Day and Khoshgoftaar *J Big Data* (2017) 4:29
DOI 10.1186/s40537-017-0089-0

Journal of Big Data

SURVEY PAPER Open Access

A survey on heterogeneous transfer learning

Oscar Day* and Taghi M. Khoshgoftaar

*Correspondence:
oday2015@fau.edu
Florida Atlantic University
College of Engineering
and Computer Science, Boca
Raton, USA

Abstract

Transfer learning has been demonstrated to be effective for many real-world applications as it exploits knowledge present in labeled training data from a source domain to enhance a model's performance in a target domain, which has little or no labeled target training data. Utilizing a labeled source, or auxiliary, domain for aiding a target task can greatly reduce the cost and effort of collecting sufficient training labels to create an effective model in the new target distribution. Currently, most transfer learning methods assume the source and target domains consist of the same feature spaces which greatly limits their applications. This is because it may be difficult to collect auxiliary labeled source domain data that shares the same feature space as the target domain. Recently, heterogeneous transfer learning methods have been developed to address such limitations. This, in effect, expands the application of transfer learning to many other real-world tasks such as cross-language text categorization, text-to-image classification, and many others. Heterogeneous transfer learning is characterized by the source and target domains having differing feature spaces, but may also be combined with other issues such as differing data distributions and label spaces. These can present significant challenges, as one must develop a method to bridge the feature spaces, data distributions, and other gaps which may be present in these cross-domain learning tasks. This paper contributes a comprehensive survey and analysis of current methods designed for performing heterogeneous transfer learning tasks to provide an updated, centralized outlook into current methodologies.

Keywords: Transfer learning, Heterogeneous transfer learning, Knowledge transfer, Supervised learning, Semisupervised learning, Unsupervised learning

Introduction

Machine learning is of increasing importance due to its success and benefit in real-world applications. Models used in machine learning are trained from a series of examples comprised of features/attributes that are associated with a single label. This label can be a class value for classification tasks or a numerical value for regression tasks [1]. When faced with unsupervised tasks, the class labels are not provided during training which can make the training process more challenging. Once these models are trained we can then apply them to predict the value for a newly arriving, unseen instance. Also, if the ground truth label is available, we can compare it to the predicted value as to calculate performance metrics [2] for the model.

Traditional machine learning operates under the assumption that the training and testing data are taken from the same input feature space and the same data distribution [3]. However, this assumption may not hold when faced with real-world scenarios. The feature space may differ in the manner that the training data may hold a specific set of features, but the testing data may have a feature space of different dimensions, or its features may represent different attributes entirely. The data distribution may also differ in the manner that, if the training and testing data were collected from different domains, the marginal and/or conditional probability distributions [4] may differ. As an example, consider the case described in [5], where a model is trained to classify records of Book A into predefined categories and is then tested using records from Book B. In this case, the model will have degraded performance because the training and testing data were taken from different data distributions, since each book has a different variety of words, sentences, etc. When the distribution changes, most statistical models need to be rebuilt completely using new labeled training data from the new distribution. This process is often expensive and difficult due to the effort of collecting sufficient labeled data to train an effective new model [6]. Thus, there is a need for a method to create a high-performance model for a target domain in a different distribution without such a significant amount of labeling effort. This can be achieved using transfer learning.

Transfer learning

Transfer learning (TL) [3, 6, 7] aims to produce an effective model for a target task with limited or no labeled training data by leveraging and exploiting knowledge from a different, but related source domain to predict the truth label for an unseen target instance. Due to insufficient labeled instances, training a model in this target task would result in degraded performance as compared to a model trained with sufficient labeled data. However, by enhancing the training with supplementary labeled data from a related source domain, the model's ability to classify target instances can be improved. The challenge becomes how to distinguish beneficial knowledge in a source domain from the inherent cross-domain noise, due to the varied distributions, and apply it to a target domain. Transfer learning can be split into two main categories when it comes to the feature spaces: homogeneous and heterogeneous transfer learning.

Homogeneous transfer learning

In homogeneous transfer learning, the feature spaces of the data in the source and target domains are represented by the same attributes ($X_s = X_t$) and labels ($Y_s = Y_t$) while the space itself is of the same dimension ($d_s = d_t$). This method thus focuses on bridging the gap in the data distributions between the domains as experienced in cross-domain transfer [7]. Overall, as described in [3] and [7], homogeneous transfer learning solutions can be organized into five categories: Instance-based, feature-based (symmetric or asymmetric), model-parameter-based, relational-informational-based and hybrid-based approaches. An in-depth review of these approaches is described in [7]. It can be also noted that much of the current literature addresses the problem of homogeneous domain adaptation. In this scenario one is performing a single, common task, but under a different domain and the goal is to reduce the accuracy drop due to the distribution shift.

Heterogeneous transfer learning

The other category of transfer learning is Heterogeneous Transfer Learning (HTL). In this scenario, the feature spaces between the source and target are nonequivalent and are generally non-overlapping. In this case, $X_s \neq X_t$ and/or $Y_s \neq Y_t$ as the source and target domains may share no features and/or labels, while the dimensions of the feature spaces may differ as well. This method thus requires feature and/or label space transformations to bridge the gap for knowledge transfer, as well as handling the cross-domain data distribution differences. This case is more challenging, as there are fewer representational commonalities between the domains. In other words, knowledge is available from source data but it is represented in a different way than that of the target. The challenge becomes how to extract it.

Most heterogeneous transfer learning solutions fall into two categories when it comes to transforming the feature spaces: symmetric and asymmetric transformation. Symmetric transformation, illustrated in Fig. 1a, takes both the source feature space X_s and target feature space X_t and learns feature transformations as to project each onto a common subspace X_C for adaptation purposes. This derived subspace becomes a domain-invariant feature subspace to associate cross-domain data, and in effect, reduces marginal distribution differences. Performing this brings the feature spaces for both domains together into a common feature representation where one can then apply traditional machine learning models such as Support Vector Machines (SVM). Optimally, one can also apply models built for homogeneous transfer learning which consider the distribution differences and domain transfer ability observed in the subspace. Asymmetric transformation mapping, illustrated in Fig. 1b, transforms the source feature space to align with that of the target ($X_t \rightarrow TX_s$) or the target to that of the source ($X_s \rightarrow TX_t$). This, in effect, bridges the feature space gap and reduces the problem into a homogeneous transfer problem when further distribution differences need to be corrected. This approach is most appropriate when the source and target have the same class label space and one can transform X_s and X_t without context feature bias. Context feature bias occurs when there are conditional distribution differences between the domains as a feature in one domain may have a different meaning in another. In either category, once the issue of varied feature spaces is resolved we may need to solve marginal and/or conditional distribution differences. This can be done through homogeneous adaption solutions which account for these distribution differences observed during cross-domain tasks.

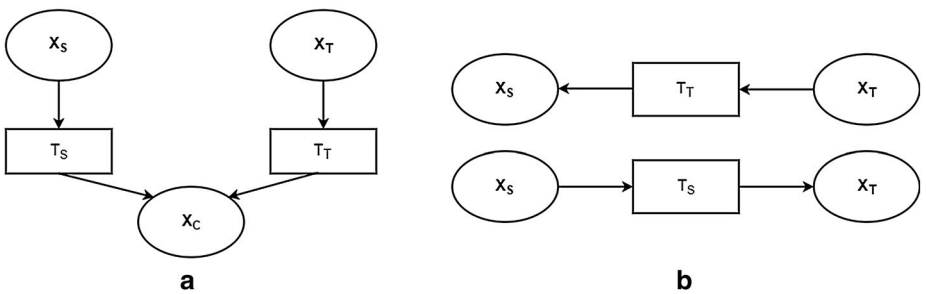

Fig. 1 Illustration of symmetric (**a**) and asymmetric (**b**) feature transformation for HTL

Domain Adaptation (DA) tasks, as previously discussed, may also be present as HTL problems and is known as Heterogeneous Domain Adaptation (HDA). This task is addressed by most of the current methods.

Negative transfer

The goal for transfer learning is to enhance the performance of a target task using an auxiliary/source domain. The issue becomes that sometimes transferring knowledge from a source domain may have a negative impact on the target model. This is also known as negative transfer. In such case, if one were to create a standard classifier, using only the limited labeled target data available for the application, then it would have better performance than a transfer model using both the limited labeled target data and the source data. This occurs mostly when the source domain has very little in common to the target. For transfer learning to be successful it is assumed that the source and target are related in some way. The less related they are, the more cross-domain noise is experienced. Thus, less performance enhancing knowledge is identified and extracted. Rosenstein et al. [8] demonstrated this by showing how transferring between dissimilar domains causes performance loss. Some homogeneous transfer learning solutions employ safeguards against negative transfer as reviewed in [7]. On the other hand, most current heterogeneous transfer learning methods do not address this issue as to be later discussed.

Big data application

The theoretical foundation of transfer learning is data-size independent and, although not extensively investigated, may be applied to big data [9, 10] to achieve the same benefits as within normal data environments. Specifically, both heterogeneous and homogeneous transfer learning methods are applicable to big data scenarios. This is because one can leverage these methods to enhance a target task in a big data environment with a source domain. As described by [7], transfer learning is especially attractive in the big data environment because, due to the growth of big data repositories, one can enhance their machine learning task by using an available dataset from a similar domain. In doing so, one can avoid the costly effort of collecting new labeled data which is especially apparent in the big data scope.

Paper overview/contributions

In this section, we will review the notations and organization used in this paper. The motivation for this survey paper is to provide a comprehensive, centralized overview of current *heterogeneous* transfer learning methodologies. The following outlines the contributions of our paper.

- Our main contribution is a comprehensive survey of 38 methods for heterogeneous transfer learning which operate under varied settings, requirements, and domains. This provides the centralized outlook into current methodologies.
- Second, we present an in-depth discussion and analysis of the surveyed methods. This is done via review and comparison performed amongst the methodologies as well as their limitations.

- Third, we present the shortcomings of current research in this domain as well as future work for further investigation.

Unlike previous surveys related to transfer learning, we focus distinctly on *heterogeneous* transfer learning and its related challenges. Specifically, [6] only briefly introduces the idea of HTL without surveying any of its methodologies while our survey goes in-depth into HTL, its challenges, and its current methods. Also, the survey conducted by [7] has a greater focus on general transfer learning methodologies and reviews only a few of the current methodologies for HTL, while we survey and analyze over three times as many. Thus, our unique focus on heterogeneous transfer learning provides a more comprehensive study across state-of-the-art methodologies and provides greater insight over other related surveys into this growing domain.

When faced with a cross-domain transfer learning task it is important to understand one's requirements as to select an appropriate method. Because labeling efforts are often expensive, one may have little or no available training labels in the target domain. Thus, this paper intuitively organizes the reviewed methods around target and source label requirements so a machine learning practitioner or researcher can easily select an appropriate HTL method based on the availability of labels in their target task.

In the literature, it can be noted that there are opposing definitions for terms used to describe label requirements [7]. For example, Liu et al. [11] use the notation of supervised transfer learning to denote a fully-labeled source domain and a limited labeled target, while unsupervised transfer learning denotes a mostly labeled source with no target domain labels. Conversely, the work of Cook and Feuz [12] who use supervised and unsupervised transfer learning to denote the presence or absence of labels respectively only in the source domain. Furthermore, they use the notation of informed and uninformed to denote the presence or absence of labels respectively in the target domain. Due to these inconsistencies, we denote the categories explicitly by their label requirements to avoid confusion. The surveyed HTL methods are organized into the following categories: "Methods which require limited target labels", "Methods which require limited target labels and accept unlabeled target instances", "Methods which require no target labels", "Methods which require limited target labels and no source labels", "Methods which no target or source labels", as well as an "HTL preprocessing method". In Tables 1, 2, 3, 4, we provide an overview of the surveyed methods in these categories. Furthermore, we present a deeper analysis into these methods in the "Discussion" section, followed by the "Conclusion", and finalized with "Future work".

Methods which require limited target labels

In this section, we survey various techniques which require labeled source data and limited labeled target data. Most of the current literature does not define how many instances would be considered "limited", but they assume it to be too few to create an effective standard classifier. In such case, transfer learning is required to enhance performance using an auxiliary, label-abundant domain.

Table 1 Surveyed HTL methods which require limited target labels

Methods	Characteristics	Sections
COTL [13]	Online tasks, multiview ensemble	"COTL, OHTWC"
OHTWC [19]	Requires co-occurrence data, online tasks, weighted ensemble	"COTL, OHTWC"
HFA [23]	Symmetric transformation w/augmentation	"HFA"
HMCA [29]	Requires source classifier, model-transfer	"HMCA"
SHFR [31]	Asymmetric, multi-class	"SHFR"
TTI [36]	Requires co-occurrence data, translator, image classification	"TTI"
TLRisk [37]	Requires co-occurrence data, translator, Markov chain, risk minimization	"TLRisk"
ARC-t [28]	Asymmetric, image classification, kernel methods	"ARC-t"
SHDA-RF [41]	Asymmetric, random forest and label pivots	"SHDA-RF"
IFSR [12]	Asymmetric, domain-independent metafeatures to compute similarity	"FSR (IFSR, UFSR, ELFSR)"
ELFSR [12]	Asymmetric, FSR ensemble, multi-source, stacking method	"FSR (IFSR, UFSR, ELFSR)"
SMVCCAE [64]	Symmetric, multi-view ensemble, CCA analysis	"SMVCCAE, SSMVCCAE"

Table 2 Surveyed HTL methods which require limited target labels and accept unlabeled target data

Methods	Characteristics	Sections
DAMA [27]	Symmetric, manifold alignment with labels, multi-source	"DAMA"
CDLS [46]	Symmetric and landmark weights	"CDLS"
IDL for HDA [49]	Online tasks, symmetric eigenanalysis-based	"IDL for HDA"
MOMAP [51]	Asymmetric, mapping by rotation and translation, multi-class, multi-source	"MOMAP"
HeMap [26]	Symmetric, spectral mapping Bayesian method, cluster-based sampling	"HeMap"
Proactive HTL [54]	Symmetric, label embeddings, proactive learning	"Proactive HTL"
SHFA [47]	Symmetric transformation w/augmentation for semi-supervised	"SHFA"
CT-Learn [57]	Requires co-occurrence data, joint transition probability graph, Markov random walk, multi-source	"CT-Learn"
SSKMDA [60]	Instance-based asymmetric, kernel matching	"SSKMDA"
SCP-ECOC [62]	Symmetric, multi-class, ECOC scheme	"SCP-ECOC"
MMDT [48]	Asymmetric, image, max-margin, multi-class	"MMDT"
SSMVCCAE [64]	Symmetric, multi-view ensemble, CCA analysis, SRKDA	"SMVCCAE, SSMVCCAE"
TNT [65]	Neural network-based mapping and classification	"TNT"
HDANA [67]	Symmetric, deep learning, autoencoder mapping	"HDANA"

