Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

From Shared Subspaces to Shared Landmarks: A Robust Multi-Source Classification Approach

Sarah M. Erfani[†*], Mahsa Baktashmotlagh[‡*], Masud Moshtaghi[†], Vinh Nguyen[†], Christopher Leckie[†], James Bailey[†], Kotagiri Ramamohanarao[†]

[†]Department of Computing and Information Systems, The University of Melbourne, Australia
[‡]Department of Science and Engineering, Queensland University of Technology, Australia
sarah.erfani@unimelb.edu.au

Abstract

Training machine leaning algorithms on augmented data from different related sources is a challenging task. This problem arises in several applications, such as the Internet of Things (IoT), where data may be collected from devices with different settings. The learned model on such datasets can generalize poorly due to distribution bias. In this paper we consider the problem of classifying unseen datasets, given several labeled training samples drawn from similar distributions. We exploit the intrinsic structure of samples in a latent subspace and identify *landmarks*, a subset of training instances from different sources that should be similar. Incorporating subspace learning and landmark selection enhances generalization by alleviating the impact of noise and outliers, as well as improving efficiency by reducing the size of the data. However, since addressing the two issues simultaneously results in an intractable problem, we relax the objective function by leveraging the theory of nonlinear projection and solve a tractable convex optimisation. Through comprehensive analysis, we show that our proposed approach outperforms state-of-the-art results on several benchmark datasets, while keeping the computational complexity low.

Introduction

The primary objective of supervised learning algorithms is to learn a function f from a training set (X, Y) that can generalize well on unseen test data X'. Traditional classifiers generalize well if X and X' are well behaved and follow the same (or very similar) distribution. However, this common fact may not hold in many applications, especially in the case of data collected from heterogenous sources. For example, in sensor monitoring networks, data may be collected from devices (i.e., domains) with different types, placements, orientations, and sampling frequencies (Stisen et al. 2015). In such applications, the classification model f may fail to generalize well due to distribution bias (or shift) in the collected samples. Therefore, developing a classification algorithm that generalizes well on acquired knowledge from various related sources and can be applied to unseen sources is an important and compelling problem.

With the advent of the IoT and the proliferation of smart devices there is a need for efficient algorithms that accommodate cross-platform data analysis. A common approach to improve the generalization ability of machine learning algorithms is to provide more training examples. However, when the examples are augmented data from multiple sources, inserting more data may only consume memory, rather than yielding better performance (Torralba and Efros 2011). The main reason behind the ineffectiveness of this common approach to modeling generalization is that the input space of the training set dramatically deviates from the test set, i.e., the datasets are biased. Consequently, the challenge is to build a system that is robust to the underlying distribution bias and performs well on unseen datasets.

Domain adaptation and domain generalization address the above problem by finding a shared subspace for related sources. The aim of domain adaptation is to produce robust models on a specific target (test) source, by leveraging supplementary information during training from this source, as well as taking labeled samples from multiple training sources. Domain adaptation produces target-specific models, indicating that the training process should be repeated for each target. Moreover, the target samples may not always be available. Domain generalization, in contrast, generates a model independent of targets. It only assumes that samples from multiple sources can be accessed, and makes no further assumption regarding the target. More specifically, domain generalization aims to cope with the deviations in the marginal distribution (X) and conditional distribution $(Y|X)$ among different sources. Blanchard et al. (2011) first introduced the notion of domain generalization. Muandet et al. (2013) developed a source invariant feature representation incorporating the distributional variance across sources to reduce the dissimilarity. However, more recently it has been shown that in many real world applications the shift between different source distributions may not be corrected by projecting the data to a latent space (Aljundi et al. 2015; Gong, Grauman, and Sha 2013), and the impact of noise and outliers may still persist.

The goal of our work is to improve classification accuracy across different sources. We introduce an efficient method that enhances classification by combining subspace projection and landmark selection. Our algorithm learns a shift-invariant latent space that minimizes the difference be-

[*]Both authors contributed equally to this work.

tween the marginal distributions (X) of the sources, while maintaining their functional relationship $(Y|X)$. To determine the similarity of two distributions, we make use of the Maximum Mean Discrepancy (MMD) (Gretton et al. 2012), which compares the means of two empirical distributions in a Reproducing Kernel Hilbert Space (RKHS). Based on the MMD formulation, we derive a new objective function and incorporate sample (or landmark) section. Landmarks are defined as a subset of a common space where the sources are closer to each other. However, since this new optimisation problem is intractable, we leverage the theory of nonlinear random projections and arrive at a relaxed objective function. The algorithm learns a lower rank representation of data, and exploits label information from the training sources to extract the landmarks that reduce the discrepancy between their distributions. The intuition behind landmark selection is that not all samples are equally amenable to generalization. More specifically, only certain samples, owing to their statistical properties, provide a bridge among the sources.

