathcal{Z}}$ and k^f: For given basis sets $\{\mathbf{b}_j^f\}_{j=1}^B \subset \mathcal{X}^f$ and $\{\mathbf{b}_j^r\}_{j=1}^B \subset \mathcal{X}^r$, the approximate kernels $\hat{k}^f$ and $\hat{k}_{\mathcal{Z}}$ are defined as

$$\begin{aligned}\hat{k}^f(\mathbf{x},\mathbf{x}') &= (\mathbf{k}_{\mathbf{x}}^f)^\top (\mathbf{K}_{BB}^f)^{-1}\mathbf{k}_{\mathbf{x}'}^f, \\ \hat{k}_{\mathcal{Z}}(\mathbf{z},\mathbf{z}') &= [\mathbf{k}_{\mathcal{Z}}]_{\mathbf{z}}^\top ([\mathbf{K}_{\mathcal{Z}}]_{BB})^{-1}[\mathbf{k}_{\mathcal{Z}}]_{\mathbf{z}'}, \\ \mathbf{k}_{\mathbf{x}}^f &= [k^f(\mathbf{x},\mathbf{b}_1^f),\dots,k^f(\mathbf{x},\mathbf{b}_B^f)]^\top, \\ [\mathbf{k}_{\mathcal{Z}}]_{\mathbf{z}} &= [k_{\mathcal{Z}}(\mathbf{z},\mathbf{b}_1^r),\dots,k_{\mathcal{Z}}(\mathbf{z},\mathbf{b}_B^r)]^\top\end{aligned} \tag{15}$$

with $[\mathbf{K}_{BB}^f]_{ij} = k^f(\mathbf{b}_i^f,\mathbf{b}_j^f)$ and $[[\mathbf{K}_{\mathcal{Z}}]_{BB}]_{ij} = k_{\mathcal{Z}}(\mathbf{b}_i^r,\mathbf{b}_j^r)$. The basis sets $\{\mathbf{b}_j^f\}$ and $\{\mathbf{b}_j^r\}$ are obtained as the cluster centers of X^f and X^r, respectively using k-means clustering with $k = B$. We fix the kernel rank B at 500. Now, substituting Eq. 15 into Eqs. 13 and 14, and subsequently to Eq. 12, we obtain the final energy for chart map learning:

$$\widehat{\mathcal{E}}^{\text{chart}}(A^{fr}) = \widehat{\mathcal{E}}^{\text{MMD}}(q^{fr},q^r) - \lambda^{\text{HSIC}}\widehat{\mathcal{O}}^{\text{HSIC}}(q^{fr}), \tag{16}$$

which can be minimized using the standard conjugate gradient method. A single evaluation of the gradient $\nabla_{A^{fr}}[\widehat{\mathcal{E}}^{\text{chart}}]$ takes linear time with respect to the numbers of data points N and M: $O(N \times M \times \dim(\mathcal{X}^r))$.

f-approximation. Similarly to previous PC approaches (Kim & Chang, 2019; Kim et al., 2020), we approximate predictors $\{f^I, g^r\}_{r=1}^R$ based on their evaluations on the test set X^f and improve the sample evaluation of f^I: $\mathbf{f}^I = [f^I(\mathbf{x}_1^f),\dots,f^I(\mathbf{x}_N^f)]^\top$. The sample reference $\mathbf{g}^r$ is defined based on the chart map $(q^{fr})^{-1}$: $\mathbf{g}^r = [g^r((q^{fr})^{-1}(\mathbf{x}_1^f)),\dots,g^r((q^{fr})^{-1}(\mathbf{x}_N^f))]^\top$. The resulting discretization of Eq. 8 is presented as

$$\widehat{\mathcal{O}}(\mathbf{f}) = \mathbf{f}^\top\mathbf{f}^t + \lambda\mathbf{f}^\top\mathbf{Q}_G\mathbf{f} \tag{17}$$

with a positive definite matrix $\mathbf{Q}_G$: Adopting (Kim et al., 2020)'s framework, we define it such that $\mathbf{f}^\top\mathbf{Q}_G\mathbf{f}$ becomes the accuracy of predicting $\mathbf{f}$ using $\mathbf{G} = [\mathbf{g}^1,\dots,\mathbf{g}^R]$ as input features based on Gaussian process (GP) regression.