Table 3 Surveyed HTL methods which require no target labels

Methods	Characteristics	Sections
CT-SVM [70]	Symmetric, CCA, transfer SVM	"CT-SVM"
HHTL [71]	Asymmetric, requires source-target correspondence data, deep learning, mSDA	"HHTL"
HDCC [76]	Symmetric, CCA, group-weighing, video annotation, multi-source	"HDCC"
CL-SCL [78]	Symmetric, structural correspondence learning, text classification	"CL-SCL"
HDP [80]	Asymmetric through metric selection and matching	"HDP"
FuzzyTL [85]	Fuzzy logic, intelligent environments, FIS	"FuzzyTL"
UFSR [12]	Asymmetric, domain-dependent metafeatures to compute similarity	"FSR (IFSR, UFSR, ELFSR)"
ELFSR [12]	Asymmetric, FSR ensemble, multi-source, voting method	"FSR (IFSR, UFSR, ELFSR)"
RLG, GLG [11]	Symmetric, LMM, Grassmann manifold	"RLG, GLG"

Table 4 Other surveyed HTL methods

Methods	Category	Characteristics	Sections
HTLIC [22]	Unlabeled source, limited target	Image classification, bipartite graph, matrix factorization, unlabeled source of text and annotated images	"HTLIC"
aPLSA [96]	Unlabeled source and target	Unsupervised, image clustering, PLSA extension, annotated auxiliary images	"aPLSA"
DCN [103]	Preprocessing	Determine relatedness of domains using co-occurrence data, multi-source	"DCN"

COTL, OHTWC

Zhao et al. [13] proposed a framework called Online Transfer Learning (OTL) to address online transfer learning tasks under homogeneous and heterogeneous scenarios. Specifically, the heterogeneous approach is called Co-regularized OTL (COTL). Standard machine learning assumes that training data is provided all at once in a batch manner, but this assumption may not always hold in real-world applications. Online learning [14] tasks are designed for when training instances arrive in an online/sequential manner. Online transfer learning aims to transfer knowledge from an offline source domain to an online target learning task which is represented by a different feature space. During these online HTL tasks, OTL assumes that the feature space of the source is a subset of the target domain and the procedure to solve these tasks is as follows. First, a classifier is built from the labeled source data, denoted as $h(x)$, using regular supervised learning with SVM or online learning via the Perceptron algorithm [15, 16]. The method then proposes to use a multi-view approach for the target data. Multi-view [17] theory describes how the same data can be viewed through different representations or "views". Two classifiers are built for two newly constructed views for the target data (denoted as f_1 and f_2) where f_1 is initialized to the source prediction function h and f_2 is initialized to 0. The predicted label is calculated as a function of these two views thus creating an ensemble classifier. For each new arriving training instance, both functions are updated using co-regularization optimization. If an arriving target training instance is incorrectly predicted then the ensemble is updated through the optimization procedure which aims to classify the next new example correctly without deviating too much from the original ensemble through regularization terms. This method thus exploits source domain knowledge and enhances an online learning target task having both domains represented by differing feature spaces. Experiments were performed comparing the proposed COTL against the Passive Aggressive algorithm (PA) [18], PA initialized to the source, and COTL with both views initialized to 0. The proposed COTL demonstrated superior performance by producing lower mistake rates, i.e. higher accuracy, on benchmark datasets which proves the method as an effective technique for knowledge transfer for online learning tasks.

Yan et al. [19] also proposed a method for online HTL learning tasks called Online Heterogeneous Transfer with Weighted Classifiers (OHTWC). This method proposes using unlabeled co-occurrence data to serve as a bridge between the two domains. The way this method works is by first creating an offline source classifier based on the similarity relationship between the instances of the source and target through the co-occurrence data. Using Pearson correlation [20], the similarity is measured between the new

arriving target instance and its co-occurred counterpart. Then, the similarity between the source and its co-occurred counterpart is measured. Therefore, combining these two measures will provide us with the similarity of the arriving target instance with the source from a different feature space. The prediction for this classifier is done through a weighted sum of the k-nearest neighbors of the source instances deemed most similar.

A second classifier is then constructed for the target instances using the online Passive Aggressive [18] algorithm. This classifier is updated for each arriving training instance based on whether it predicted the instance correctly or not. The updating is done through a positive regularization parameter and a hinge loss function. Having these two classifiers, the final prediction for arriving instances is done as an ensemble of the two. A Hedge [21] strategy is used to dynamically update weights assigned to each of the classifiers. This is done by changing the voting weights based on the loss suffered from each classifier respectively. The larger the loss, the larger the weight decrease occurs to such classifier causing the more accurate classifier to have a higher weight for prediction purposes. Figure 2, adapted from [19], provides an illustration of this method for using weighted online and offline classifiers to perform an online task. An experiment for text-to-image classification was performed to test the effectiveness of the proposed algorithm. The proposed algorithm was compared against PA [18], SVM, HTLIC [22], and PA using k-NN on co-occurrence data. The results show that the proposed OHTWC algorithm achieved best performance against the baselines for most cases by having the lowest error rates.

OHTWC differs from COTL in the manner that OHTWC uses a weighted ensemble rather than a co-regularized multi-view approach. It also uses co-occurrence data to link the source and target domains while COTL does not since COTL only uses the source and target data.

HFA

Duan et al. [23] proposed an HDA method called Heterogeneous Feature Augmentation (HFA). It aims to solve scenarios when the source and target domain are represented

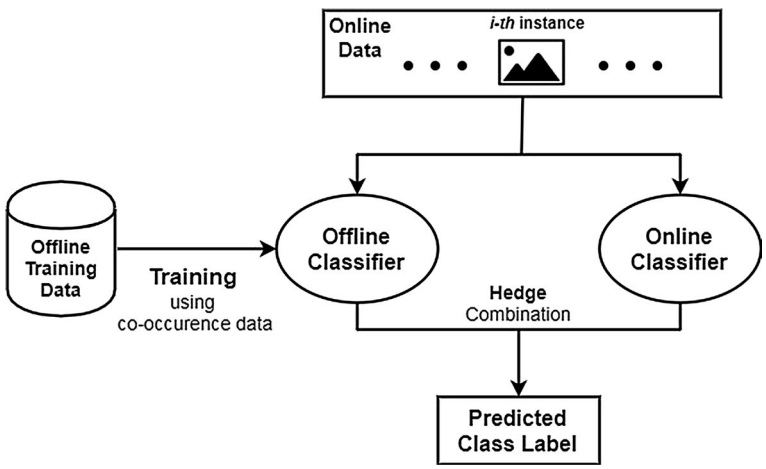

Fig. 2 Illustration of OHTWC adapted from [19] showing the use of offline and online classifiers to predict labels for an online task

by heterogeneous features and dimensions. This method uses a labeled source with a limited labeled target along with a symmetric transformation for the feature space differences. First, the data from the source and target is transformed into a common subspace using projection matrices P and Q respectively. These are found using standard SVM with hinge loss in both the linear and nonlinear cases. Then, two new feature mapping functions are proposed to augment the transformed data in the common latent space with their original features. This is illustrated in Eq. (1), adapted from [23]. These new feature mappings, ϕ_s and ϕ_t, consist of the transformed features (Px^s and Qx^t) to create the common subspace, the original features (x^s and x^t) used to augment the subspace, and zeroes (0_{ds} and 0_{dt}) to account for dimensional differences. In other words, the original source features will be added to the common subspace along with zeroes for the original target features and the original target features will be added to the subspace along with zeroes for the original source features as to match the differences in dimensions. Thus, creating an augmented common feature space.

$$\varphi_s\left(x^s\right) = \begin{bmatrix} Px^s \\ x^s \\ 0_{d_t} \end{bmatrix}, \quad \varphi_t\left(x^t\right) = \begin{bmatrix} Qx^t \\ 0_{d_s} \\ x^t \end{bmatrix} \tag{1}$$

This method is expanded from Feature Replication (FR) which was first proposed by Daume III [24] to offset the conditional distribution differences between the domains. FR aims to solve the HDA problem by padding with extra zeroes to make the dimensions from the two domains the same while HFA uses this methodology but extends it by utilizing the latent common features. Duan et al. [23] also proposed a transformation metric H which combines P and Q in efforts to simplify the optimization problem. Therefore, it is only necessary to solve for H making the subspace a latent subspace. The optimization problem is solved to develop a final target prediction function to predict newly arriving target instances. This is done through an alternating optimization procedure which simultaneously solves the dual problem of the SVM and finds the optimal transformation metric H. Experiments were performed for object recognition and text categorization tasks. The proposed HFA method was compared with a standard SVM trained on the target, KCCA [25], HeMap [26], DAMA [27], and ARC-t [28]. The proposed method demonstrated effectiveness for these HDA tasks by having higher classification accuracy averaged over ten rounds of the experiments.

HMCA

Mozafari et al. [29] proposed a framework called Heterogeneous Max-margin Classifier Adaptation (HMCA) which addresses heterogeneous and homogeneous domain adaptation problems. This method uses model-transferring for these DA tasks and is an extension from their previous work [30]. Model-transferring domain adaptation methods use a previously trained source classifier for adaptation purposes on a target domain. These methods require trained source classifiers and the target data during training thus the source training samples are not required for building the classifier in the target. This is because, in this case, the parameters from the source classifier are used in the adaptation process rather than the source samples. HMCA learns a max-margin classifier for the target domain, and adapts it according to the source classifier's pre-learned offset.

This adaption occurs through a modification in the target's SVM objective function which minimizes the distance of the target's offset from the source's offset. To do this, the source classifier and the labeled target domain data are projected onto a one-dimensional space. This offset can be seen as the discrimination point which the classifier uses to separate and distinguish the two classes in the one-dimensional space. Thus, to utilize source knowledge and adapt the target classifier utilizing the source model, a linear SVM is found whose offset discriminates the target samples correctly in the one-dimensional space. This is done under the condition that it also must have the minimum distance to the offset of the source as to maximize the similarity between the domains.

Since only the classifier offset is required from the source domain to adapt for the target, this method works for domains represented by heterogeneous features. Even so, HCMA has limitations though as it faces issues with overfitting (when there is high dimensionality and low target samples), noise, and outliers. It also makes two assumptions. The first is that the source and target classifiers be built with the same ratio of positive and negative samples, an assumption which may not always hold in real-world situations. The second is that the two domains hold an Intra-Sample Distance Pattern (ISDP) condition. ISDP is used here as a measure of correspondence between samples in the source and target domains. Two domains with similar ISDP mean they are highly similar as a sample from one domain corresponds to that of another. In other words, it is assumed that the source and target have similar conditional distributions in the one-dimensional space. This ISDP assumption does not always hold in real-world scenarios, though it does act as a method to measure the domain adaptability of source classifiers as to select the most similar source domain to that of the target when multiple source classifiers are available. Experiments were conducted under homogeneous and heterogeneous scenarios and the proposed HMCA showed increased accuracy rates for pedestrian detection and image classification tasks.

SHFR

Zhou et al. [31] proposed a method for HDA tasks when more than two classes are present. The proposed method is called Sparse Heterogeneous Feature Representation (SHFR) and it aims to solve the HDA problem through a sparse feature transformation matrix G. This is an asymmetric transformation to map the weight vectors learned from the binary classifiers of the source domain (as multi-class problems are commonly decomposed into multiple binary classifiers) to those of the target. The transformation G, is learned using multi-task learning from [32] and the goal is to minimize the distance between the weight vectors as to reduce the difference between the domains after transformation. In other words, both the source and target have assigned weight vectors to their classifiers and a transformation metric is constructed which minimizes the difference between the transformed source and target vectors. Learning this feature mapping is based on two assumptions: (1) the feature mapping G is highly sparse and (2) the transformation is class-invariant meaning all classes share the same mapping. This sparse, class-invariant transformation helps reduce bias of weak binary classifiers and improve overall performance. The learning task is conducted as a compressed sensing [33] problem (CS).

Through CS theory, it can be shown that the estimation error of the transformation matrix decreases with an increased number of classifiers. The issue becomes how to

create enough binary classifiers for optimal performance without making redundant, extra classifiers. The method proposes using the Error Correcting Output Codes [34] (ECOC) scheme to generate the binary classifiers. This method increases the robustness and accuracy of SHFR as it uses a voting scheme for prediction. This is particularly useful as not all the learned binary target classifiers will have high predictive capabilities due to limited labeled training data. As long as there are more correct classifiers than incorrect, then the prediction will be accurate. Linear SVM's are used during the learning process and are trained using the target and transformed source. In the study, experiments were performed for text document classification, two being cross-lingual and one being cross-view. The proposed method was compared with DAMA [27], ARC-t [28], HFA [23], standard SVM trained using only target data using one-vs-one and one-vs-all strategy for multi-class classification, and SHFR using one-vs-one as to test the ECOC scheme. Principal Component Analysis [35] (PCA) was used to reduce the number of dimensions as the baseline methods cannot handle such high-dimensional features. The proposed method SHFR with ECOC performed best for all tests and demonstrated to be an effective method for multi-class HDA.

TTI

Qi et al. [36] proposed an HTL method for image classification. This method uses a Translator for Text to Images (TTI). TTI uses labeled source text data, text-image co-occurrence data, and limited labeled image target data. Image classification currently has two challenges: (1) labeled image data is relatively scarce and expensive to collect and (2) features of image data lack semantic meaning for class prediction as they represent visual features rather than conceptual ones. On the other hand, labeled text data is often more available than labeled image data and text features have more semantic meaning for predicting a class label. With this observation, this method proposes using transfer learning to exploit such text data to improve image classification. The issue becomes how to relate the text to the images for semantic knowledge transfer. To close this gap, this method uses a text-image co-occurrence matrix which contains images along with the text that occurred with them on the same webpage. Co-occurrence information is effective in this case because of the assumption that the text around an image is describing the concepts in such image. This co-occurrence information is relatively inexpensive to collect and serves as a bridge to learn the correspondence for translating the semantic information between the features of the text source domain to that of the image target domain. This translation is done through a form of a feature transformation called a "semantic translator function." This translator takes into account the source, target, and co-occurrence data and learns the correspondence between the text from the source and the images of the target through the co-occurrence bridge. Each translator for the source text contains a "topic space" which is a common subspace to associate the data for translation.