Through a comprehensive analysis on several image and sensor benchmark datasets, we demonstrate that our algorithm outperforms state-of-the-art results, while being computationally efficient. Unlike many existing approaches that are built based on nonlinear kernels (Blanchard, Lee, and Scott 2011; Muandet, Balduzzi, and Schölkopf 2013; Khosla et al. 2012), the proposed algorithm exploits random features in an invariant subspace to reveal nonlinear patterns in the data. It enables large-scale data processing of computationally expensive machine learning algorithms by significantly reducing the size of the data. Moreover, to the best of our knowledge this is the first attempt to integrate lower rank representation and landmark selection.

Related Work

As our proposed algorithm performs domain generalization method based on the use of randomized kernels, we briefly review these two lines of research in this section.

Domain generalization: Given several labeled training samples drawn from different sources with biased distributions, domain generalization assigns labels to target sets. Fluctuations in the distributions arise in a variety of applications due to technical, environmental, biological, or other sources of variation. This problem has been addressed in other areas of machine learning such as domain adaptation (Jiang 2008; Aljundi et al. 2015) and transfer learning (Pan and Yang 2010). However, they require the incorporation of target samples or even access to some of the target labels, while domain generalization can be performed independent of the target set.

Blanchard et al. (2011) first raised the domain generalization problem and proposed a kernel-based approach that identifies an appropriate RKHS and optimizes a regularised empirical risk over the space. Two projection-based algorithms, Domain-Invariant Component Analysis (DICA) and Unsupervised DICA (UDICA), were then developed by Muandet et al. (2013) to solve the same problem. DICA and UDICA extend Kernel PCA by incorporating the distributional invariance across domains to reduce the dissimilarity.

Domain generalization algorithms have also been studied by the computer vision community for object recognition. Khosla et al. (2012) proposed Undoing Dataset Bias (UDB), a multi-task max-margin classifier exploiting dataset-specific biases in feature space. The encoded biases are used to push each dataset's weight to be aligned with the global weights. Xu et al. (2014) proposed an exemplar SVM based method by exploiting the low-rank structure in the source. They formulated a new optimisation problem as a nuclear norm-based regularizer that captures the likelihoods of all positive samples. Niu et al. (2015) extended Xu et al. (2014) and proposed a multi-view domain generalization approach for visual recognition by fusing multiple SVM classifiers. They built upon exemplar SVMs to learn a set of SVM classifiers by using one positive sample and all negative samples in the source each time. Ghifary et al. (2015) proposed a multi-task autoencoder that leverages naturally occurring variation in sources as a substitute for the artificially induced corruption, and learns a transformation from the original image into analogs in multiple related sources. More recently, to overcome distribution variance across sources Erfani et al. (2016a) introduced ESRand, which incorporates random projection with elliptical data summarisation. While ESRand is reasonably efficient and delivers high accuracy, it requires at least $d+1$ samples for each source, where d is the number of features in the projected space. In some applications where the dimensionality of data is high, e.g., in images, collecting $d+1$ samples by each source may not be feasible.

Kernel randomisation: Various nonlinear kernel-machine formulations have been used to improve the capacity of learning machines while making learning feasible, e.g., quadratic programming (QP) solvers. In particular, these kernel-based methods rely on the computation of a kernel matrix over all pairs of data points, which limits the scalability of the algorithm on large datasets, and also can limit its effectiveness on high dimensional inputs, given the need to have sufficiently large training samples spanning the variation in the high dimensional space.

To address the scalability problems of kernel-machines, techniques have been proposed that either preprocess the data, e.g., by using dimensionality reduction techniques such as PCA or deep learning, or alleviate the QP problem, e.g., by breaking the problem into smaller pieces, for example by using chunking. A more recent trend explores the use of randomisation, such as linear random projection (Blum 2006) as a substitute for the computationally expensive step of kernel matrix construction. The work of Rahimi and Recht (2007; 2009) made a breakthrough in this approach. They replicated a Radial Basis Function (RBF) kernel by randomly projecting the data to a lower dimensional space and then used linear algorithms. Random projection avoids the complexity of traditional optimisation methods needed for nonlinear kernels. Recently, randomisation has been applied to other kernel methods, such as dot-product kernels (Kar and Karnick 2012), and one-class SVM (Erfani et al. 2015; 2016b).