In our model space, all predictors are centered and normalized (Eq. 3). As this is not the case in practice, we explicitly normalize them. Incorporating this into $\widehat{\mathcal{O}}$ in Eq. 17 we obtain a Rayleigh quotient for a hyperparameter $\lambda^{\text{GP}} > 0$:

$$\begin{aligned}\widehat{\mathcal{O}}(\mathbf{f}) =& \frac{\mathbf{f}^\top\mathbf{C}(\mathbf{f}^t(\mathbf{f}^t)^\top + \lambda\mathbf{S})\mathbf{C}\mathbf{f}}{\mathbf{f}^\top\mathbf{C}\mathbf{f}}, \\ \mathbf{S} =& 2\mathbf{H}(\mathbf{H}+\lambda^{\text{GP}}\mathbf{I})^{-1} \\ & - (\mathbf{H}+\lambda^{\text{GP}}\mathbf{I})^{-1}\mathbf{H}\mathbf{H}(\mathbf{H}+\lambda^{\text{GP}}\mathbf{I})^{-1}.\end{aligned} \tag{18}$$

The matrix $\mathbf{H}$ consists of anisotropic Gaussian kernel evaluations of the references:

$$[\mathbf{H}]_{i,j} = \exp\left(-(\mathbf{G}_{[i,:]} - \mathbf{G}_{[j,:]})\mathbf{\Sigma}(\mathbf{G}_{[i,:]} - \mathbf{G}_{[j,:]})^\top\right), \tag{19}$$

where $\mathbf{G}_{[i,:]}$ is the i-th row of $\mathbf{G}$ and $\mathbf{\Sigma}$ is a diagonal matrix of non-negative entries. The maximum of $\widehat{\mathcal{O}}$ is attained at the eigenvector corresponding to the largest eigenvalue of $\mathbf{f}^t(\mathbf{f}^t)^\top + \lambda\mathbf{S}$. Similarly to the case of kernel approximations in the HSIC and MMD evaluations (Eq. 15), $\mathbf{H}$ is approximated based on a low-rank factorization, leading to a computationally efficient algorithm to find the desired eigenvector: Readers are referred to the accompanying supplemental material for details of the $\mathbf{H}$ approximation and the derivation of $\widehat{\mathcal{O}}$ (Eq. 18) from $\mathcal{O}$ (Eq. 8). We fix the rank of this approximation at 300 following (Kim et al., 2020).

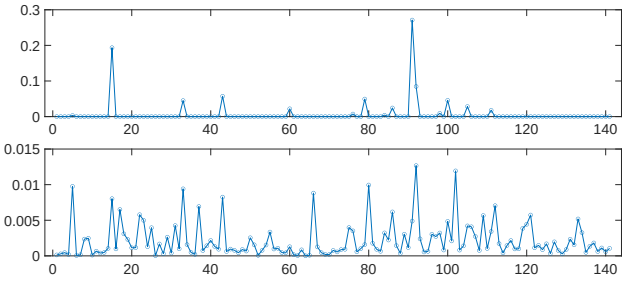

Figure 2. Diagonal values of $\mathbf{\Sigma}$ estimated for *OSR* dataset: Maximizing the marginal likelihood can yield highly skewed distribution of $\mathbf{\Sigma}$ values focusing only on few references (top). Adding the entropy regularizer (Eq. 20) leads to a more balanced selection of references (bottom). x-axis represents the reference index.