This translator function is optimized through proximal gradient based optimization to find effective transformation matrices and complies with the parsimony principle. This principle states that the least complex, effective model is preferred as this avoids overfitting issues when faced with low training samples. Figure 3, adapted from [36], illustrates the semantic label propagation from the source texts to the test image. As shown

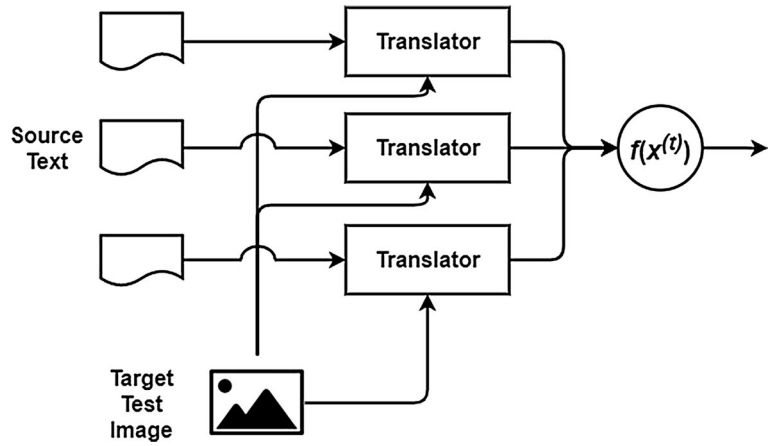

Fig. 3 Illustration for TTI adapted from [36] showing the process of semantic label propagation from source texts to test image using the learned translator. Output is a final prediction function

in the figure, each of the translators are aggregated to create a final prediction function $f(x^{(t)})$. Experiments were conducted for image classification and the proposed method is compared against HTLIC [22], TLRisk [37], and a baseline classifier trained only on the images without any source knowledge. The proposed TTI had the best performance in most cases, even when the number of labeled target samples were reduced.

TLRisk

Dai et al. [37] proposed a method for heterogeneous transfer learning called Translated Learning via Risk Minimization (TLRisk). This method proposes using a translator to perform a form of an asymmetric mapping by "translating" the features from a source feature space to that of a target feature space as to learn within a single feature space. It uses the language model of [38] and combines the use of such feature translation with nearest neighbor learning. This proposed method is modeled by a Markovian chain. The source feature space can be modeled as an initial Markov chain $c \rightarrow y_s \rightarrow x_s$ where the source data x_s is represented by features y_s. Also, the target can be similarly modeled as $c \rightarrow y_t \rightarrow x_t$. Thus, the proposed method can be modeled by a new Markovian chain of $c \rightarrow y_s \rightarrow y_t \rightarrow x_t$ which illustrates the transfer of knowledge from a source feature space to that of the target using $y_s \rightarrow y_t$ as a feature-level translation. This translation is done by learning a probabilistic model which uses co-occurrence data as a bridge between the source and target feature spaces. Furthermore, the proposed model performs translated learning through an extension of the risk minimization framework of [38]. The objective of this is to minimize an expected risk function with respect to the labeled training data and the feature translator. This risk function is formulated as an expected loss of classifying a test instance as a particular class thus we want to minimize the risk function as to have a higher probability of correct classification. Due to computational costs of these operations, the risk function is approximated and the algorithm is implemented using dynamic programming.

This method has similarities to TTI [36] as it uses a feature translator to solve the differing feature spaces and also uses co-occurrence data as the link. However, they differ

in properties such that TLRisk uses a Markovian chain model and risk minimization and TTI uses the "topic-space" as a common subspace in the translator. Experiments were performed on text-aided image classification and cross-language classification to test the effectiveness of the proposed TLRisk method. This method was evaluated using three different dissimilarity functions when approximating the loss function. The results show that TLRisk outperformed the baselines no matter which dissimilarity function was used demonstrating it as an effective method.

ARC-t

Kulis et al. [28] proposed a flexible framework for supervised HDA tasks for image domains called Asymmetric Regularized Cross-domain transfer (ARC-t). This method is based off using asymmetric, non-linear transformations learned in kernel space to map the target visual domain to the source. This asymmetric transformation is based on metric learning and this framework is an extension from [39] as it was made to function for domains with different dimensions through changes of the regularizer. Similar to DAMA [27], this method uses label information to construct the similarity and dissimilarity constraints between instances in the source and those projected from the target domain. It uses labeled data from the source and limited labeled data from the target. This method can be applied to situations where the target category differs from the training categories through the use of the correspondence constraints. The objective function for the transformation W contains the regularizer, in this case it is the squared Frobenius norm, but can extend to others such as trace-norm, and cost terms. These are minimized through an alternating optimization procedure. The transformation is learned in a non-linear Gaussian radial basis function kernel.

After the transformation, the target is mapped to the source and one can apply many classification methods such as SVM or k-Nearest Neighbors (k-NN). In the case of this work, k-NN was used. This method, in essence, also provides support for multi-class scenarios. Experiments were conducted to test this approach for object recognition with benchmark datasets under two scenarios: (1) source and target contains labeled training instances for all categories and (2) source contains labeled training instances for all categories but the target only contains half of the categories and shall be tested with data containing all categories. The proposed kernelized method was compared with metric learning [40], ITML [39], SVM, feature augmentation [24], and variations of k-NN. Since these methods do not support heterogeneous feature spaces with different dimensions, Kernel Canonical Correlation Analysis (KCCA) was used as pre-processing only on the baselines to project the heterogeneous feature spaces into a common subspace. The results show the proposed ARC-t had best performance for most cases over the baselines especially when faced with the challenge of adapting to new categories, features, and codebooks. It can be also noted that the kernelized version of ARC-t had significant improvements over the linear variant thus demonstrating the effectiveness of the kernelization proposed.

SHDA-RF

Sukhija et al. [41] proposed an HDA solution called Supervised Heterogeneous Domain Adaptation via Random Forest (SHDA-RF). This method works for any standard HDA

tasks but was motivated by its application for activity recognition in smart homes. In short, these smart home tasks learn the daily activities of the home owner through sensors but, when building a model for a new target home, there are limited labels available in this new home. Because of this, the intention is to transfer knowledge from a source home to a target home. Due to the varied positioning of the sensors and home layouts, it creates heterogeneous feature spaces but the overall activities are the same. Therefore, this algorithm assumes that the source and target have different features that categorize data partitions with similar label distributions. With this assumption, the shared label distributions can act as a pivot to learn a mapping between the feature spaces of the source and target. The generated sparse mapping in this process represents target features as linear combinations of source features. The goal of this is to create a mapping of the source domain features to the target domain features with the shared labels acting as the bridge.

To find these pivots, the method proposes using the leaf nodes of random forest [42] models created for the source and target. Each the target and the source domains will have their own random forest model built for them which estimates the pivots along with relationship matrices for each. These measure the contribution of the source and target domain's specific features to the shared label distribution. The contribution matrices can be found through the structure of the random forest. The leaf nodes of the decision trees hold a label distribution which are associated with a particular data partition. The path of each decision tree in the random forest from the root to this leaf contains a sequence of features chosen as split functions. Thus, using the features from this path we now have the label distribution and the domain specific features that are associated with it. For the case of duplicate label distributions in leaf nodes, the algorithm takes an average of the feature contribution vectors. The sparse mapping to map the source to the target is derived by running the least absolute shrinkage and selection operator method from the relationship matrices and minimizes the difference between them during optimization. The target random forest model is then re-trained using the mapped source data and the target data as before it was trained only on the target data. Random forest is particularly useful for this method since it reduces overfitting and complexity as only a single model is needed for each domain to find the relationship between the features of the source and target to the shared pivot. Experiments were performed for home activity recognition and text categorization. The results showed that the proposed SHDA-RF is effective by having the lowest error rates in most of the tests.

Methods which require limited target labels and accept unlabeled target instances

In this section, we survey various semi-supervised techniques which require labeled source data and limited labeled target data but they also accept unlabeled target instances during training. These techniques aim to enhance performance by exploiting knowledge from the unlabeled instances by incorporating them into different aspects of the target model training process.

DAMA

Wang and Mahadevan [27] proposed an HDA solution called Domain Adaption using Manifold Alignment (DAMA). This method uses a manifold alignment based approach and is an extension from their previous framework [43]. Standard manifold alignment techniques require a small amount of cross-domain correspondence relationships to learn mapping functions, but these correspondences are often difficult to obtain and sometimes require manual translation which can be highly expensive. DAMA makes use of the label information rather than correspondence to align the input domains. This HDA approach learns mapping functions to project multiple source domains and the target into a latent common subspace. When creating the subspace, each input domain is treated as a manifold and the mapping function is constructed for each input domain while preserving the topology of each one. Instances with the same label are matched and forced to be neighbors while those with different labels are separated. To do this, each manifold is represented by a Laplacian matrix which models the similarity of instances with others who share the same label. These are then joined into a larger matrix representation which serves as a joint manifold for the union of all input domains. Since this joint manifold has features from all the input domains, a dimensionality reduction process is performed to remove redundant features and is solved through a generalized eigenvalue decomposition. Once this common subspace is derived, one can combine this approach with other existing DA approaches as to support multiple source domains.

The DA approach used in this work to apply on this subspace is built in two stages. First, a linear regression model is trained using the labeled source data in the latent subspace. Then manifold regularization [44] is applied with a second linear regression model. This is then aggregated with the first to minimize prediction error for the labeled instances in the target. In other words, the first regression model uses the data from the source while the second adapts the first model to the target domain. During manifold regularization, unlabeled target instances are used to reduce overfitting issues caused by having very limited labeled target data. Experiments were performed on text document categorization and ranking tasks. The proposed method was compared with Canonical Correlation Analysis (CCA) [45], correspondence-based manifold alignment [43], and manifold regularization [44] (target only). The results showed the proposed method performed the best followed by manifold regularization, correspondence-based manifold alignment, and CCA being the worst performing in this case.

CDLS

Tsai et al. [46] proposed a semi-supervised HDA solution called Cross-Domain Landmark Selection (CDLS). This method derives a domain-invariant common subspace and learns representative cross-domain landmarks for adaptation purposes. To initialize the process, first all the target domain data is projected into an m-dimensional subspace. Here, $m \leq min(d_s, d_t)$ as to prevent overfitting from mapping a low dimensional space into a higher one. Then a feature transformation A is learned to project the source data onto the m-dimensional space creating a common subspace for the source and target data. A linear SVM is trained on the labeled cross-domain data to predict pseudo-labels for the unlabeled target instances. Once this is initialized the optimization process begins which updates the transformation A and landmark weights $\{\alpha, \beta\}$. Then, with these new

landmark weights, an updated SVM is trained to predict new pseudo-labels for the unlabeled target instances. This process repeats until convergence is reached. From this, final predicted labels are the output thus exploiting unlabeled target data during training.

These landmark weights are assigned to instances which are most representative of the labeled target instances and thus are more suited to be used for adaption. The higher the α, the more similar a labeled source instance is to a labeled target instance of the same class. The higher the β, the more similar an unlabeled target instance is. Those with lower weights have a higher chance to be misclassified. From this observation, one should use the instances with the higher weights for adaptation purposes as they are most representative of the target domain thus filtering out noisy examples. A parameter $\delta \in [0, 1]$ is then introduced to set the cut-off point for the portion of source labeled and target unlabeled instances to be used. If $\delta = 0$, then no cross-domain data will be used and the classifier will be trained only on the available target data. Note that when $\delta = 0$, this a supervised variant of CDLS used in the experiment called CDLS_*sup*. If $\delta = 1$, then all the cross-domain data will be used including those with low weights. Both of these are non-optimal as the first exploits no cross-domain data and the latter uses too much which introduces noisy examples, thus $\delta = 0.5$ was used in the study. Experiments were performed for object recognition and cross-lingual text categorization. The proposed method was compared with SHFA [47], DAMA [27], MMDT [48], CDLS_*sup*, and an SVM trained only on the labeled target data. The proposed method performed significantly better for all tests.

IDL for HDA

Han and Wu [49] proposed an Incremental Discriminant Learning (IDL) method for online HDA tasks. Online tasks, as described previously, are applicable when training samples are acquired sequentially. Standard models need to be completely rebuilt if new training data becomes available but to save time and space resources, this method incrementally optimizes its projection matrices as to account for such new data. This symmetric feature transformation method computes projection matrices to map the data from both domains into a common subspace. In this subspace, class labels are exploited to build a discriminative subspace. This is done through the transformations during which the variance of the samples of different classes are maximized and the variance of samples within the same class are minimized. In other words, the distance between instances of the same class is reduced and instances of different classes are separated to create such discriminative subspace. From this subspace, a standard SVM can be applied for classification using the projected data from both domains. When a new instance arrives, it is projected onto this discriminative subspace where it can then be classified based on where it landed in relation to the other instances in the subspace.

To allow for incremental training data, the existing projection matrices are updated with an eigenspace merging algorithm from Hall [50]. In doing so, this method updates the principal components of the total scatter matrix and the between-class scatter matrices as to the compute the projection matrices from them both. When a new training instance arrives, first the total scatter matrix is updated, then the between class scatter matrix, and finally the projection matrices are updated. Thus, one only needs to store the principal components of these matrices rather than all of the original training

data. When updating the scatter matrices, a sufficient spanning set concept is used to reduce the dimension of the eigenvalue problem. To utilize unlabeled target instances to enhance performance, this method introduces a criterion in the objective function as to use these instances to reduce the data distribution difference between the source and target. Experiments were performed for cross-view action recognition, object recognition, and multilingual text categorization which analyzed classification accuracy and computation time. The proposed method was compared with KCCA [25], HeMap [26], DAMA [27], ARC-t [28], and HFA [23]. The results showed that the proposed method had promising results by having good prediction accuracy and fast computational speeds in most cases. In the other cases, the results showed it was outperformed and/or slower depending on the dataset/test performed.

MOMAP

Harel and Mannor [51] proposed a method called Multiple Outlook MAPing (MOMAP). The motivation is to learn for a single task having data from multiple sources which may be represented by heterogeneous feature spaces. In their paper, each feature space is defined as an "outlook" by the authors. Specifically for HTL tasks, each feature space from the source or target is defined as an outlook and the goal is to map each one of the source outlooks to that of the target outlook through asymmetric transformation. This process is done by first applying a scaling to each of the outlooks as to normalize the features to the same range of [0,1]. During such normalization, the process of Winsorization [52] is applied to remove sensitivity when scaling outliers which is done by collapsing the extreme two percentile of the data to the high ends of the rest of the data. To map the two outlooks, a process of rotation and translation is performed to match the source to the target. This is done by first grouping the classes together from each outlook, translating the means of the features for each class group to zero, and then matching the source groups with the corresponding target groups. From this, a transformation matrix for each group can be created.