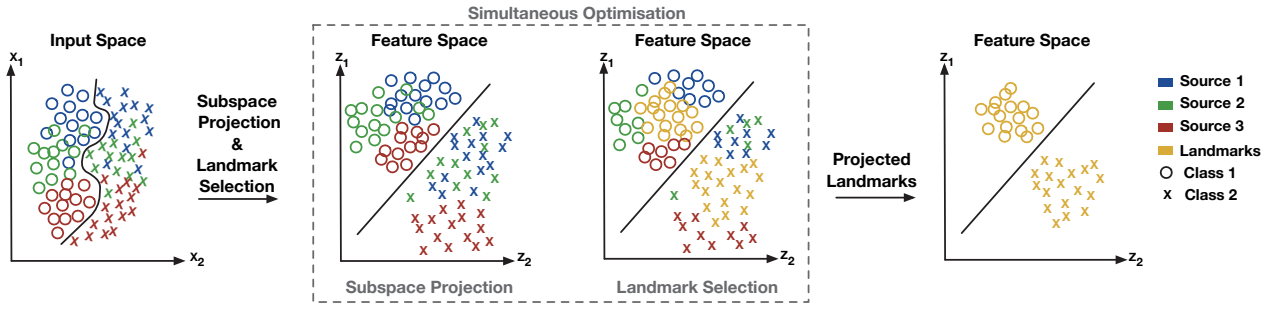

Figure 1: (Best shown in colour) Overall workflow of our approach. Generally, our objective is to find a projection to a subspace such that it minimizes the distribution bias between similar classes, while maintaining the distance between dissimilar ones. To further improve the generalization we select the observations that maximize the overlap between similar classes. Then a classifier (typically an SVM or k-NN) is trained on the reduced set.

Proposed Approach

In this section, we introduce our approach to multi-source classification. Our goal is to learn a representation of the data that is shared across different sources. The key idea behind our formulation is to simultaneously find a projection to a low-dimensional subspace, and select landmarks/samples from the existing data sources, such that the distance between the distributions of multiple sources is minimized. Intuitively, with such a representation, a classifier trained on the existing sources should perform equally well on the unseen data. Fig. 1 illustrates the overall workflow of our proposed approach. Before formally presenting our algorithms, we first elaborate the idea of Maximum Mean Discrepancy, which provides the foundation for the proposed algorithm.

Maximum Mean Discrepancy (MMD)

In our work, we are interested in measuring the distribution difference between multiple sources of data. Generally speaking, we can compare two probability distributions either through parametric models, or non-parametric ones. In the former methods, the probability distributions are first modelled, e.g., using Gaussian Mixture Models, and then the models are compared to measure the dissimilarity among the distributions. In non-parametric approaches, e.g., kernel density estimation, the probability distribution is estimated from observations without modelling the distributions explicitly.

To compare data distributions, we exploit a non-parametric approach, mainly because in our problem, modelling probability distributions is extremely difficult if not impossible. More specifically, our data (sensor and visual):

- exhibits very complex probability distributions. As such, an accurate model requires a large set of parameters to tune which is restrictive.
- is by default high-dimensional. Training effective probabilistic models for high-dimensional data is difficult and demands a large number of observations.

Therefore, we use the MMD between the distributions as a means to measure their dissimilarity.

Given $\boldsymbol{X}_p = \{\boldsymbol{x}_p^1, \cdots, \boldsymbol{x}_p^n\}$ and $\boldsymbol{X}_q = \{\boldsymbol{x}_q^1, \cdots, \boldsymbol{x}_q^m\}$ as two sets of i.i.d. observations from sources p and q, with m and n observations, respectively, the MMD criterion determines whether $p = q$ in RKHS.

Definition 1 *(Gretton et al. 2006) Let $\mathfrak{F}$ be a class of functions $f : \mathcal{X} \to \mathbb{R}$. Then the maximum mean discrepancy (MMD) and its empirical estimate are defined as:*

$$\mathrm{MMD}(\mathfrak{F}, p, q) = \sup_{f \in \mathfrak{F}} \left(E_{x \sim p}[f(x)] - E_{x \sim q}[f(x)] \right),$$

$$\mathrm{MMD}(\mathfrak{F}, \boldsymbol{X}_p, \boldsymbol{X}_q) = \sup_{f \in \mathfrak{F}} \left(\frac{1}{n} \sum_{i=1}^{n} f(\boldsymbol{x}_p^i) - \frac{1}{m} \sum_{j=1}^{m} f(\boldsymbol{x}_q^j) \right).$$

Clearly, the value of MMD depends on the function set $\mathfrak{F}$.

Theorem 1 *(Gretton et al. 2006) Let $\mathfrak{F}$ be a unit ball in RKHS, defined on compact metric space $\mathcal{X}$ with associated kernel $k(\cdot, \cdot)$. Then $MMD(\mathfrak{F}, p, q) = 0$ if and only if $p = q$.*

In short, the MMD between the distributions of two sets of observations is equivalent to the distance between the means of the two sets mapped into a high-dimensional, nonlinear feature space. We note that recently the characteristic RKHS (a more general RKHS compared to universal RKHS) has been used to assess the MMD (Sriperumbudur, Fukumizu, and Lanckriet 2011).