Identification of relevant tasks. The magnitudes of the entries in $\mathbf{\Sigma}$ (Eq. 19) represent the relevance of references: For large $[\mathbf{\Sigma}]_{r,r}$, the corresponding reference $\mathbf{g}^r$ makes significant contributions to kernel evaluations $\mathbf{H}$ and thereby has a large impact on predicting $\mathbf{f}$.

(Kim et al., 2020) determined $\mathbf{\Sigma}$ by maximizing the *marginal likelihood* $p(\mathbf{f}|\mathbf{G},\mathbf{\Sigma})$ under the GP prior. This enabled to automatically identify relevant references in (Kim et al., 2020). However, unlike their application scenario, our algorithm uses a much larger number of references and further, our references are designed for tasks that differ significantly from the target. In this case, directly maximizing the marginal likelihood tends to select only few spurious references ignoring less significantly related (to the target), but still potentially useful references (Fig. 2). Therefore, we regularize the estimation of $\mathbf{\Sigma}$ using a balancing term: Our algorithm maximizes

$$\mathcal{O}^{\text{ML}} = p(\mathbf{f}|\mathbf{G},\mathbf{\Sigma}) + \lambda^{\text{Ent}}\sum_{r=1}^R [\widetilde{\mathbf{\Sigma}}]_{r,r}\log([\widetilde{\mathbf{\Sigma}}]_{r,r}) \tag{20}$$

with $[\widetilde{\mathbf{\Sigma}}]_{r,r} = [\mathbf{\Sigma}]_{r,r}/\sum_j[\mathbf{\Sigma}]_{j,j}$. The second term of $\mathcal{O}^{\text{ML}}$ is the negative entropy when the diagonal terms of $\widetilde{\mathbf{\Sigma}}$ are inter-

preted as a probability distribution, and it encourages even reference contributions. Figure 2 demonstrates the effectiveness of this regularizer. In our preliminary experiments on *OSR* dataset (see Sec. 3), adding this regularizer improved the average accuracy of the final predictors by around 5.4% with negligible additional computational overhead.

Hyperparameters. The hyperparameters of our algorithm include the kernel parameters $(\sigma^f)^2$ and $\sigma^2_{\mathcal{Z}}$ for k^f (Eq. 5) and $k_{\mathcal{Z}}$ (Eq. 4), respectively, regularization parameter λ^{HSIC} for chart map estimation (Eq. 12), entropy regularizer λ^{Ent} (Eq. 20), GP parameter λ^{GP} (Eq. 18), combination parameter λ (Eq. 17), and the number of iterations T of the averaging process (Eq. 17): $(\sigma^f)^2$ is decided as the square of the average distances of data points in X^f to its mean. $\sigma^2_{\mathcal{Z}}$ is determined similarly based on X^r. λ^{HSIC} is decided at a small value of 10^{-8} while λ^{Ent} is fixed at 100 which guarantees that no value of $\mathbf{\Sigma}$ is 50 times larger than the mean values of the diagonal entries of $\mathbf{\Sigma}$ on *PubFig* dataset (See Sec. 3). The remaining parameters λ^{GP}, λ, and T are shared by (Kim et al., 2020)'s algorithm and they are tuned based on validation sets following their experimental protocol.

3. Experiments

To assess the effectiveness of our approach, we performed experiments on five datasets in visual attributes ranking: For a target visual attribute, our goal is to learn a ranking function f such that $f(\mathbf{x}_i) > f(\mathbf{x}_j)$ if image $\mathbf{x}_i$ has a stronger attribute presence than $\mathbf{x}_j$. The initial predictor f^I was constructed based on training rank pairs $S = \{[\mathbf{x}_{i(p)}, \mathbf{x}_{j(p)}]\}_{p=1}^{P}$ where $[\mathbf{x}_i, \mathbf{x}_j] \in S$ indicates that the attribute is stronger in $\mathbf{x}_i$ than $\mathbf{x}_j$, and the rank loss (Chapelle & Keerthi, 2010):

$$l([\mathbf{x}_i, \mathbf{x}_j]) = \max\left(0, 1 - (f(\mathbf{x}_i) - f(\mathbf{x}_j))\right)^2.$$

We initially constructed DNN rankers, and linear and nonlinear rank SVMs (Chapelle & Keerthi, 2010; Parikh & Grauman, 2011), and selected SVMs which consistently achieved the highest accuracy.