To build the matrices, each outlook has a utilization matrix constructed for it and then singular value decomposition (SVD) is performed, as to align the marginal distributions, using the utilization matrices. This is done during the matching by rotation process which is performed to derive the transformation matrices for the class groupings. When presented with outlooks of different dimensions, the smaller utilization matrices are padded with zeros to equalize the dimensions overall. After deriving the transformations, one can apply a standard classification algorithm using the target outlook and the transformed source outlooks. Experiments were performed using homogeneous and heterogeneous environments for activity recognition using data from wearable sensors. These experiments are multiclass in nature, thus a multi-class SVM was used as the classifier. Balanced error rate was used as the performance metric due to the uneven class distribution of the dataset. For the heterogeneous experiments, the proposed MOMAP algorithm was compared with an SVM trained only on the limited target data and an SVM trained on a fully-labeled target dataset. The results showed that the proposed MOMAP algorithm was outperformed by the SVM trained on a fully-labeled target as one would expect. On the other hand, the proposed method outperformed the SVM

trained only on the limited target data in most cases by having a lower balanced error rate. This, in effect, demonstrated the effectiveness of the proposed method.

HeMap

Shi et al. [26] proposed an HTL method called Heterogeneous spectral Mapping (HeMap). This method may be used when the source and target have different input feature spaces/dimensions, different data distributions, and/or different output label spaces. A transfer learning task containing all of these can be very challenging. First, to address the issue of different input feature spaces, the method proposes using a symmetric transformation process to project the source and target data onto a common subspace through the use of spectral mapping to unify the input spaces. To learn the projection matrices and create the common subspace this method does not utilize labels, as the source and target may be of different label spaces. Instead, it tries to discover correspondences between the source and target data points in the optimization. The spectral mapping technique uses linear transformations such as, rotation, scaling, etc. on the target matrix, and is modeled through an optimization objective which aims to maximize the similarity between the source and target while preserving their original structure. If they are too unrelated to each other, the projected data may still exhibit differences in distributions when preserving the data's original structure. In this case, the process is aborted when they are too unrelated and no source data will be utilized as the risk it may degrade performance is too high. To handle the case of different data distributions, this method extends the approach from the author's previous work [53] and utilizes a clustering based sample selection method on the latent subspace as to select source data that is most similar to the target data. This is then used as new training data, which in effect aims to resolve the marginal distribution differences and improve performance. To handle the case of different output label spaces, this method also extends work from [53] and uses a Bayesian-based method to model the relationship between the differing output spaces as to unify them through re-scaling and calibration of the two output variables.

This method requires having the same number of instances for both the source and target. To satisfy this, the paper proposes randomly duplicating the smaller input to match the larger, as to maintain original distribution, with the purpose of making the projection matrices of the same size. Classification and regression experiments, which include image classification and drug efficacy prediction, were performed to test the effectiveness of the proposed algorithm. The proposed HeMap had better performance, measured via error rate, compared to the single baseline tested though details of this baseline were not provided in the paper.

Proactive HTL

Moon and Carbonell [54] proposed a framework which aims to solve multi-class text HTL tasks which we denote as proactive HTL. This method may be used when faced with a source and target which exhibit both different feature and label spaces. This method uses a symmetric transformation technique to map the source and target onto a common subspace, as to resolve the differing feature space issue. Then to resolve the differing label spaces, this method simultaneously learns a shared projection to map

the data into a final embedded label space. Specifically, the proposed method utilizes a skip-gram based language model [55] which learns semantically meaningful vector representations of the words as to map the source and target labels into a word embedding space. From this space, we can obtain the source and target's semantic class relations. The obtained label term embeddings are used as "anchors" for the source and target as to derive instances belonging from semantically related categories. From this, a simultaneous optimization objective is performed where, first, two linear transformations are learned to map the source and target data into a common subspace. Note that these mappings can be learned with deep neural networks, kernel machines, etc. Simultaneous to this, the proposed method then learns a shared projection from the common subspace to map the joint features into a final embedded label space. By performing this we are resolving the issues of heterogeneous feature and label spaces. A 1-NN classifier is used to look for the category embedding which is closest to the projected instance in the final space.

These tasks are done under a proactive learning framework which exploits unlabeled target instances as to expand the original limited target training data and improve performance. Sampling of the unlabeled instances is done by iteratively selecting "bridge" instances in the target which utilize source knowledge and maximizes the expected utility function of the target model. The sampling is done using two objectives: (1) maximal marginal distribution overlap (MD) which selects unlabeled target instances where the marginal distributions of the source and target have the highest overlap, and therefore are more semantically related; and (2) maximum projection entropy (PE) which selects unlabeled target instances which maximizes entropy of the dot product similarities between a projected instance and its possible label embeddings. Experiments for heterolingual text classification were conducted where the source and target datasets contain both heterogeneous feature and label spaces. The results demonstrated the proposed method is effective compared to the baselines as it lowered error rates.

SHFA

Li et al. [47] directly extended the HFA [23] algorithm and proposed Semi-supervised Heterogeneous Feature Augmentation (SHFA). This method exploits the knowledge from unlabeled target instances to enhance a target HTL task with limited target labels. The method utilizes ρ-SVM with squared hinge loss trained on the limited target as to infer pseudo-labels for the unlabeled instances. The task of finding the optimal labels is computationally expensive and is denoted as a Mixed Integer Programming problem. Because of this, an optimal linear combination of the feasible labeling candidates is found for the instances leading to a less expensive optimization problem. These pseudo-labels are used during the building of the final classifier and are estimated when learning the optimal nonlinear transformation metric H. As described in Duan [23] and "HFA", H is a combination of the projections P and Q used to map the source and target into a common subspace and by using H one does not need to solve for P and Q directly but rather only optimize the H transformation matrix. For SHFA, H is decomposed into a linear combination of a set of rank-one positive semi-definite matrices. Optimization of such is then solved using Multiple Kernel Learning as defined in [56] with the l_1-norm constraint. Experiments were performed for object recognition, text categorization, and

sentiment classification. The proposed SHFA was compared with the original HFA [23], HeMap [26], DAMA [27], ARC-t [28], and an SVM trained only on the labeled target instances. The proposed SHFA performed significantly better than the other baselines for these experiments by having better classification accuracy, including better performance over standard HFA.

CT-Learn

Ng et al. [57, 58] proposed a Co-Transfer Learning (CT-Learn) framework which aims to transfer knowledge between two or more heterogeneous domains. In such case one can have a target domain and several source domains, each represented by different feature spaces. The proposed method uses a co-transfer strategy which builds relations to link the feature spaces, through co-occurrence data, as to co-transfer knowledge amongst them simultaneously. First a joint transition probability graph P, shown in Eq. (2) which was adapted from [57], is constructed using intra-relationships and inter-relationships for all the co-occurrence, labeled, and unlabeled instances across both domains. The intra-relationships are calculated through the affinity of the intrinsic manifold structure between instances of the *i* th domain. In (2), the diagonal block matrix $P^{(i,i)}$, which is an n_i by n_i matrix, indicates these intra-relationships. The off-diagonal block matrix $P^{(i,j)}$, which is an n_i-by-n_j matrix, indicates the inter-relationships between the *i* th and *j* th domains and is calculated with the co-occurrence data. The weighting parameter $\lambda_{i,j}$ controls the amount of knowledge to be transferred from the *j* th instance space to the *i* th during the learning process. For a standard binary domain transfer learning task $\lambda_{target,source} \neq 0$ and $\lambda_{source,target} = 0$ so that knowledge is transferred from the source to the target and not the target to the source.

$$P = \begin{pmatrix} \lambda_{1,1}P^{(1,1)} & \lambda_{1,2}P^{(1,2)} & \cdots & \lambda_{1,N}P^{(1,N)} \\ \lambda_{2,1}P^{(2,1)} & \lambda_{2,2}P^{(2,2)} & \cdots & \lambda_{2,N}P^{(2,N)} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{N,1}P^{(N,1)} & \lambda_{N,2}P^{(N,2)} & \cdots & \lambda_{N,N}P^{(N,N)} \end{pmatrix} \tag{2}$$

The learning process is modeled as a coupled Markov-chain based random walk with restart [59] where each instance is regarded as a node. The proposed method uses this random walk during the learning process on the joint transition probability graph to propagate the ranking score of labeled instances as to calculate the importance of a set of labels to an unlabeled instance. In other words, the proposed method calculates ranking scores of labels and generates the possible labels for a test instance through propagation of these ranking scores. The use of a Markov chain differs from TLRisk [37] as for CT-Learn the Markovian principles are used for the learning process while in TLRisk it is used to estimate parameters. Binary and multi-class experiments were performed comparing the proposed CT-Learn against TTI [36], HTLIC [22], and an SVM model trained only on the target data. The results showed the proposed method is statistically superior, as it increased accuracy over the baselines when applied to the experiment's cross-language and text-image classification tasks.

SSKMDA

Xiao and Guo [60] proposed a semi-supervised method called Semi-Supervised Kernel Matching Domain Adaptation (SSKMDA) for homogeneous and heterogeneous domain adaptation. This method utilizes kernel matching techniques to map target instances to the source rather than using a feature-based feature transformation. This method also utilizes labeled and unlabeled data from the source and target domains. To address the issue of heterogeneous features, the proposed method creates a kernel matrix for the target and source data respectively and simultaneously learns a prediction function while mapping instances of the target to similar source instances by matching portions of these matrices. The motivation for mapping instances rather than features is because when performing text classification tasks, the application in the study, the feature spaces are often of much higher dimension than the number of instances. Thus, projecting features requires much more computational resources. The learning objective of the proposed method consists of a combination of three parts: the kernel matching criterion, prediction losses, and graph Laplacian regularization. From the combination of these, the goal is to learn the kernel mapping of the instances and the kernelized prediction model as to minimize the regularized training losses in both domains. The kernel matching criterion is used to map the individual target instances to source instances based on their geometric similarities which are expressed in the kernel matrices. This is done using the Hilbert–Schmidt Independence Criterion (HSIC) [61].

Through the use of HSIC, the target kernel matrix is mapped to a submatrix of the source kernel matrix while using available labeled target instances as pivots for class separation. These pivots instances are mapped to source instances of the same class. The unlabeled instances are also mapped to their expected class in the source instance set using the pivots and kernel affinity measures. Prediction losses are incorporated into the learning process as to minimize training loss on both the labeled source and mapped labeled target instances for a prediction model trained on the source data. This model was generated in a supervised manner. To exploit unlabeled instances within the proposed framework, graph Laplacian regularization terms are incorporated to utilize information about the unlabeled data's geometric structures of the marginal distributions for each domain. For this method to work, a manifold assumption is made that, when mapping the instances, if two points which have similar intrinsic geometry then it implies their conditional distributions are similar. Otherwise if they are not similar, the matching will be noisy and/or irrelevant. Experiments were performed for cross-language text classification comparing the proposed SSKMDA to HeMap [26], DAMA [27], ARC-t [28], and HFA [23]. The results demonstrated the proposed method to be more effective than these baselines for addressing this task.

SCP-ECOC

Xiao and Guo [62] also proposed a semi-supervised HTL method for multi-class DA called Subspace Co-Projection with ECOC (SCP-ECOC). This method aims to solve the issue of disjoint feature spaces by learning a set of symmetric transformation matrices to project the source and target data into a common subspace. While projecting the instances, the proposed method simultaneously learns cross-domain prediction models from the projected labeled instances in the co-located latent subspace. While this

subspace allows for cross-domain knowledge transfer, it may have poor discriminative information as it was generated with little labeled target instances. Therefore, this method proposes exploiting unlabeled instances by incorporating them into the co-projection process and minimizing the means of the projected instances between source and target domains. This is due to the assumption that the two domains have similar empirical marginal distributions in the subspace which is enforced with a maximum mean discrepancy criterion [63]. This, in effect, reduces the marginal distribution differences between the domains in the subspace.

To handle the challenge of multi-class tasks, the proposed method utilizes Error-Correcting Output Code (ECOC) [34] schemes, specifically exhaustive ECOC. This is used rather than the commonly used one-vs-all (OVA) scheme. By using ECOC the multi-class task can be casted into a larger number of cross-domain binary classification problems than in the OVA scheme. This thus creates a more robust model because adding more of these binary classifiers which increases the stability and informative ability of the subspace co-projection when faced with a multi-class scenario. Combining all these methods, the learning process becomes a joint minimization process and is solved with an alternating optimization procedure. Thus, by combing these methods into one framework, this framework ensures an informative common subspace with high discriminative ability and minimal marginal distribution difference for knowledge transfer across multi-class domains. Experiments were performed on cross-lingual text classification and cross-domain digit image recognition tasks. The proposed SCP-ECOC was compared with HeMap [26], DAMA [27], MMDT [48], SHFA [47], SCP with OVA, and a baseline trained only on the target instances. The results proved the proposed method effective by having the highest classification accuracy for all tests.

MMDT

Hoffman et al. [48] proposed a heterogeneous domain adaptation method for multi-class image classification called Max-Margin Domain Transforms (MMDT). This method allows for multiple classes, requires limited labeled target training instances, and utilizes unlabeled target instances though one can have classes which contain only unlabeled instances. This is possible as it adapts all the points through a linear asymmetric feature transformation to map the target domain to that of the source creating a feature transformation shared across all classes. To establish multi-class support, the proposed method adapts a max-margin classifier by learning a shared component of the domain shift though the feature transformation. MMDT simultaneously learns the projection matrices along with the classifier parameters and performs optimization on both through a classification loss based cost function. In other words, in this framework both the classification objective and the feature transformation are updated together based on the prediction results of the previous training instance in relation to its truth label. By performing such optimization through classification loss, accuracy can be increased. This method also provides for less computational cost as it reduces training time by using hyperplane, rather than similarity, constraints. By doing so, one can optimize in linear space and avoid computational complexities of kernel techniques, thus making the proposed method suitable for scaling to larger quantities of training data.

The main idea behind this method is to jointly learn affine hyperplanes which separate the classes consisting of source and target instances. This is done while learning the feature transformation to map the points of the target to the source such that it projects the points onto the correct side of each source hyperplane. In comparison, ARC-t [28] uses similarity constraints to map points of the same category close to each other followed by a classification step but, on the other hand, MMDT proposes to project the target points to the correct side of the already learned hyperplane as for better classification accuracy. Experiments were performed to test the performance of the proposed method for image classification tasks. Four tests were conducted which include multi-class, homogeneous domain adaption, heterogeneous domain adaption, multi-task DA, as well as scaling the test to larger datasets for computational performance analysis. For the heterogeneous tests, the proposed MMDT was compared with ARC-t [28], HFA [23], and SVM trained only on the target data. The results showed that the proposed method was effective and it had the highest classification accuracy in most cases, while also having the fastest computation time for the larger dataset test while maintaining highest accuracy.