Multi-Source Classification

There are two popular approaches to multi-source classification: projecting all source samples to a common subspace where they share similar distributions (Baktashmotlagh et al. 2013; Pan et al. 2011), or selecting landmarks/samples from the source data in a way that the distribution distance between different sources will be minimized (Gretton et al. 2006). Here, we follow similar ideas, but unfiy the subspace projection and sample selection into a single optimisation problem, and show that this unified approach improves the classification accuracy.

More specifically, we simultaneously learn a subspace ($\boldsymbol{W}$) and identify landmarks ($\boldsymbol{\alpha}$) that minimize the distribution difference between multiple sources. We exploit MMD as a measure of the distance between the distribution of multiple sources, which lets us write our optimisation problem

as:

$$\min_{\boldsymbol{\alpha},\boldsymbol{W}} \left\| \frac{1}{\sum_{i=1}^{n}\alpha_i}\sum_{i=1}^{n}\alpha_i\phi(\boldsymbol{x}_i^p)\boldsymbol{W} - \frac{1}{m}\sum_{j=1}^{m}\phi(\boldsymbol{x}_j^q)\boldsymbol{W} \right\|_{\mathcal{H}} \quad \text{s.t.} \quad \alpha_i \in \{0,1\} \tag{1}$$

where $\phi(\cdot)$ is the mapping from $\mathbb{R}^D$ to the Hilbert space RKHS $\mathcal{H}$, $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_n]$ is the vector of binary variables indicating if a sample from source p is selected as landmark (e.g., $\alpha_i = 1 \rightarrow \boldsymbol{x}_i \in \text{Landmarks}$), and $\boldsymbol{W}$ is a subspace projection applied to the source samples.

We try to find a subspace and select samples shared among different sources of data, so that the distribution distance between multiple sources will be minimized. To enforce choosing the appropriate proportion of all the classes of the source data, we add another constraint to our optimisation problem: $\frac{1}{\sum_n \alpha_n}\alpha_n y_{n,c} = \frac{1}{n}\sum_n y_{n,c}$, where C is the number of classes and $y_{n,c}$ is a variable that determines whether the i^{th} source sample is a member of class c or not (Gong, Grauman, and Sha 2013).

It is intractable to solve the optimisation problem in (1) because of the binary constraints. Therefore, we solve the relaxed problem which can be expressed as

$$\min_{\boldsymbol{\beta},\boldsymbol{W}} \left\| \sum_{i=1}^{n}\beta_i\phi(\boldsymbol{x}_i^p)\boldsymbol{W} - \frac{1}{m}\sum_{j=1}^{m}\phi(\boldsymbol{x}_j^q)\boldsymbol{W} \right\|_{\mathcal{H}}, \quad \text{s.t.} \quad \beta_i \in [0,1]\,, and \quad \sum_{i=1}^{n}\beta_i = 1 \tag{2}$$

where the variable β_i replaces a binary variable $\alpha_i/(\sum \alpha_i)$.

The optimisation problem in (2) can be expressed in terms of a kernel function $k(\cdot,\cdot)$. We make use of the Gaussian kernel function which satisfies the universality condition of the MMD:

$$\begin{aligned}\min_{\boldsymbol{\beta},\boldsymbol{W}} \; & \sum_{i,j=1}^{n}\beta_i\boldsymbol{W}^T k(x_i^p, x_j^p)\boldsymbol{W} \\ & + \frac{1}{m^2}\sum_{i,j=1}^{m}\boldsymbol{W}^T k(x_i^q, x_j^q)\boldsymbol{W} \\ & - \frac{2}{m}\sum_{i,j=1}^{n,m}\boldsymbol{W}^T k(\beta_i x_i^p, x_j^q)\boldsymbol{W}\,,\end{aligned} \tag{3}$$

where $k(\cdot,\cdot)$ is the Gaussian kernel function $\exp\left(-\frac{(\boldsymbol{x}_i-\boldsymbol{x}_j)^T(\boldsymbol{x}_i-\boldsymbol{x}_j)}{\sigma}\right)$, and σ is the bandwidth in the Gaussian kernel.

The resulting optimisation problem is a non-convex problem, and also is cumbersome to solve for large scale datasets. To overcome this limitation, instead of solving the optimisation problem for $\boldsymbol{W}$, we refer to the results of (Lopez-Paz, Muandet, and Recht 2015), and project all source data to a random subspace.