Datasets and settings. The Animals with Attributes 2 dataset (*AWA2*) consists of 37,322 images of 85 attributes and 50 classes (Xian et al., 2019). The goal is to estimate rankings on each of the target attributes where the labels are provided as class-wise comparisons, i.e. all images in a class has a stronger or weaker presence of certain attributes than other classes. The Caltech-UCSD Birds dataset (*CUB*) provides 312 attributes labeled based on 200 bird classes (Wah et al., 2011). Each image in these datasets is represented by features extracted using ResNet101 pre-trained on ImageNet. The Public Figure Faces (*PubFig*) and *Shoes* datasets respectively contain 772 images of eight people (classes) with 11 attributes, and 14,658 images of 10 shoe attributes and 10 categories. For these datasets, images are represented as GIST features and color histograms provided by (Kovashka et al., 2012). The Outdoor Scene Recognition (*OSR*) dataset provides six attributes from eight scene categories. We use GIST features from (Parikh & Grauman, 2011).

Table 1. Results of statistical significance tests of our method compared to the baseline initial targets ($\mathbf{f}^I$), and the PC algorithms that use the identity chart maps (*Id*), image translation (*IT*), and (only) *MMD*, based on a t–test with $\alpha = 0.95$. For each method, we show the numbers of target attributes where our algorithm is statistically significantly better (first column) and worse (second column).

Dataset	$\mathbf{f}^I$		*Id*		*IT*		*MMD*		# targets
AWA2	55	0	41	0	53	0	48	0	80
CUB	23	0	24	0	22	0	20	0	40
PubFig (ResNet)	9	0	8	0	7	0	5	0	11
PubFig	9	0	N/A		N/A		8	0	11
Shoes (ResNet)	9	0	10	0	5	0	4	0	10
Shoes	8	0	N/A		N/A		7	0	10
OSR (ResNet)	6	0	6	0	5	0	3	0	6
OSR	6	0	N/A		N/A		2	0	6
Total (%)	71.84	0	60.54	0	62.59	0	55.75	0	174

For each dataset, the initial predictor $\mathbf{f}^0$ of each target attribute was constructed based on 300 labels and it was improved by combining it with the references predictors constructed for *the other* datasets. All datasets are different in their image categories: *PubFig*, *OSR*, *Shoes*, *AWA2*, and *CUB*, respectively contain images of human faces, outdoor scenes, shoes, animals, and birds, and we are not aware of any instance-level connection among these datasets. To ensure that sufficient numbers of training and testing labels are presented for the experiments, we selected 80 and 40 attributes from *AWA2* and *CUB*, respectively.

The accuracies of the initial predictors on *PubFig*, *Shoes*, and *OSR* can be significantly improved when advanced ResNet features are used instead of the classical GIST and color histogram features. We also performed experiments with these features to evaluate the performance of our PC algorithm when the initial predictors have higher accuracy.

We evaluated the accuracy of ranking results using 100 times Kendall's τ coefficients measuring the percentage difference of correctly and incorrectly estimated pairs. For each dataset, experiments were repeated 10 times with different training and test set combinations and the results were averaged.