SMVCCAE, SSMVCCAE

Samat et al. [64] recently proposed two new HTL methods called Supervised Multi-View CCA Ensemble (SMVCCAE) and Semi-supervised SMVCCAE (SSMVCCAE) for remote sensing image classification and pattern recognition tasks. These methods use a multi-view CCA ensemble approach. Specifically, the supervised variant (SMVCCAE) first takes the target training data and splits it into N views based on multi-view learning, thus creating multiple views which may be disjoint or partially disjoint feature subsets. Each of these views provide unique but complementary information which can be individually compared with the source domain to provide for more enhanced knowledge transfer. From this, each view is projected with the entire source data onto a correlation subspace. Therefore, this will result in N common subspaces each of which compare a target view with the source data. The transformation matrices used for such projections are obtained through Canonical Correlation Analysis (CCA) as to project the data symmetrically onto these correlation subspaces. Then a base classifier, in this case Random Forest [42], is trained on the transformed data on each of these subspaces, thus creating N classifiers. These N classifiers are combined to create an ensemble learner where a class for an instance of the target testing data is predicted by each of the classifiers in the ensemble, thus creating N predictions. To combine these predictions into a final output, one may use voting schemes such as majority vote where the class value that was predicted the most becomes the final output. This may be suitable for some cases but this method proposes a more enhanced voting method which uses a weighted vote scheme based on the correlation coefficients from the subspaces.

As discussed in "CT-SVM", the correlation coefficients ρ from the subspace can be used to indicate the similarity of the source to the target along such dimension. In this case, the higher the ρ value, the greater correlation the source has to a particular target view meaning there is greater transfer ability for that subspace. From this, we see that a higher weight should be placed on the prediction results from a model built on a subspace that has a higher correlation coefficient ρ value rather than one with poor correlation. This thus becomes the proposed weighted voting strategy used which is based

on these correlation coefficients. This method is the supervised version, SMVCCAE, which requires limited target labels but this method is extended to accept unlabeled target instances in the second proposed method, SSMVCCAE. This method utilizes unlabeled instances to reduce the data distribution gap by incorporating multiple speed-up spectral regression kernel discriminant analysis (SRKDA) into the original supervised method. These unlabeled instances are used during the process of building the affinity matrix which is utilized in the KDA problem. Experiments were performed for hyperspectral image classification which demonstrated that the proposed methods were effective as they achieved higher classification accuracy and computational efficiency compared to the baselines.

TNT

Chen et al. [65] proposed a neural network based framework called Transfer Neural Trees (TNT) for semi-supervised HDA tasks. This framework is divided into two layers: mapping and prediction. The first layer is that of mapping the source and target into a domain-invariant representation while the second layer performs adaptation and classification. This process of mapping, adaptation and classification are all solved in a joint manner. Mapping of the source domain data is performed separately to that of the target domain data while each using a single-layer neural network. For the case of this study, these neural networks apply the hyperbolic tangent as the activation function with an output dimension of 100. To account for the unlabeled target instances in semi-supervised tasks, an embedding loss term is incorporated into the target domain feature mapping. When performing adaptation and classification, minimizing this loss term can increase predictive consistency for the outputs of the individual trees and the forest as a whole. This process also preserves the structural consistency between the labeled and unlabeled target instances.

For adaptation and classification, Chen et al. [65] also proposed Transfer Neural Decision Forest (Transfer-NDF) for use in the TNT framework. Inspired from deep neural decision forest [66] and random forest [42], Transfer-NDF uses neural networks as decision trees as to build a forest of neural decision trees. To build this forest, first the source domain data is observed to build the individual trees. Both the source mapping and the forest are updated via backpropagation. Then, once the trees are obtained, the target domain data is observed to perform distribution adaptation. Also, during the learning of Transfer-NDF, a process of stochastic pruning is applied which adapts representative neurons for better generalization from the learned source domain to that of the target. In the case of this study, Transfer-NDF consists of 20 trees with a depth of 7. Similar to the process of Random Forest, each tree samples 20 dimensions from the final mapping output for a diverse representation of the data.

Experiments were performed for cross-domain object recognition and text-to-image classification. The proposed TNT method was compared with MMDT [48], HFA [23], SHFR [31], SCP-ECOC [62], dNDF [66], SVM trained only on the labeled target data, and a two-layer neural network trained only on the labeled target data. The proposed method outperformed these baselines for almost all of the experiments thus demonstrating its effectiveness for these tasks.

HDANA

Wang et al. [67] recently proposed a semi-supervised HDA method called Heterogeneous Domain Adaptation Network based on Autoencoder (HDANA). Through the use of a Deep Learning [68] approach, this method performs symmetric transformation to project the source and target data onto a shared feature space. This process is realized by using two autoencoder networks for the source and target domain data respectively. The authors claim that, by performing such multi-layer non-linear mapping methodology, one can obtain a more abstractive shared feature space which better represents the probability distributions of the data compared to shallow transfer learning methods. Each autoencoder network in this method consists of $n + 1$ layers, n of which are feature layers and the final layer is the classification layer.

At the last feature and classification layer of each autoencoder network, the marginal and conditional distributions are matched by introducing a Maximum Mean Discrepancy (MMD) metric. This is to reduce the cross-domain data distribution differences exhibited in the shared subspace. Along with the aforementioned metric, a manifold alignment term based on label information is also introduced into the objective function. The addition of such term is to preserve the geometric structure and label consistency of the data. This manifold alignment term contains three parts: geometric, similarity, and dissimilarity terms. The geometric term based on graph ensures that the manifold structure of the data within each domain remains unchanged though it does not account for label information. Thus, the similarity term is maximized during optimization for intra-class examples and the dissimilarity term is maximized for inter-class examples. This ensures that instances of the same class have a similar shared feature representation while those from different classes share a different one. These terms allow for a more discriminatory shared feature representation. Finally, to improve classification performance, a loss term for the softmax classifier used in the classification layer is incorporated into the objective function. The output obtained from this loss term is used during optimization.

To optimize the terms of the autoencoders and the objective function, the gradient descent method with backpropagation is used to learn updated parameters. For each iteration, new pseudo-labels are calculated for the unlabeled target instances and the process of mapping, classifying, and updating continues until convergence is reached. The output obtained from the final softmax classifier becomes the final predicted labels for the unlabeled target domain data.

Experiments were performed for cross-domain object recognition and cross-lingual text categorization. The proposed HDANA was compared with DAMA [27], MMDT [48], SHFA [47], G-JDA [69], TNT [65], SVM trained only on the labeled target data, and an Autoencoder trained only on the labeled target data. The results showed that the proposed method had superior performance over these benchmarks for all of the tests conducted.

Methods which require no target labels

In this section, we survey various techniques which require labeled source data but do not require any labeled target data. These techniques are useful when collecting even limited labeled target data is too expensive. These methods utilize the unlabeled target instances during the training process and generally either infer their labels, align

the source data (or source prediction function) to that of the target domain, incorporate them into the mapping process, or utilize them otherwise to represent the target domain.

CT-SVM

Yeh et al. [70] proposed a framework for cross-domain pattern recognition tasks which uses Kernel CCA for symmetric feature transformation and proposed a modified SVM named Correlation-Transfer SVM (we denote as CT-SVM). This method utilizes Canonical Correlation Analysis (CCA) [45] for deriving a domain-invariant joint feature space to associate data from the source and target. Here, CCA aims to maximize the correlation, through a coefficient ρ, between two variable sets and learns projection vectors to derive a subspace which can be used as a common feature representation to solve cross-domain problems. Reduced Kernel CCA is utilized when non-linear subspaces are desired. The common subspace dimension is bounded by the minimum feature dimension whether it from the source or target feature space as to prevent overfitting. Once this common subspace is derived, one can project unseen test data onto the subspace for classification by a model trained from the source view data, as in this case no labeled target instances are available. Rather than directly using this subspace, this method proposes to exploit the domain adaption ability that can be observed in the subspace.

Each dimension in the derived CCA is associated to its own ρ value. The higher the correlation coefficient ρ, the closer the source and target domains are to each other in such dimension thus it is more suited for domain transfer. An example of this can be seen in Fig. 4 which was inspired from [70]. As illustrated in Fig. 4a, in the dimension with the higher ρ value the source and target data are more similar and their points on the subspace are more clustered together. Therefore, a classifier projected onto the subspace will have an easier time discriminating between the classes for the unseen target test data. On the other hand, in Fig. 4b, in the dimension with the lower ρ value the source and target are dissimilar, as there is a higher difference in distribution, which will result in lower recognition performance. Standard CCA-based approaches for

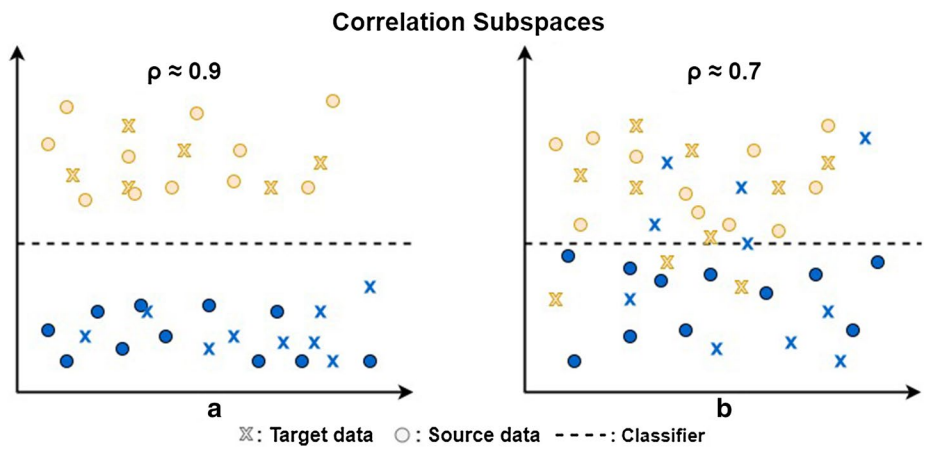

Fig. 4 Illustration for CT-SVM, inspired from [70], showing the difference between a ρ value with higher correlation (**a**) versus a ρ value with lower correlation (**b**) when projecting source and target instances onto correlation subspaces

cross-domain tasks do not take the domain transferability into account when learning classifiers in the correlation subspace which can lead to degraded performance when lower ρ values are present.

This method thus proposes using a modified SVM called Correlation-Transfer SVM which takes such transferability into account by enforcing a suppression of the learned SVM along such dimensions with lower ρ values as they have little or no beneficial knowledge for transfer to the target task. This, in effect, will provide more focus on the attributes with higher correlation between the source and target domains where beneficial discriminatory knowledge is more likely to be transferred from. This method does not require labeled target data and it utilizes the available unlabeled target and source pairs to select a kernel parameter which results in the highest correlation performance. Experiments were performed for cross-view action recognition, handwritten digit recognition, and image-to-text/text-to-image classification to analyze the effectiveness of the proposed algorithm. The proposed method was compared with combinations of linear and kernel SVM and CCA where the proposed method was shown to be effective. The results also indicate that deriving a common subspace better suited for domain adaption is more important than designing a complex classifier in a standard feature space.

HHTL

Zhou et al. [71] proposed a method for HTL called Hybrid Heterogeneous Transfer Learning (HHTL). This method creates an asymmetric mapping from the target to the source and takes into account bias issues of cross-domain correspondences. HHTL proposes using a Deep Learning [68] approach to perform cross-domain feature mapping and distribution bias reduction. It uses a labeled source, unlabeled target, and unlabeled correspondence data. When using correspondence data between source and target domains the method assumes that they are statistically representative (though this assumption may not always hold in real-world scenarios). An example of using correspondence data is when one has a cross-lingual, cross-domain review sentiment classification task in which we could use review-correspondences to learn a mapping between the two languages. The issue becomes that transforming the data from one language to another may not be effective due to distribution bias caused by the difference in product domains. Thus, this method aims to discover a latent feature representation to reduce such bias after transformation.

To do so, this method proposes applying a marginalized Stacked Denoised Autoencoder (mSDA) [72, 73] on the source domain data with its corresponding unlabeled data and the target with its corresponding data as to learn high-level feature representations. Then from these high-level representations, a feature mapping is learned to create a latent common representation as to bridge the gap between heterogeneous feature spaces. These two steps of mSDA and heterogeneous feature mapping are recursively applied to the source and target data in each layer as to generate the different levels and feature transformations for the K layers used in this deep learning approach. After such process, standard classification models can be used on the source data with the latent representation to build a target classifier. Experiments were performed for multi-lingual sentiment classification with biased and unbiased correspondence instances. The proposed HHTL method was compared with HeMap [26], multimodal deep learning [74]

with CCA, cross-lingual KCCA [75], and SVM-based method. This SVM-based method transfers labels predicted from the source correspondence data to the target as to train a target SVM on these pseudo-labels. Results showed the proposed method had the highest accuracy rates for the unbiased tests and for most cases in the biased tests.

HDCC

Wang et al. [76] proposed a framework called Heterogeneous Discriminative Analysis of Canonical Correlation (HDCC) to handle HDA for the application of video annotation, as well as other HTL, tasks. For video annotation tasks, it is often expensive and difficult to collect sufficient labeled videos to build an effective classifier. The goal, in this case, is to use transfer learning to transfer knowledge from a label abundant image source domain to enhance a target video annotation task which, in this case, has no labels. The issue becomes that source image data are represented by static image features while target video data are represented by spatial–temporal video features. Thus, this method adopts Canonical Correlation Analysis (CCA) to create a common subspace using projection matrices for the source and target by incorporating both the discriminative information exhibited in the source, and the topology information in the target. The topology information in the target is explored by taking each target video and splitting it into smaller clips where then a frame is selected randomly, thus creating a static image from the video as to relate to the static source images. The discriminative information is extracted from the source by maximizing the similarity (reducing the variance) of infra-class samples while minimizing the similarities (maximizing the variance) of inter-class samples to create a more discriminative space. From this subspace, we can then apply standard learning algorithms to create a target model.

In this work, DASVM [77] is used due to the distribution differences between the captured keyframe and the source image. This method initializes the model using the source labeled data and then they are gradually replaced by target instances to learn the final separation hyperplane. Wang et al. [76] also extends HDCC by introducing a joint group weighing methodology to learn from multiple heterogeneous sources. The idea behind this is to organize source images into groups in a joint manner based on their semantic meanings rather than origin. Then, weights are assigned to each group of input source data based on their relatedness to the target video. This in turn, allows for selection of more performance enhancing source data which contains greater semantic knowledge for transferring to the target task. The training with these groups can be split into two phases: (1) a classifier is learned for each input source group and (2) weights are assigned and optimization of such generate the target classifier. In the test phase, image frames are extracted from the videos and inputted into the classifier to predict a final label. Experiments were performed for image-aided video annotation which demonstrated the effectiveness of the proposed HDCC and group weighing methods.