We propose to exploit a lower rank approximation of (3) using nonlinear random Fourier features, which serves as a good approximation of the Gaussian (non-linear) kernel. For shift-invariant kernels we can exploit Bochner's theorem to generate h dimensional random features $\boldsymbol{Z} \in \mathbb{R}^{m\times h}$, and for $i = 1, \ldots, m$

$$\boldsymbol{z}_i = [\cos(\boldsymbol{r}_i^T\boldsymbol{x}_1 + b_i), \ldots, \cos(\boldsymbol{r}_i^T\boldsymbol{x}_h + b_i)]. \tag{4}$$

The vectors $(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_h)$ are sampled from the Fourier transformation, and $(b_1, \ldots, b_h) \sim \mathcal{U}(0, 2\pi)$.

Then (2) can be written as

$$\min_{\boldsymbol{\beta}} \left\| \sum_{i=1}^{n}\beta_i\boldsymbol{z}_i^p - \frac{1}{m}\sum_{j=1}^{m}\boldsymbol{z}_j^q \right\|_{\mathcal{H}}, \quad \text{s.t.} \quad \beta_i \in [0,1]\,, and \quad \sum_{i=1}^{n}\beta_i = 1 \tag{5}$$

where the mean in the RKHS reduces to

$$\tilde{\mu} = \frac{1}{m}\sum_{i=1}^{m}\boldsymbol{z}_i \in \mathbb{R}^h. \tag{6}$$

In practice, we project the data from all sources to a random subspace $\boldsymbol{W}$, solve (3) for the weights $\boldsymbol{\beta}$, and then enforce a threshold on the output variable $\boldsymbol{\beta}$ to obtain the binary weights $\boldsymbol{\alpha}$.

Empirical Evaluation

In this section, we compare the performance and efficiency of the proposed algorithm with state-of-the-art methods through classification tasks on multiple sensor and image benchmark datasets. In the implementation, we project all samples from the different sources to a shared random subspace, and find the samples/landmarks that are most similar among all the data sources. We use the resulting representation (projected landmarks) as the input to train a linear SVM classifier and k-NN.

Sensor Datasets: We use four real life activity recognition datasets from the UCI Machine Learning Repository: (*i*) Daily and Sport Activity (DSA), (*ii*) Heterogeneity Activity Recognition (HAR), (*iii*) Opportunity Activity Recognition (OAR), (*iv*) PAMAP2 Physical Activity Monitoring, with the number of 19, 6, 5, 13 activities collected from 8, 9, 4, 8 subjects, respectively[1].

Image Datasets: We use four set sof images from the Caltech-101 (C), LabelMe (L), SUN09 (S), and PASCAL VOC2007 (P) datasets. Each of the datasets represents a data source and they share five object categories: bird, car, chair, dog, and person. Instead of using the raw features as inputs to the algorithms, we used the DeCAF6 extracted features with the dimensionality of 4,096 [2].

[1] DSA, HAR and PAMAP2 are large datasets including millions of samples. We used a subset of these datasets. For DSA and PAMAP2 the first 1000 samples of each activity from each user were used, and for HAR the first 2000 samples were used.

[2] Available at: `http://www.cs.dartmouth.edu/~chenfang/proj_page/FXR_iccv13/index.php`

Table 1: Comparison of the leave-one-source-out classification accuracies for the sensor datasets. Bold-face values indicate the best performance for each dataset.

	$k-NN$				*l-SVM*							
Dataset	DICA	AE	CAE	Ours	DICA	AE	CAE	Ours	k-NN	SVM	UDB	LRE
DSA	87.81	90.11	**95.02**	94.16	87.18	91.63	93.68	94.67	87.46	86.15	88.61	91.73
HAR	68.44	76.25	**84.11**	**84.03**	63.31	76.69	83.15	**84.68**	65.27	73.95	75.86	80.41
OAR	73.35	78.88	84.92	87.18	74.42	76.17	86.35	**88.86**	71.57	71.42	76.68	79.15
PAMAP2	81.41	91.23	94.61	**96.70**	82.44	90.63	**96.53**	**96.08**	79.45	83.21	84.86	88.56
Avg.	77.75	84.12	89.67	90.52	76.84	83.78	89.93	**91.10**	75.94	78.68	81.50	84.96

Table 2: Comparison of the leave-one-source-out classification accuracies for the image datasets. Bold-face values indicate the best performance for each dataset.