Baselines. We are not aware of any existing algorithm that can be applied to the scenario of combining predictors across heterogeneous task categories. Therefore, our experiments focused on assessing alternative algorithm design possibilities using different chart map learning strategies (Sec. 2.2.1): We compare with 1) baseline initial predictors

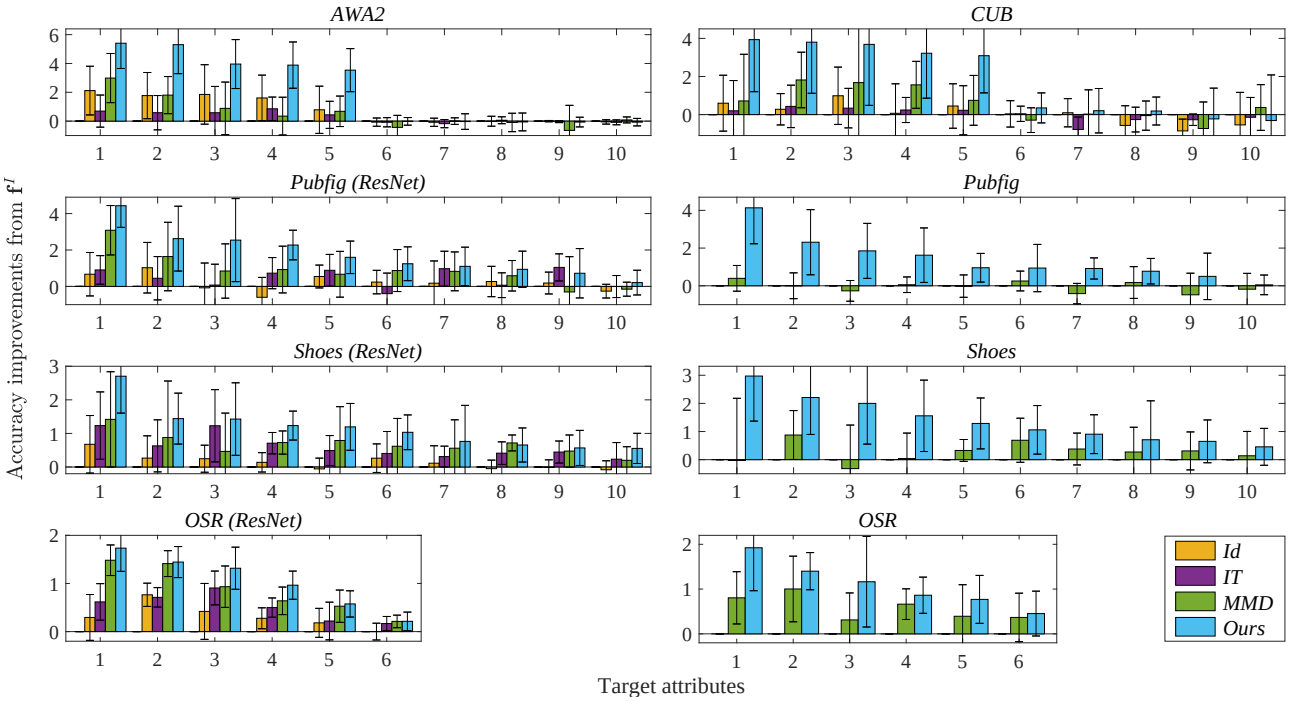

Figure 3. Average accuracy improvements (from the initial predictors $\mathbf{f}^I$) of different PC algorithms that use 1) the identity chart maps (*Id*); 2) image translation networks (*IT*); 3) *MMD*; 4) a combination of MMD and HSIC (*Ours*). The error bars represent twice the standard deviations. The absolute accuracy values of all predictors are provided in the supplemental document.

$\mathbf{f}^I$; and PC algorithms that use 2) identity chart maps, i.e. evaluating the references directly on the target data (*Id*); 3) image translation (*IT*) via cycle-consistent translation networks (Zhu et al., 2017) as image-level domain adaptation; 4) only *MMD* as commonly used in existing domain adaptation work (Pan et al., 2009; Saito et al., 2018; Long et al., 2015; 2017) (obtained by removing the HSIC energy in our chart map learning scheme; Eq. 16). It should be note that *Id* is applicable only when the target and reference domains coincide (but with possibly different probability distributions), hence it cannot be applied to improving predictors of *PubFig*, *OSR*, and *Shoes* as they are constructed on classical features whereas the references use advanced ResNet features. While *IT* can be applied independently of the feature representations, our experiments with *IT* also focused on combining ResNet feature-based predictors as we do not have the exact same code that were used in extracting features for *PubFig*, *OSR*, and *Shoes* (Parikh & Grauman, 2011; Kovashka et al., 2012).