CL-SCL

Prettenhofer and Stein [78] proposed a new HTL method called Cross-Language Structural Correspondence Learning (CL-SCL) for cross-language text classification that is built upon structural correspondence learning (SCL) [79]. These transfer learning tasks are inherently heterogeneous in nature as the feature space of a source document

is written in a different language as the target. They also have non-overlapping feature spaces due to each language having a distinct word set in their vocabulary. In this case, the goal of such task is to enhance the performance of a target text classification task which contains only unlabeled instances utilizing knowledge from a source domain with amble labeled and unlabeled documents. As mentioned previously, these documents do not share any common features/words as they are from different languages; however, one can link the two of them through the words' semantic meaning. Using this concept as well as SCL theory, this method proposes using word pairs $\{w_s, w_t\}$ as pivots which capture correspondences between the source and target languages as they are semantically equivalent. From these correspondences, we can learn a symmetric transformation mapping to discover a common latent feature space which reduces the task into a standard classification problem.

This CL-SCL method is comprised of three steps. The first step selects a small set of word pairs to be used as pivots by querying a translation oracle, such as Google Translate, with a source word to finds its corresponding translation in the target vocabulary. With this relation, w_s captures the correspondence with the source vocabulary and w_t with the target. Both words from these pivots must have good predictive value and occur frequently to be effective. The second step removes these pivot words from a copy of each feature space where then a linear classifier is trained for each pivot using the data from both domains. These pivot classifiers are applied to this copy as to predict whether a pivot word occurs in a document based on the other words in it. This allows us to model the relationship of a pivot word to other words in documents associated to each class. From these, we then compute correlations across pivots with singular value decomposition in step three as to then finally learn a transformation mapping to apply to each input space to discover the latent feature subspace. This method uses less resources than current approaches as it does not require extensive multi-lingual dictionaries or parallel corpus. Experiments performed for cross-language sentiment classification showed the proposed CL-SCL had competitive results.

HDP

Nam and Kim [80] proposed an HTL method called Heterogeneous Defect Prediction (HDP) for software defect prediction, though it can also be applied for other tasks. Most software defect prediction studies have been performed to train a model for a project and test it within the same project while using the same set of metrics/attributes. This is known as within-project defect prediction (WPDP). When a new project is developed, it contains little or no historical defect data. Therefore, one can use transfer learning techniques to apply data/models from other source projects and to detect defects in the new target project. This is otherwise known as cross-project defect prediction (CPDP). Most CPDP approaches require both source and target projects to be represented by the same set of metrics as for homogeneous transfer learning though this becomes a limitation as it may be challenging to find source projects which are characterized by the exact same set of metrics as one's new target project. To address such limitation, the proposed method can accept heterogeneous metric sets to transfer knowledge from a source to a target software defect detection task.

This approach uses two steps before creating a target classifier: metric selection and metric matching. First, metric selection (a.k.a. feature selection [81] denoted in other domains) is applied to the source data to remove redundant and irrelevant features. Techniques applied for this include gain ratio, Chi squared, relief-F, and significance attribute evaluation [82, 83]. During the experiments, the top 15% of metrics were selected as suggested by [82]. Second, metrics are matched between the source and target based on similarities of distribution or correlation. The manner this is done is by measuring the similarity between a source and target metric pair using one of three analyzers (PAnalyzer, KSAnalyzer, or SCoAnalyzer) to calculate a "matching score." From this, dissimilar pairs are eliminated based on a threshold value and the remaining are organized into groups of unique pairing combinations. Each of these groups has all the combinations of pairings without duplicate metrics. The best combination of metric matching is found by selecting the group with the highest sum of matching scores using weighted bipartite matching [84], thus leaving one remaining group which matches the metrics that are most similar. These scoring techniques include: (1) The PAnalyzer which uses percentile based matching, (2) The KSAnalyzer which uses the p value from the Kolmogorov–Smirnov Test, and (3) The SCoAnalyzer which uses the Spearman's rank correlation coefficient to measure sample correlation. After such procedure, a standard classification algorithm may then be trained on the transformed source to detect defects for target projects. Experiments performed for cross-project defect prediction showed the proposed HDP method effective in most cases, as described by the authors. These results may be misleading though due to certain results not being included in calculations as described by Weiss et al. [7].

FuzzyTL

Shell and Coupland [85] proposed an HTL method called Fuzzy Transfer Learning (FuzzyTL) for learning in intelligent environments (IE). Intelligent environments consist of a wide-range of applications where sensors of diverse kinds are placed to gather information about an environment. IE's are often dynamic and transient in nature due to fluid environments and changing conditions which causes uncertain datasets. Each IE is unique due its specific implementation which includes characteristics such as variations in the environments themselves as well as sensor placement, construction, type, and quantity. Thus, when trying to learn a model for a new IE it is challenging to collect sufficient labeled training data to learn an effective model for such IE. Transfer learning can be used in this case to transfer knowledge from labeled data of an existing source IE to enhance building a model for a target IE.

The proposed FuzzyTL method utilizes fuzzy logic [86] as to transfer knowledge from contextually varying environments to model a target task in an IE using labeled source data. Using fuzzy logic as a base, this framework can absorb the inherent uncertainty and dynamic nature of IE's by incorporating approximation and greater expressiveness of such uncertainty exhibited within the data. This proposed method consists of two main parts: learning and adaptation. During the learning phase, the Fuzzy Interface System (FIS) is constructed using the source labeled data, which is the basis for capturing knowledge from the source, and transferring it to the target model. The structure of the FIS consists of fuzzy sets and fuzzy rules which are formulated using an Ad-Hoc

Data Driven Learning (ADDL) process and the algorithm by [87, 88]. During the Adaptation phase, the generated source FIS and the target unlabeled data are used to adapt the fuzzy rule set and the fuzzy rule base from the FIS as to capture variations in the data and assist in bridging the contextual gap between the source and target. Figure 5, adapted from [85], provides an illustration of the FuzzyTL process to transfer knowledge from the source to the target and predict a value in the target. Experiments were performed using data from a real IE sensor network as for comparing FuzzyTL to ADDL. The results showed that the proposed method outperformed this baseline as it was able to adapt to changes in temporal and situational contexts inherent in IE's.

FSR (IFSR, UFSR, ELFSR)

Feuz and Cook [12] developed three variants of their proposed Feature-Space Remapping (FSR) method for HTL tasks when one has either limited target labels or optionally no target labels as well as an ensemble technique. This work extends from the previous version of their paper [89]. These methods do not assume the source and target domains contain the same feature spaces, probability distributions, dimensions, or label spaces.

The proposed method, called Feature-Space Remapping, handles these issues by remapping the target data from their original feature spaces to that of the source feature space through an asymmetric transformation. This mapping is found through the use of metafeatures which are used to find the similarity between a feature pair. More specifically, first metafeatures are created, based on the availability of target labels, from the source and target data. Then, from these metafeatures, a similarity matrix is constructed by computing a similarity score between the source and target metafeatures. To construct the mapping, feature pairs with the highest similarity scores are selected as to decrease the error when applying target domain data to a previously learned source domain model. This mapping is many to one as many target features can map to one source feature but not the reverse. If multiple target features are mapped to a single source feature, then their values are combined with an aggregation protocol such as

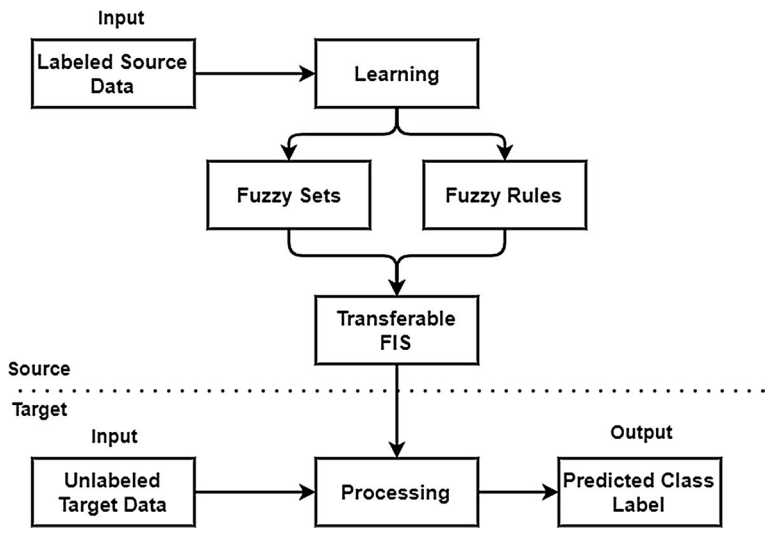

Fig. 5 Illustration for FuzzyTL, adapted from [85], which provides an overview of the methodology

minimum, maximum, average, etc. This mapping is applied to the target test data which maps its feature space to the most related source features and a final prediction is done with the prediction function trained on the source data. This FSR method can also be used as a preprocessing step to obtain a common feature space for later use with homogeneous transfer learning techniques.

Three variants of this method are proposed IFSR, UFSR, and ELFSR. Informed Feature-Space Remapping (IFSR) is used when limited target labels are provided. Thus, domain-independent metafeatures can be constructed with the aid of these labels. The metafeatures are found for this method by computing a feature-label co-occurrence value for each feature from the source and target space by calculating the expected value of a feature based on the given label from the labeled training data. Uniformed Feature-Space Remapping (UFSR) is used when no labeled target training data is provided. Thus, domain-dependent features are constructed due to the lack of labeled to assist in the mapping process. The metafeatures built for this method are created to model the conditional distributions such that the mapping process will select a feature pair with similar conditional distributions as this implies they are more similar for classifying a label and thus should be mapped together. Ensemble Learning via Feature-Space Remapping (ELFSR) extends these models by accepting multiple source domain data and creates an ensemble learning scenario. Here, an individual mapping from the target domain is learned to each one of the input source domains and a separate base classifier is built for each of these source domains. Combining these base classifiers creates an ensemble model. The final prediction for a class label in this model is calculated by a voting or stacking [90] scheme. Experiments were performed for activity recognition and document classification tasks which showed promising results for the proposed methods. These promising results are especially noted for the ensemble classifier as ensemble classification often provides better performance over singular models.

RLG, GLG

Liu et al. [11] proposed two models called Random Linear monotonic map Geodesic flow kernel model (RLG) and Grassmann Linear monotonic map Geodesic flow kernel model (GLG). These methods are designed for unsupervised HTL and use a labeled source domain along with an unlabeled target domain. Both models use a symmetric transformation to map the source and target data into a latent common subspace using a proposed Linear Monotonic Map (LMM) method. To ensure reliability and guarantee that knowledge is transferred appropriately, the resulting mappings must satisfy two conditions: (1) the mappings are monotonic and (2) performing the inverse of the maps results in the original feature spaces. Also, when transferring knowledge, it is important to maintain the same label space of the target domain after transformation as a significant change implies an error must exist. Therefore, in this method, the mappings must also satisfy the condition that the label space remains unchanged in this process. By using these rules when learning the mappings, one avoids particular conditions for negative transfer. From this latent subspace, both models employ the Geodesic Flow Kernel (GFK) [91, 92]. This is a homogeneous transfer learning method which is being used in this case to transfer knowledge between the projected domains.

The difference between the two proposed models lies within the generation of the transformation parameters for the LMM method. Random LMM GFK (RLG) randomly selects the parameters for the mappings. While this is computationally effective, it does not ensure a reliable model. Because of this, GLG is proposed which optimizes these parameters. Grassmann LMM GFK (GLG) uses a Grassmann manifold to measure the distance between the heterogeneous source and target domains as to optimize the parameters of the LMM through the construction of a cost function. The incorporation of the Grassmann manifold also ensures the geometric properties of the data remains unchanged during this process. The downside is that this method is computationally complex and has poor scalability. Optimization of such is done with an evolutionary algorithm through micro-analysis. Experiments were performed for credit assessment, text classification, and cancer detection to test the effectiveness of the proposed RLG and GLG. The proposed methods were compared with variations of KCCA [70], and GFK. The results indicated that both of the proposed methods achieve superior accuracy over the baselines for the tests conducted.

Methods which require limited target labels and no source labels

In this section, we present an HTL technique for the category under which one has limited target labels but aims to enhance the performance of the target classifier using only unlabeled source data. This is useful when one wants to gain the benefits of transferring knowledge but cannot collect a suitable fully-labeled domain. It is also attractive as unlabeled data is relatively cheap, thus it is to one's benefit to exploit it for a target task.

HTLIC

Zhu et al. [22] proposed an HTL method for image classification tasks called HTL for Image Classification (HTLIC). As discussed previously, image feature spaces contain visual attributes and often lack deterministic semantic meaning thus it may be costly to collect sufficient labeled image data to train an effective classifier on such features. For example, SIFT [93] descriptors, which are commonly used to represent visual data from images, model pixel data and lack the conceptual knowledge that can be found in other feature spaces, such as from text documents, to classify an instance. The proposed HTLIC method aims to solve such challenge by enhancing a target image classification task with limited labeled data by exploiting semantic knowledge derived from unlabeled text documents and unlabeled annotated images from an auxiliary source. This unlabeled auxiliary data is relatively inexpensive to collect and can enhance target image classification performance with this process.

First, a connection is found between the unlabeled text documents and unlabeled annotated images from an auxiliary source of related concepts. This is done using a two-layer bipartite graph where the top layer represents the relationship between the images and the tags while the bottom layer represents the relationship between the tags and the documents. Thus, from this we can form connections between the images and the text using the annotating tags to bridge the heterogeneous feature spaces gap. Using this connection, we can discover a common semantic space between the text and images by extracting latent semantic features from low-level image features using Latent Semantic Analysis (LSA) [94] and learning high-level features through collective matrix

factorization (CMF) [95]. Once such semantic view is constructed, the images in the target domain can be transformed to match this new enhanced feature space for image classification as it contains greater knowledge for classification than the original. This in turn is expected to provide greater classification performance. A standard classification model can then be trained from the transformed target images to make predictions on test images. In this case, linear SVM's were applied for the experiment. This method differs from aPLSA [96] for input requirements, as HTLIC also accepts unlabeled auxiliary documents rather than just unlabeled annotated images. Additionally, HTLIC requires limited target labels while aPLSA does not. Experiments were performed on image classification tasks to compare the proposed HTLIC to PCA, Tag [97], and an SVM trained only on the labeled target images. The proposed method was shown to be effective by having the highest classification accuracy for most tests.

Methods which require no target or source labels

In this section, we present an unsupervised HTL method which does not require any source or target labeled data. These unsupervised learning tasks aim to find a hidden structure from the data rather than directly determining the probabilities an instance can be categorized as a particular label. This is because, in this case, these labels are not given so one can only analyze patterns in the unlabeled data provided. Due to this, full classification tasks cannot be performed under this scenario, as in the previous categories, since no labels are provided. However, other tasks such as clustering [98] can still benefit from the use of transfer learning and an auxiliary source domain.

aPLSA

Yang et al. [96] proposed a method for unsupervised HTL for image clustering tasks called annotation-based PLSA (aPLSA). aPLSA performs these tasks through the use of user-annotated auxiliary images from social websites. This method does not require a labeled source or target, but rather utilizes such annotated images to find a correlation between the text and image features. This is to improve the performance of image clustering tasks even when the data is sparse as to learn a latent feature representation. Image clustering tasks aim to organize images into groups, or clusters, based on their similarities or differences. Similar images are grouped into the same cluster, while dissimilar ones are separated into other clusters. A major application for this task is when organizing search-engine results for a query-based image search. Motivated by such application, this method proposes to use heterogeneous transfer learning as to transfer knowledge from socially annotated auxiliary source images to enhance a target clustering task. Correspondences between the instances of the domains are not assumed in this case when using the heterogeneous auxiliary data. Therefore, the images may be unrelated and still improve performance.