		$k-NN$				*l-SVM*							
Train Set	Test Set	DICA	AE	CAE	Ours	DICA	AE	CAE	Ours	k-NN	l-SVM	UDB	LRE
C, L, S	P	59.26	60.01	**62.16**	61.80	59.14	59.10	61.86	**62.45**	59.03	58.86	54.29	60.58
C, L, P	S	56.34	57.50	**58.00**	**58.24**	55.81	57.86	**58.02**	57.31	55.09	49.09	54.21	54.88
C, S, P	L	53.47	57.63	59.32	**60.11**	55.11	58.20	59.67	59.80	52.64	52.49	58.09	59.74
L, S,P	C	85.89	86.44	88.12	88.43	86.05	86.67	**89.88**	**89.07**	84.73	77.67	87.50	88.11
Avg.		63.74	65.40	66.90	**67.15**	64.03	65.46	**67.36**	**67.16**	62.87	59.53	63.52	65.83

Baselines: To evaluate the performance and efficiency of our algorithm, we compare it with the following baseline methods: (*i*) **DICA** and **UDICA**: kernel-based optimisation algorithms that learn an invariant transformation to minimize the dissimilarity across domains, (*ii*) **AE** (Autoencoder) (Bengio et al. 2007): a basic autoencoder trained by stochastic gradient descent, (*iii*) **CAE** (Contractive Autoencoder) (Rifai et al. 2011): an autoencoder with an additional penalty, the Frobenius norm of the Jacobian matrix of the encoder activations with respect to the input, to yield robust features on the activation layer, (*iv*) k**-NN**: k Nearest Neighbour, we use $k = 1$, (*v*) l**-SVM**: Support Vector Machine with linear kernel, (*vi*) **UDB** (Khosla et al. 2012): a max-margin SVM-based framework for reducing dataset bias, (*vii*) **LRE-SVM** (Xu et al. 2014): a non-linear exemplar-SVMs model with a nuclear norm regularisation to impose a low-rank likelihood matrix.

The hyper-parameters of all the algorithms are adjusted using grid search based on their best performance on a validation set. Algorithms $i-iii$ are used for feature extraction. For classification purposes, the learnt features from these algorithms are used with k-NN and multi-class SVM with a linear kernel *l*-SVM. Since the focus of the experiment is to evaluate the effectiveness of the studied methods, we utilize simple classification algorithms, otherwise more advanced approaches can be employed. For algorithms $iv-vii$ no feature extraction has been conducted, and the algorithms have been applied directly on the (normalised) raw datasets.

Metric: We use the Receiver Operating Characteristic (ROC) curve and the corresponding Area Under the Curve (AUC) to measure the performance of all the methods. The reported training times are in seconds using MATLAB on an Intel Core i7 CPU at 3.60 GHz with 16 GB RAM. The stated AUC values and training times are the average of 10 folds for each experiment. For SVM based methods LIBSVM was used.

Table 3: Wilcoxon test to compare the performance of the top four algorithms regarding the p-values. The values in bold indicate that the null hypothesis is rejected for the corresponding method. R^+ corresponds to the sum of the ranks for the method on the first column, and R^- for our method. The $X(k)$ and $X(l)$ indicate the result of algorithm X using k-NN and l-SVM, respectively.

	Ours(k)			Ours(l)		
Method	R^+	R^-	p	R^+	R^-	p
AE(l)	0	36	**0.0078**	0	36	**0.0078**
AE(k)	0	36	**0.0078**	0	36	**0.0078**
DAE(l)	9	27	0.2500	11	25	0.3828
DAE(k)	18	27	0.2500	7	29	0.1484
UDB	0	36	**0.0078**	0	36	**0.0078**
LRE	0	36	**0.0078**	0	36	**0.0078**

Accuracy Evaluation

To assess the generalization ability of our algorithm across sources, we conduct our experiments on sensor and image datasets. All the records in each dataset are normalized between [0,1]. For each dataset, we take one subject (i.e., source) as the test set and the remaining subjects as the training set, i.e., leave-one-source-out, and repeat this for all the sources. In Table 1, due to the high number of sources in the sensor datasets, the average classification accuracy over all the sources is reported, while in Table 2 the accuracy for each individual source is reported. The stated values are the percentage accuracy. Since the accuracy results of UDICA and DICA are similar on these dataset, only the results of

DICA have been included in the tables.

Comparing the performance of the proposed algorithm with conventional machine learning algorithms, l-SVM and k-NN, the large performance gap, i.e., about 5% on average, for both sensor image datasets indicates that our method is effective in reducing distribution bias. This improvement can also be observed in the comparison with the domain generalization approaches, DICA, UDB, and LRE, however, our algorithm outperforms these approaches as well. Over all, our algorithm delivers the best performance on the benchmark datasets with an average accuracy of 91% and 67% for sensor and image datasets. The closest results are from CAE, with respectively 90% and 67% accuracy.