In the supplemental, we also compare with 1) a parametric adaptation of (Kim et al., 2020)'s PC algorithm and 2) and (Mejjati et al., 2018)'s nonparametric MTL algorithm: All PC algorithms considered here outperform these adaptations, demonstrating that existing PC and MTL algorithms cannot be straightforwardly extended to combining heterogeneous task predictors.

Results. Figure 3 and Table 1 summarize the results. In Fig. 3, we show the results of up to 10 different target attributes per dataset corresponding to the best (five) and worst (five) improvements from the baseline initial predictors $\mathbf{f}^I$ achieved by our algorithm (*Ours*). The complete results are provided in the accompanying supplemental which show a similar tendency.

While not all attributes showed marked improvements, all four PC algorithms (*Id*, *IT*, *MMD*, *Ours*) frequently achieved significant performance gain over $\mathbf{f}^I$'s, confirming the utility and possibility of predictor combination across heterogeneous task categories. Apart from one dataset (*AWA2*) where *Id* ranked second best (after *Ours*), *IT* and *MMD* outperformed *Id* demonstrating that when the target and reference domains are significantly different, aligning the target data with the distributions of references can help extract *useful* information from the references.

As demonstrated in Fig. 1, using MMD only in chart map learning can fail to preserve the local structure of the original target data. By explicitly addressing this with the additional HSIC regularizer, our final algorithm (*Ours*) showed further significant improvements. While in principle, image translation has the capability of matching images across different domains (e.g. matching aerial photographs to maps (Zhu et al., 2017)), we observed that training such translation networks is challenging when the two domains differ sub-

stantially as in our PC scenario (see supplemental document for example translation results). As a result, *IT* recorded only a comparable level of performance to *MMD*. Further, our preliminary experiments suggested that the *IT* training process is prone to mode collapse, making the transfer networks generate identical images, requiring human intervention, e.g. to restart with a new initialization or stop before mode collapse starts.

Overall, *Ours* statistically significantly improved $\mathbf{f}^I$, *Id*, *IT*, and *MMD* for 71.84%, 60.54%, 62.59%, and 55.75% of the total target attributes, respectively as shown in Table 1. Importantly, *Ours* did not significantly degraded performance in any of the total 174 target attributes.

4. Conclusions

Existing predictor combination (PC) approaches require all predictors to be jointly evaluated on a shared dataset. This limits their application domain to combining predictors in a single task category. We presented a new algorithm that overcomes this limitation and enables PC across heterogeneous task categories. Our algorithm maps the input data of the target task into a latent space where the reference predictors can be directly evaluated. To facilitate this process, we proposed a new data alignment scheme that combines the maximum mean discrepancy and Hilbert-Schmidt independence criterion. With experiments on five datasets that represent diverse task categories, we demonstrated that predictor combination across heterogeneous task categories (e.g. to combine predictors of shoe and animal attributes) can indeed significantly improve the initial predictors.

Limitations and future work. Our final algorithm as well as its variations *IT* and *MMD* requires knowledge of reference domains in the form of example inputs. While this should not be a severe limitation for most visual understanding problems since often, obtaining images (but not the corresponding labels) of specific task categories is not difficult. However, when the reference predictors are trained on proprietary or confidential data, our approach might not be directly applicable. Also, as our algorithm builds upon (Kim et al., 2020)'s predictor combination framework, it inherits some of its limitations: Our algorithm assesses the relevance of a reference based on how effective it is in predicting the target: We use Bayesian relevance determination with an additional entropy-based regularizer for this task. However, it is possible that our method is misled by spurious correlations: A simple failure case is when a copy of the initial target $\mathbf{f}^I$ is included in the reference set. In this case, our algorithm will identify this feature as the most relevant reference, but it might not help improve the target.