The proposed method extends the use of Probabilistic Latent Semantic Analysis (PLSA) [99, 100] to incorporate annotated auxiliary information. The idea is to perform PLSA on both the source and target simultaneously and link them through common latent variables. First, PLSA is applied to the target images as to obtain an image instance-to-feature co-occurrence matrix. Along with this, PLSA is conducted on the annotated image data at the same time to obtain a text-to-image feature co-occurrence

matrix. These matrices are used to estimate the clustering function. While performing these simultaneously, the common latent variables are used to combine the two to learn high quality latent variables for use during target clustering. In other words, the image features are clustered into latent variables and then, simultaneously, the annotated auxiliary information is clustered in to the same latent variables. Thus, providing a connection for the transferring of knowledge with an enhanced feature representation for a target image clustering task. Experiments were performed for image clustering comparing the proposed aPLSA to K-means [101], PLSA [99], and STC [102]. For K-means and PLSA, experiments were performed applying these methods to the target data only, as well as combining the target and auxiliary data. The results showed that the proposed method had better results than the baselines as it had lower entropy, indicating less randomness for better classification results, under all experiments. This shows that one can use aPLSA to improve an image clustering task by transferring knowledge from unrelated auxiliary annotated images.

Methods for HTL preprocessing

In this section, we discuss a preprocessing method which may be applied before using the surveyed HTL methods as to have them operate with optimal performance. In this case, preprocessing methods are used to set the appropriate parameters required in each algorithm for optimal final performance. Currently, for methods that require such parameters, one has to manually select the proper values or test with a set of values, both of which are expensive and time consuming. This provides motivation for the use of a preprocessing method, as it can aid in selecting the proper parameter values.

DCN

Yang et al. [103] proposed an HTL preprocessing method based on using a Directed Cyclic Network (DCN). This method is proposed to be used before performing an HTL task to improve the overall effectiveness of the transfer learning method used. This method aims to measure the relatedness between $N \geq 2$ heterogeneous domains through co-occurrence data by learning transferred weights in order to determine whether a source domain is suitable for transferring knowledge to a target domain, as well as how much of that knowledge should be transferred. This co-occurrence data represents the same data from the target and source domains, but in different feature spaces. Therefore, one can use it independently to determine similarities amongst them to see whether a domain is suitable for knowledge transfer before applying computationally complex HTL methods. This is done by performing PCA on the co-occurrence data to compute their principal components for each feature space. The coefficients of such components have the same discriminatory order of significance for prediction purposes as the original domain's feature spaces. Using these coefficients, a directed cyclic network is built for each principal component to model the relationship amongst the domains and capture strong or weak relations amongst them. Here, each node represents a domain and the directed edge contains the transferred weight measuring the conditional dependence from one domain to another. The larger the weight value, the more knowledge can be transferred from one domain to another. Note that this graph can be asymmetric, as the i th domain may affect the j th more than the j th affects the i th domain. The optimal

network structure is computed through a Markov Chain Monte Carlo (MCMC) [104] method. Once a DCN is constructed for each principal component, they are merged into a single DCN by computing a weighted sum of the edge weights. More weight is given to the components used to construct the DCN that have higher contributions which is based on their corresponding eigenvalues. A low final edge weight means little knowledge can be transferred from such source domains implying it is unrelated and not suitable for knowledge transfer thus aiding in selecting the best source domain to use. The final transferred weights can also be used to set the parameters in the HTL methods for better overall performance. This was proven in the experiments under which the proposed method was incorporated into aPLSA [96], HTLIC [22], and CT-Learn [57]. Note that these each fall into different label requirement categories, thus implying the proposed method may be used for any HTL scenario.

Discussion

In this section, we compare and analyze various characteristics of the surveyed HTL methods as well as their empirical studies. This includes presenting patterns and differences exhibited amongst the different methodologies along with shortcomings of current research in this domain.

Comparative analysis

All of the surveyed HTL methods have the commonality that they propose a methodology for knowledge transfer when faced with differing feature spaces. They each propose a unique process to address such issue. One of the patterns that can be noticed amongst them to accomplish this is the use of co-occurrence data. TTI [36], TLRisk [37], OHTWC [19], CT-Learn [57], and DCN [103] use co-occurrence data as a bridge between the source and target feature spaces. Co-occurrence data is most often used for text aided image classification and is relatively cheap to collect. It is effective because of the assumption that the text surrounding an image on a webpage is semantically related to the image. Therefore, one can use this co-occurrence information as a bridge to relate text features and image features. This bridge, in essence, can be used to solve the issue of differing feature spaces. This has similarities to the pivot-based methods which bridge the feature spaces through individual pairs of instances from each domain that are linked together. This is done in CL-SCL [78] where each pivot pair contains an instance from each domain but they are directly semantically related. In cross-lingual text classification, a pivot is comprised of a word from each language and the link is modeled as a direct translation of the words which serves as a bridge between the languages. The pivot methods differ from the use of co-occurrence data as pivots are taken directly from instances in the input datasets while the co-occurrence data is its own independent input data set.

Another pattern that can be noted is the use of Canonical Correlation Analysis (CCA) to solve the differing feature space issue. Surveyed methods that employ this include CT-SVM [70], HDCC [76], SMVCCAE [64], and SSMVCCAE [64]. CCA aims to maximize the similarity between two variable sets by projecting them onto a correlation subspace. Both CT-SVM and SMVCCAE/SSMVCCAE utilize the ρ correlation coefficient calculated in CCA which measures the correlation/similarity between the two variable sets

in each dimension. CT-SVM uses this metric to suppress learned SVM's on dimensions that have poor correlation to increase focus on those that are more similar. On the other hand, SMVCCAE/SSMVCCAE propose using a weighted voting system based on the ρ values.

Many of these methods are also application-specific and it would be difficult or impossible to port them to other applications. Most of the surveyed methods are designed for, or applied to, image and cross-lingual text classification tasks which are common applications for HTL. Methods designed specifically for image classification such as HTLIC [22] and TTI [36] would be impossible for use on non-image domains. Also, porting application-specific methods such as FuzzyTL [85] may not have optimal performance outside of their domain. Co-occurrence data, as required by some previously discussed methods, is also very application-specific and it may be difficult to obtain or source for certain applications.

Additionally, most of the surveyed methods also do not address the issue of differing label spaces or differences in marginal/conditional distributions. When using a heterogeneous auxiliary dataset, the class labels may not match in real-world scenarios. Specifically, FSR [12], HeMap [26], and Proactive HTL [54] are the only surveyed methods that directly address this issue of differing label spaces. Also, when drawing data from a different domain, this auxiliary data may exhibit data distribution differences and further adaptation is required in this case to correct this after resolving the feature space gap.

Moreover, the surveyed methods also lack investigation into many common real-world machine learning challenges that may arise during transfer learning tasks. This includes noise (either class or attribute), class imbalance, outliers, high-dimensionality, as well others [105–107]. These issues have a unique impact on transfer learning tasks as they may present in either the source or the target domain which causes changing circumstances [108–110].

Performance analysis

All of the surveyed HTL methods conduct an empirical study to investigate the effectiveness of their proposed algorithm for different tasks. Most of the surveyed methods conduct their study with a uniquely designed experimental testbed and compare their proposed algorithm with few (if any) other HTL algorithms. All of the proposed methods claim they have better performance than the selected baseline algorithms of the study. These baseline algorithms are often not originally designed for HTL tasks and are expected have inferior performance when applied to these tasks. The most common baselines used were variations of SVM trained on only the available source or target data. One can note that it is not meaningful to compare an algorithm with others that would inherently have inferior performance in the designed testbed. This, in essence, would not provide strong evidence towards the superiority of the proposed method. On the other hand, a more valuable discovery would be if the proposed algorithm demonstrated better performance over other HTL algorithms that have the *same label requirements*. Only few of the studies conducted utilized other HTL methods in the same label requirement category.

Specifically, none of the algorithms in the "Methods which require no target labels" section compare the performance of their algorithm to others in the same category.

Currently, the issue is that one cannot directly compare the performance of the surveyed algorithms using the data provided. This is because there are little commonalities observed in the experimental testbeds of the surveyed methods in each category. Comparing these methods using the little overlap that may exist would not provide meaningful conclusions due to the varied nature of the studies. Thus, these are shortcomings of the current research. This is also the limitation of current research in this domain as we cannot provide a direct performance comparison of the surveyed methods. It would be insightful to have a comparative study conducted to evaluate the effectiveness of the surveyed algorithms on a common, fair experimental testbed. To do this, a standardized test framework could be developed for heterogeneous transfer learning algorithms, as was done for homogeneous transfer learning algorithms [111, 112].

Conclusion

When faced with little or no labeled training data, a model trained on such data will have insufficient discriminatory ability and would be unable to predict accurately. In order to handle this issue, transfer learning techniques have received significant focus in research communities. Specifically, heterogeneous transfer learning has been recently studied as it broadens the application of current transfer learning methods.

In this paper, we provided a comprehensive survey of 38 methods which are designed to handle these heterogeneous transfer learning tasks. We organized these methods based on labeling requirements. This is because the motivation for transfer learning is to improve performance for a target task when faced with little or no labeled data thus this organization is intuitive as one must select an appropriate algorithm based on one's available resources. In addition, we also provided an in-depth discussion and analysis.

For heterogeneous transfer learning tasks, one may face the issue of differing feature spaces along with any combination of differing feature dimensions, label spaces, or data distributions. Because of this, these tasks are often more challenging but are present in many real-world applications such as cross-language text classification, image classification, as well as many others.

Overall, the surveyed methods aim to find a commonality between the source and target domains to bring the feature spaces to a common representation as to extract performance-enhancing knowledge without the inherent cross-domain noise.

Future work

First, as HTL is a relatively recent area in research, there are still many areas of interest to investigate. One of these is the issue of scalability. The rise of big data has led to many real-world applications that could benefit from transfer learning on very large datasets. Unfortunately, most HTL techniques are highly complex and scale very poorly because of this. Thus, there is a need to investigate and develop highly computationally efficient methods which can be applied to large datasets.

Second, another area to investigate is negative transfer. Most of these methods do not employ safeguards against negative transfer. More specifically, none of the surveyed techniques have such safeguards except RLG/GLG. This technique claims to avoid negative transfer by setting the right environment for knowledge transfer by ensuring the label spaces do not change after transformation; though this method does not necessarily prevent

negative transfer. Thus, there is a need for research of methods which implement such safeguards or techniques that can be incorporated into current HTL methods for this task.

Third, further investigation into correcting differences in marginal and conditional distributions for HTL tasks should be conducted, as most of these techniques focus on the issue of bridging the feature space gap but do not effectively consider data distribution issues once the feature spaces have been aligned.

Fourth, most of these methods assume the label spaces are the same between the source and target but under certain HTL tasks this assumption does not hold thus further investigation into such problem would be beneficial.

Finally, these methods do not investigate the effect of the presence of noise or class-imbalance in either the source or target domain on the performance of a transfer learning task thus an investigation into such may prove useful for real-world applications where these issues are common.

Authors' contributions
TMK introduced this topic to OD and provided technical guidance throughout the research and writing phases of the manuscript. OD performed the literature review, analysis, and writing of the manuscript. Both authors read and approved the final manuscript.

Acknowledgements
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Availability of data and materials
Not applicable.

Consent for publication
Not applicable.

Ethics approval and consent to participate
Not applicable.

Funding
Not applicable.

Publisher's Note

Received: 6 July 2017 Accepted: 18 September 2017
Published online: 26 September 2017

References

1. Witten IH, Frank E, Hall MA, Pal CJ. Data mining: practical machine learning tools and techniques. Burlington: Morgan Kaufmann; 2016.
2. Seliya N, Khoshgoftaar TM, Van Hulse J. A study on the relationships of classifier performance metrics. In: 21st international conference on tools with artificial intelligence, 2009. ICTAI'09. New York: IEEE; 2009. p. 59-66.
3. Pan SJ. Transfer learning. In: Aggarwal CC, editor. Data classification: algorithms and applications. CRC Press; 2014. p. 537–70. http://www.crcnetbase.com/doi/abs/10.1201/b17320-22.
4. Ieong S. Probability theory review for machine learning. Rep. Stanford University, 06 Nov. 2006. Web.
5. Li Y, Wei B, Chen H, Jiang L, Li Z. Cross-domain learning based traditional Chinese medicine medical record classification. In: 10th international conference on intelligent systems and knowledge engineering (ISKE), 2015. New York: IEEE; 2015. p. 335–40.
6. Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng. 2010;22(10):1345–59.
7. Weiss K, Khoshgoftaar TM, Wang DD. A survey of transfer learning. J Big Data. 2016;3(1):1–40.
8. Rosenstein MT, Marx Z, Kaelbling LP, Dietterich TG. To transfer or not to transfer. In: NIPS 2005 workshop on transfer learning, vol. 898. 2005.
9. Landset S, Khoshgoftaar TM, Richter AN, Hasanin T. A survey of open source tools for machine learning with big data in the Hadoop ecosystem. J Big Data. 2015;2(1):24.
10. Najafabadi MM, Villanustre F, Khoshgoftaar TM, Seliya N, Wald R, Muharemagic E. Deep learning applications and challenges in big data analytics. J Big Data. 2015;2(1):21.