To statistically assess the significance of the performance between our algorithm and the top four algorithms, we use the Wilcoxon test. Table 3 summarizes these results. The p-value associated with each comparison represents the lowest level of significance of a hypothesis that results in a rejection. This value allows one to identify if two algorithms have significantly different performance and to what extent. The returned p-values for all the algorithms, except CAE, reject the null hypothesis for the accuracy measure with a level of significance of $\alpha = 0.05$, indicating the superiority of our algorithm over the compared methods. Although our algorithm is not statistically better than CAE, it delivers much higher ranks (R^-) in this comparison. Fig. 2 illustrates the behaviour of our algorithms on the five image classes. From the left, the first column shows some example of images misclassified by all the approaches, the second and third columns show the images that were misclassified by SVM and CAE, but correctly labeled with our algorithm.

A possible explanation for effectiveness of our algorithm can relate to the dimensionality of the manifold in feature space where samples concentrate. We hypothesise that if features concentrate near a low dimensional sub-manifold, then the algorithm has found invariant features and will generalize well. Moreover, the landmark selection step eliminates the noisy records and outliers, giving a boost to the generalization.

Efficiency Evaluation

To improve generalization, our algorithm substantially reduces the number of features as well as the number of samples. To study this impact, we compare the training time of our algorithm with CAE, which has the second best accuracy, and l-SVM. In this experiment we use a sensor dataset, OAR, and a set of image samples including the L, P, and S datasets. Figure 3 demonstrates the result of this comparison. Comparing the training time of l-SVM with our algorithm, the advantage of our subspace landmark selection is immediately revealed. Reducing both the number of features and samples substantially diminishes the training time. Even in comparison with CAE, the training time of our algorithm increases at a lower rate. The training time of these three methods are comparable only when the size of data is small, e.g., when in OAR the number of records is less than 5000. But in larger datasets like images where the dimensionality of data is usually high, even for small numbers of records our algorithm runs about twice as fast.

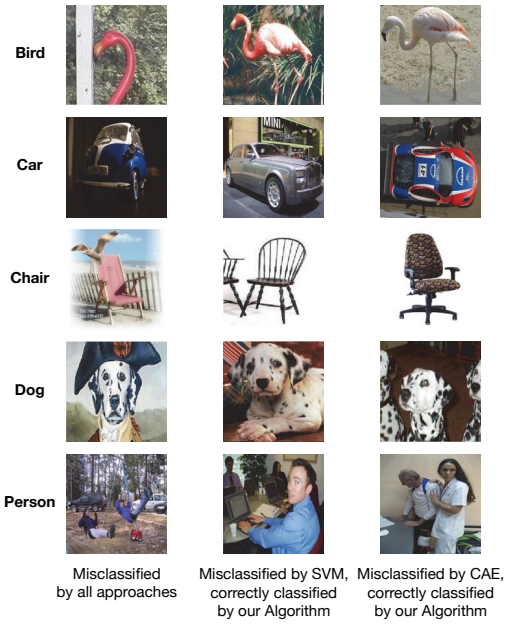

Figure 2: Some examples of misclassified test samples from different sources.

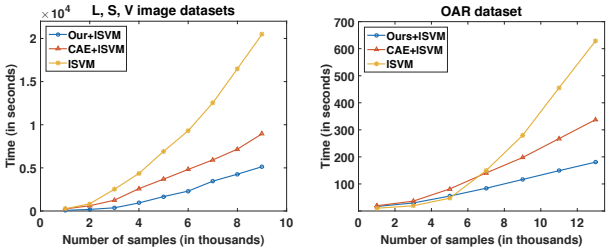

Figure 3: Comparison of the training time of our algorithm with CAE and l-SVM.

Conclusion

Distribution similarity is central to the multi-source learning problem. The need for adaptive classifiers arises in many application domains, especially in IoT applications involving a variety of devices. While existing approaches focus on correcting shift-distribution between data sources by learning a projection to a latent space, we have advanced the field by proposing a unified approach to subspace learning and landmark selection. At the core is the idea of exploiting landmarks in a lower dimensional space, and identifying samples from the training sources that share grater statistical similarity within this space. We applied the model to sensor and visual benchmark datasets and empirically verified the convergence of the training algorithm. The results are very promising, they are on par or better than state-of-the-art methods in classification accuracy, and with significant gains in terms of training time. In future work, we will explore the performance of our algorithm in other areas of machine learning such as streaming video data, where the appearance of objects transforms in real-time.

References

Aljundi, R.; Emonet, R.; Muselet, D.; and Sebban, M. 2015. Landmarks-based kernelized subspace alignment for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 56–63.

Baktashmotlagh, M.; Harandi, M.; Lovell, B.; and Salzmann, M. 2013. Unsupervised domain adaptation by domain invariant projection. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 769–776.