Future work should explore 1) the possibility of combining image translation with our MMD and HSIC-based chart map learning approach (as they are complementary) and 2) the application of our approach to combine predictors across different data modalities, e.g. images, sound, and text data.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant (No. 2021R1A2C2012195) and Institute of Information & Communications Technology Planning & Evaluation (IITP) grant (No. 2020–0–01336, Artificial Intelligence Graduate School Program, UNIST) both funded by the Korea government (MSIT). KIK thanks James Tompkin, Insoo Kim, and Suyeong Park for insightful discussions.

References

Agarwal, A., Gerber, S., and Daume, H. Learning multiple tasks using manifold regularization. In *NIPS*, pp. 46–54, 2010.

Ammar, H. B., Eaton, E., Ruvolo, P., and Taylor, M. E. Unsupervised cross-domain transfer in policy gradient reinforcement learning via manifold alignment. In *AAAI*, pp. 2504–2510, 2015.

Argyriou, A., Evgeniou, T., and Pontil, M. Convex multi-task feature learning. *Machine Learning*, 73(3), 2008.

Chapelle, O. and Keerthi, S. S. Efficient algorithms for ranking with SVMs. *Information Retrieval*, 13(3):201–215, 2010.

Csurka, G., Baradel, F., Chidlovskii, B., and Clinchant, S. Discrepancy-based networks for unsupervised domain adaptation: a comparative study. In *ICCV Workshops*, 2017.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In *CVPR*, pp. 248–255, 2009.

Gao, Y., Ma, J., Zhao, M., Liu, W., and Yuille, A. L. NDDR-CNN: Layerwise feature fusing in multi-task CNNs by neural discriminative dimensionality reduction. In *CVPR*, pp. 3205–3214, 2019.

Gong, P., Ye, J., and Zhang, C. Robust multi-task feature learning. In *KDD*, pp. 895–903, 2012.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. Measuring statistical dependence with Hilbert-Schmidt norms. In *ALT*, pp. 63–77, 2005.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *JMLR*, 13: 723–773, 2012.

Jitkrittum, W., Szabó, Z., and Gretton, A. An adaptive test of independence with analytic kernel embeddings. In *ICML*, pp. 1742–1751, 2017.

Kim, K. I. and Chang, H. J. Joint manifold diffusion for combining predictions on decoupled observations. In *CVPR*, pp. 7549–7557, 2019.

Kim, K. I., Tompkin, J., and Richardt, C. Predictor combination at test time. In *ICCV*, pp. 3553–3561, 2017a.

Kim, K. I., Richardt, C., and Chang, H. J. Combining task predictors via enhancing joint predictability. In *ECCV*, 2020.

Kim, T., Cha, M., Kim, H., Lee, J. K., and Kim, J. Learning to discover cross-domain relations with generative adversarial networks. In *ICML*, pp. 1857–1865, 2017b.

Kokkinos, I. UberNet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In *CVPR*, pp. 6129–6138, 2017.

Kovashka, A., Parikh, D., and Grauman, K. Whittlesearch: Image search with relative attribute feedback. In *CVPR*, pp. 2973–2980, 2012.

Lee, G., Yang, E., and Hwang, S. J. Asymmetric multi-task learning based on task relatedness and loss. In *ICML*, pp. 230–238, 2016.

Long, M., Cao, Y., Wang, J., and Jordan, M. I. Learning transferable features with deep adaptation networks. In *ICML*, pp. 97–105, 2015.

Long, M., Zhu, H., Wang, J.-M., and Jordan, M. I. Deep transfer learning with joint adaptation networks. In *ICML*, pp. 2208–2217, 2017.

Luo, Y., Tao, D., Geng, B., Xu, C., and Maybank, S. J. Manifold regularized multitask learning for semi-supervised multilabel image classification. *IEEE TIP*, 22(2):523–536, 2013.

Mahmud, M. M. H. and Ray, S. R. Transfer learning using Kolmogorov complexity: basic theory and empirical evaluations. In *NIPS*, 2008.

Maninis, K.-K., Radosavovic, I., and Kokkinos, I. Attentive single-tasking of multiple tasks. In *CVPR*, pp. 1851–1860, 2019.

Mejjati, Y. A., Cosker, D., and Kim, K. I. Multi-task learning by maximizing statistical dependence. In *CVPR*, pp. 3465–3473, 2018.

Meyerson, E. and Miikkulainen, R. Modular universal reparameterization: deep multi-task learning across diverse domains. In *NeurIPS*, 2019.

Meyerson, E. and Miikkulainen, R. The traveling observer model: multi-task learning through spatial variable embeddings. In *ICLR*, 2021.

Misra, I., Shrivastava, A., Gupta, A., and Hebert, M. Cross-stitch networks for multi-task learning. In *CVPR*, pp. 3994–4003, 2016.

Murez, Z., Kolouri, S., Kriegman, D., Ramamoorthi, R., and Kim, K. Image to image translation for domain adaptation. In *CVPR*, 2018.

Nguyen, T. A., Jeong, H., Yang, E., and Hwang, S. J. Clinical risk prediction with temporal probabilistic asymmetric multi-task learning. In *arXiv:2006.12777*, 2020.

Pan, S. J., Tsang, I., Kwok, J., and Yang, Q. Domain adaptation via transfer component analysis. In *IJCAI*, pp. 1187–1192, 2009.

Parikh, D. and Grauman, K. Relative attributes. In *ICCV*, pp. 503–510, 2011.

Rasmussen, C. E. and Williams, C. K. I. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006.

Rebuffi, S.-A., Bilen, H., and Vedaldi, A. Learning multiple visual domains with residual adapters. In *NIPS*, 2017.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015.

Saito, K., Watanabe, K., Ushikuand, Y., and Harada, T. Maximum classifier discrepancy for unsupervised domain adaptation. In *CVPR*, pp. 3723–3732, 2018.

Sanh, V., Wolf, T., and Ruder, S. A hierarchical multi-task approach for learning embeddings from semantic tasks. In *AAAI*, pp. 6949–6956, 2019.

Schölkopf, B. and Smola, A. J. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

Titsias, M. K. and Lázaro-Gredilla, M. Spike and slab variational inference for multi-task and multiple kernel learning. In *NIPS*, pp. 2339–2347, 2011.

Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. Adversarial discriminative domain adaptation. In *CVPR*, pp. 7167–7176, 2017.

Vandenhende, S., Georgoulis, S., Gansbeke, W. V., Proesmans, M., Dai, D., and Gool, L. V. Multi-task learning for dense prediction tasks: a survey. In *arXiv:2004.13379*, 2020.

Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Wang, F., Wang, X., and Li, T. Semi-supervised multi-task learning with task regularizations. In *ICDM*, pp. 562–568, 2009.

Wei, Y., Zhang, Y., Huang, J., and Yang, Q. Transfer learning via learning to transfer. In *ICML*, pp. 5085–5094, 2018.

Xian, Y., Lampert, C. H., Schiele, B., and Akata, Z. Zero-shot learning – A comprehensive evaluation of the good, the bad and the ugly. *IEEE TPAMI*, 41(9):2251–2265, 2019.

Yan, H., Ding, Y., Li, P., Wang, Q., Xu, Y., and Zuo, W. Mind the class weight bias: weighted maximum mean discrepancy for unsupervised domain adaptation. In *CVPR*, pp. 2272–2281, 2017.

Zhang, Z., Cui, Z., Xu, C., Yan, Y., Sebe, N., and Yang, J. Pattern-affinitive propagation across depth, surface normal and semantic segmentation. In *CVPR*, pp. 4106–4115, 2019.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017.