Bengio, Y.; Lamblin, P.; Popovici, D.; Larochelle, H.; et al. 2007. Greedy layer-wise training of deep networks. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, volume 19, 153–160.

Blanchard, G.; Lee, G.; and Scott, C. 2011. Generalizing from several related classification tasks to a new unlabeled sample. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2178–2186.

Blum, A. 2006. Random projection, margins, kernels, and feature-selection. In *Proceedings of Subspace, Latent Structure and Feature Selection*. 52–68.

Erfani, S. M.; Baktashmotlagh, M.; Rajasegarar, S.; Karunasekera, S.; and Leckie, C. 2015. R1SVM: a Randomised Nonlinear Approach to Large-Scale Anomaly Detection. In *Association for the Advancement of Artificial Intelligence (AAAI)*.

Erfani, S. M.; Baktashmotlagh, M.; Moshtaghi, M.; Nguyen, V.; Leckie, C.; Bailey, J.; and Ramamohanarao, K. 2016a. Robust domain generalisation by enforcing distribution invariance. In *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*.

Erfani, S. M.; Baktashmotlagh, M.; Rajasegarar, S.; Nguyen, V.; Leckie, C.; Bailey, J.; and Ramamohanarao, K. 2016b. R1STM: One-class support tensor machine with randomised kernel. In *Proceedings of SIAM International Conference on Data Mining (SDM)*.

Ghifary, M.; Bastiaan Kleijn, W.; Zhang, M.; and Balduzzi, D. 2015. Domain generalization for object recognition with multi-task autoencoders. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2551–2559.

Gong, B.; Grauman, K.; and Sha, F. 2013. Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation. In *Proceedings of International Conference on Machine Learning (ICML)*, 222–230.

Gretton, A.; Borgwardt, K. M.; Rasch, M.; Schölkopf, B.; and Smola, A. J. 2006. A kernel method for the two-sample-problem. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 513–520.

Gretton, A.; Borgwardt, K. M.; Rasch, M. J.; Schölkopf, B.; and Smola, A. 2012. A kernel two-sample test. *Journal of Machine Learning Research (JMLR)* 13(1):723–773.

Jiang, J. 2008. A literature survey on domain adaptation of statistical classifiers. *URL: http://sifaka. cs. uiuc. edu/jiang4/domainadaptation/survey*.

Kar, P., and Karnick, H. 2012. Random feature maps for dot product kernels. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 583–591.

Khosla, A.; Zhou, T.; Malisiewicz, T.; Efros, A. A.; and Torralba, A. 2012. Undoing the damage of dataset bias. In *Proceedings of European Conference on Computer Vision (ECCV)*. 158–171.

Lopez-Paz, D.; Muandet, K.; and Recht, B. 2015. The randomized causation coefficient. *Journal of Machine Learning Research (JMLR)* 16:2901–2907.

Muandet, K.; Balduzzi, D.; and Schölkopf, B. 2013. Domain generalization via invariant feature representation. In *Proceedings of International Conference on Machine Learning (ICML)*, volume 28, 10–18.

Niu, L.; Li, W.; and Xu, D. 2015. Multi-view domain generalization for visual recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 4193–4201.

Pan, S. J., and Yang, Q. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering (TKDE)* 22(10):1345–1359.

Pan, S. J.; Tsang, I. W.; Kwok, J. T.; and Yang, Q. 2011. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks* 22(2):199–210.

Rahimi, A., and Recht, B. 2007. Random features for large-scale kernel machines. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 1177–1184.

Rahimi, A., and Recht, B. 2009. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 1313–1320.

Rifai, S.; Vincent, P.; Muller, X.; Glorot, X.; and Bengio, Y. 2011. Contractive auto-encoders: Explicit invariance during feature extraction. In *Proceedings of International Conference on Machine Learning (ICML)*, 833–840.

Sriperumbudur, B. K.; Fukumizu, K.; and Lanckriet, G. R. 2011. Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research (JMLR)* 12:2389–2410.

Stisen, A.; Blunck, H.; Bhattacharya, S.; Prentow, T. S.; Kjærgaard, M. B.; Dey, A.; Sonne, T.; and Jensen, M. M. 2015. Smart devices are different: Assessing and mitigating-mobile sensing heterogeneities for activity recognition. In *Proceedings of ACM Conference on Embedded Networked Sensor Systems (SenSys)*, 127–140.

Torralba, A., and Efros, A. A. 2011. Unbiased look at dataset bias. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1521–1528.

Xu, Z.; Li, W.; Niu, L.; and Xu, D. 2014. Exploiting low-rank structure from latent domains for domain generalization. In *Proceedings of European Conference on Computer Vision (ECCV)*. 628–643.