11. Liu F, Zhang G, Lu H, Lu J. Heterogeneous unsupervised cross-domain transfer learning. arXiv preprint arXiv:1701.02511. 2017.
12. Feuz KD, Cook DJ. Transfer learning across feature-rich heterogeneous feature spaces via feature-space remapping (FSR). ACM Trans Intell Syst Technol. 2015;6:3.
13. Zhao P, Hoi SC. OTL: a framework of online transfer learning. In: Proceedings of the 27th international conference on machine learning (ICML-10). 2010.
14. Shalev-Shwartz S, Singer Y. Online learning: theory, algorithms, and applications. 2007.
15. Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. Psychol Rev. 1958;65(6):386.
16. Freund Y, Schapire RE. Large margin classification using the perceptron algorithm. Mach Learn. 1999;37(3):277–96.
17. Xu C, Tao D, Xu C. A survey on multi-view learning. 2013. arXiv preprint arXiv:1304.5634.
18. Shalev-Shwartz S, Crammer K, Dekel O, Singer Y. Online passive-aggressive algorithms. In: Advances in neural information processing systems. 2004. p. 1229–36.
19. Yan Y, Wu Q, Tan M, Min H. Online heterogeneous transfer learning by weighted offline and online classifiers. In: Computer vision–ECCV 2016 workshops. Berlin: Springer International Publishing; 2016. p. 467–74.
20. Benesty J, Chen J, Huang Y, Cohen I. Pearson correlation coefficient. In: Noise reduction in speech processing. Berlin: Springer; 2009. p. 1–4.
21. Freund Y, Schapire RE. A desicion-theoretic generalization of on-line learning and an application to boosting. In: European conference on computational learning theory. Berlin: Springer; 1995.
22. Zhu Y, Chen Y, Lu Z, Pan SJ, Xue GR, Yu Y, Yang Q. Heterogeneous transfer learning for image classification. In: AAAI. 2011.
23. Duan L, Xu D, Tsang I. Learning with augmented features for heterogeneous domain adaptation. 2012. arXiv preprint arXiv:1206.4660.
24. Daumé H III. Frustratingly easy domain adaptation. 2009. arXiv preprint arXiv:0907.1815.
25. Shawe-Taylor J, Cristianini N. Kernel methods for pattern analysis. Cambridge: Cambridge university press; 2004.
26. Shi X, Liu Q, Fan W, Philip SY, Zhu R. Transfer learning on heterogenous feature spaces via spectral transformation. In: 2010 IEEE 10th international conference on data mining (ICDM). New York: IEEE; 2010. p. 1049–54.
27. Wang C, Mahadevan S. Heterogeneous domain adaptation using manifold alignment. In: IJCAI proceedings-international joint conference on artificial intelligence, vol. 22, No. 1. 2011.
28. Kulis B, Saenko K, Darrell T. What you saw is not what you get: domain adaptation using asymmetric kernel transforms. In: 2011 IEEE conference on computer vision and pattern recognition (CVPR). New York: IEEE; 2011.
29. Mozafari AS, Jamzad M. A SVM-based model-transferring method for heterogeneous domain adaptation. Pattern Recogn. 2016;56:142–58.
30. Mozafari AS, Jamzad M. Heterogeneous domain adaptation using previously learned classifier for object detection problem. In: 2014 IEEE international conference on image processing (ICIP). New York: IEEE; 2014.
31. Zhou JT, Tsang IW, Pan SJ, Tan M. Heterogeneous domain adaptation for multiple classes. In: AISTATS. 2014. p. 1095–1103.
32. Ando RK, Zhang T. A framework for learning predictive structures from multiple tasks and unlabeled data. J Mach Learn Res. 2005;6(Nov):1817–53.
33. Donoho DL. Compressed sensing. IEEE Trans Inf Theory. 2006;52(4):1289–306.
34. Dietterich TG, Bakiri G. Solving multiclass learning problems via error-correcting output codes. J Artif Intell Res. 1995;2:263–86.
35. Jolliffe I. Principal component analysis. New York: Wiley; 2002.
36. Qi GJ, Aggarwal C, Huang T. Towards semantic knowledge propagation from text corpus to web images. In: Proceedings of the 20th international conference on world wide web. New York: ACM; 2011.
37. Dai W, Chen Y, Xue GR, Yang Q, Yu Y. Translated learning: transfer learning across different feature spaces. In: NIPS. 2008. p. 353–60.
38. Lafferty J, Zhai C. Document language models, query models, and risk minimization for information retrieval. In: Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval. New York: ACM; 2001.
39. Saenko K, Kulis B, Fritz M, Darrell T. Adapting visual category models to new domains. In: Proc. ECCV, September 2010, Heraklion, Greece.
40. Davis JV, et al. Information-theoretic metric learning. In: Proceedings of the 24th international conference on machine learning. New York: ACM; 2007.
41. Sukhija S, Krishnan NC, Singh G. Supervised heterogeneous domain adaptation via random forests.
42. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
43. Wang C, Mahadevan S. A general framework for manifold alignment. In: AAAI fall symposium: manifold learning and its applications. 2009.
44. Belkin M, Niyogi P, Sindhwani V. Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. J Mach Learn Res. 2006;7(Nov):2399–434.
45. Hotelling H. Relations between two sets of variates. Biometrika. 1936;28(3/4):321–77.
46. Hubert Tsai YH, Yeh YR, Frank Wang YC. Learning cross-domain landmarks for heterogeneous domain adaptation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
47. Li W, Duan L, Dong X, Tsang IW. Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation. IEEE Trans Pattern Anal Mach Intell. 2014;36(6):1134–48.
48. Hoffman J, Rodner E, Donahue J, Darrell T, Saenko K. Efficient learning of domain-invariant image representations. 2013. arXiv preprint arXiv:1301.3224.
49. Han P, Wu X. Incremental discriminant learning for heterogeneous domain adaptation. In: 2015 IEEE international conference on data mining workshop (ICDMW). New York: IEEE; 2015.
50. Hall PM, Marshall AD, Martin RR. Incremental eigenanalysis for classification. In: BMVC, vol. 98. 1998.
51. Harel M, Mannor S. Learning from multiple outlooks. 2010. arXiv preprint arXiv:1005.0027.

52. Wilcox R. Trimming and winsorization. Encyclopedia of biostatistics. 1998.
53. Shi X, Liu Q, Fan W, Yang Q, Yu PS. Predictive modeling with heterogeneous sources. In: Proceedings of the 2010 SIAM international conference on data mining. Philadelphia: Society for Industrial and Applied Mathematics; 2010. p. 814–25.
54. Moon S, Carbonell J. Proactive transfer learning for heterogeneous feature and label spaces. In: Joint European conference on machine learning and knowledge discovery in databases. Berlin: Springer International Publishing; 2016.
55. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. In: ICLR. 2013.
56. Kloft M, Brefeld U, Sonnenburg S, Zien A. Lp-norm multiple kernel learning. J Mach Learn Res. 2011;12(Mar):953–97.
57. Ng MK, Wu Q, Ye Y. Co-transfer learning via joint transition probability graph based method. In: Proceedings of the 1st international workshop on cross domain knowledge discovery in web and social network mining. New York: ACM; 2012.
58. Wu Q, Ng MK, Ye Y. Cotransfer learning using coupled Markov chains with restart. IEEE Intell Syst. 2014;29(4):26–33. doi:10.1109/MIS.2013.32.
59. Tong H, Faloutsos C, Pan JY. Random walk with restart: fast solutions and applications. Knowl Inf Syst. 2008;14(3):327–46.
60. Xiao M, Guo Y. Feature space independent semi-supervised domain adaptation via kernel matching. IEEE Trans Pattern Anal Mach Intell. 2015;37(1):54–66.
61. Gretton A, Bousquet O, Smola A, Scholkopf B. Measuring statistical dependence with Hilbert–Schmidt norms. In: International conference on algorithmic learning theory. Berlin: Springer; 2005. p. 63-77.
62. Xiao M, Guo Y. Semi-supervised subspace co-projection for multi-class heterogeneous domain adaptation. In: Joint European conference on machine learning and knowledge discovery in databases. Berlin: Springer International Publishing; 2015.
63. Borgwardt KM, Gretton A, Rasch MJ, Kriegel HP, Schölkopf B, Smola AJ. Integrating structured biological data by kernel maximum mean discrepancy. Bioinformatics. 2006;22(14):e49–57.
64. Samat A, Persello C, Gamba P, Liu S, Abuduwaili J, Li E. Supervised and semi-supervised multi-view canonical correlation analysis ensemble for heterogeneous domain adaptation in remote sensing image classification. Remote Sens. 2017;9(4):337.
65. Chen WY, Hsu TM, Tsai YH, Wang YC, Chen MS. Transfer neural trees for heterogeneous domain adaptation. In: European conference on computer vision. Berlin: Springer International Publishing; 2016. p. 399–414.
66. Kontschieder P, Fiterau M, Criminisi A, Rota Bulo S. Deep neural decision forests. In: Proceedings of the IEEE international conference on computer vision. 2015. p. 1467–75.
67. Wang X, Ma Y, Cheng Y, Zou L, Rodrigues JJPC. Heterogeneous domain adaptation network based on autoencoder. J Parallel Distrib Comput. 2017. doi:10.1016/j.jpdc.2017.06.003.
68. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436–44.
69. Hsieh YT, Tao SY, Tsai YHH, et al. Recognizing heterogeneous cross-domain data via generalized joint distribution adaptation. In: Proceedings of IEEE international conference on multimedia and expo. 2016. p. 1–6.
70. Yeh YR, Huang CH, Wang YC. Heterogeneous domain adaptation and classification by exploiting the correlation subspace. IEEE Trans Image Process. 2014;23(5):2009–18.
71. Zhou JT, Pan SJ, Tsang IW, Yan Y. Hybrid heterogeneous transfer learning through deep learning. In: AAAI. 2014. p. 2213–20.
72. Vincent P, Larochelle H, Bengio Y, Manzagol PA. Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th international conference on machine learning. New York: ACM; 2008. p. 1096-103.
73. Chen M, Xu Z, Weinberger K, Sha F. Marginalized denoising autoencoders for domain adaptation. 2012. arXiv preprint arXiv:1206.4683.
74. Ngiam J, Khosla A, Kim M, Nam J, Lee H, Ng AY. Multimodal deep learning. In: Proceedings of the 28th international conference on machine learning (ICML-11). 2011. p. 689–96.
75. Vinokourov A, Cristianini N, Shawe-Taylor J. Inferring a semantic representation of text via cross-language correlation analysis. In: NIPS, vol. 1, No. 3. 2002.
76. Wang H, Wu X, Jia Y. Heterogeneous domain adaptation method for video annotation. IET Comput Vis. 2016;11(2):181–7.
77. Bruzzone L, Marconcini M. Domain adaptation problems: a DASVM classification technique and a circular validation strategy. IEEE Trans Pattern Anal Mach Intell. 2010;32(5):770–87.
78. Prettenhofer P, Stein B. Cross-language text classification using structural correspondence learning. In: Proceedings of the 48th annual meeting of the association for computational linguistics. Association for Computational Linguistics; 2010.
79. Blitzer J, McDonald R, Pereira F. Domain adaptation with structural correspondence learning. In: Proceedings of the 2006 conference on empirical methods in natural language processing. Association for Computational Linguistics; 2006. p. 120–8.
80. Nam J, Kim S. Heterogeneous defect prediction. In: Proceedings of the 2015 10th joint meeting on foundations of software engineering. New York: ACM; 2015.
81. Van Hulse J, Khoshgoftaar TM, Napolitano A, Wald R. Feature selection with high-dimensional imbalanced data. In: IEEE international conference on data mining workshops, 2009. ICDMW'09. New York: IEEE; 2009. p. 507–514.
82. Gao K, Khoshgoftaar TM, Wang H, Seliya N. Choosing software metrics for defect prediction: an investigation on feature selection techniques. Softw Pract Exp. 2011;41(5):579–606.
83. Shivaji S, Whitehead EJ, Akella R, Kim S. Reducing features to improve code change-based bug prediction. IEEE Trans Softw Eng. 2013;39(4):552–69.
84. Matousek J, Gärtner B. Understanding and using linear programming. Berlin: Springer Science & Business Media; 2007.
85. Shell J, Coupland S. Towards fuzzy transfer learning for intelligent environments. In: International joint conference on ambient intelligence. Berlin: Springer; 2012.

86. Zadeh LA. Fuzzy logic. Computer. 1988;21:83–93.
87. Wang LX. The WM method completed: a flexible fuzzy system approach to data mining. IEEE Trans Fuzzy Syst. 2003;11(6):768–82.
88. Wang LX, Mendel JM. Generating fuzzy rules by learning from examples. IEEE Trans Syst Man Cybern. 1992;22(6):1414–27.
89. Dillon Feuz K, Cook DJ. Heterogeneous transfer learning for activity recognition using heuristic search techniques. Int J Pervasive Comput Commun. 2014;10(4):393–418.
90. Wolpert DH. Stacked generalization. Neural Netw. 1992;5(2):241–59.
91. Gong B, Grauman K, Sha F. Learning kernels for unsupervised domain adaptation with applications to visual object recognition. Int J Comput Vis. 2014;109(1-2):3–27.
92. Gong B, Shi Y, Sha F, Grauman K. Geodesic flow kernel for unsupervised domain adaptation. In: 2012 IEEE conference on computer vision and pattern recognition (CVPR). New York: IEEE; 2012. p. 2066–73.
93. Lowe DG. Distinctive image features from scale-invariant keypoints. Int J Comput Vis. 2004;60(2):91–110.
94. Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R. Indexing by latent semantic analysis. J Am Soc Inf Sci. 1990;41(6):391.
95. Singh AP, Gordon GJ. Relational learning via collective matrix factorization. In: Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining. New York: ACM; 2008.
96. Yang Q, Chen Y, Xue GR, Dai W, Yu Y. Heterogeneous transfer learning for image clustering via the social web. In: Proceedings of the joint conference of the 47th annual meeting of the ACL and the 4th international joint conference on natural language processing of the AFNLP, vol 1. Association for Computational Linguistics; 2009. p. 1–9.
97. Wang G, Hoiem D, Forsyth D. Building text features for object image classification. In: IEEE conference on computer vision and pattern recognition, 2009. CVPR 2009. New York: IEEE; 2009.
98. Zhong S, Khoshgoftaar TM, Seliya N. Clustering-based network intrusion detection. Int J Reliab Qual Saf Eng. 2007;14(02):169–87.
99. Hofmann T. Probabilistic latent semantic analysis. In: Proceedings of the fifteenth conference on uncertainty in artificial intelligence. Burlington: Morgan Kaufmann Publishers Inc.; 1999.
100. Hofmann T. Unsupervised learning by probabilistic latent semantic analysis. Mach Learn. 2001;42(1):177–96.
101. MacQueen J. Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, vol. 1, No. 14. 1967.
102. Dai W, Yang Q, Xue GR, Yu Y. Self-taught clustering. In: Proceedings of the 25th international conference on Machine learning. New York: ACM; 2008. p. 200–7.
103. Yang L, Jing L, Jian Y, Ng MK. Learning transferred weights from co-occurrence data for heterogeneous transfer learning. IEEE Trans Neural Netw Learn Syst. 2016;27(11):2187–200.
104. Giudici P, Castelo R. Improving Markov chain Monte Carlo model search for data mining. Mach Learn. 2003;50(1-2):127–58.
105. Khoshgoftaar TM, Van Hulse J, Napolitano A. Comparing boosting and bagging techniques with noisy and imbalanced data. IEEE Trans Syst Man Cybern Part A Syst Hum. 2011;41(3):552–68.
106. Van Hulse JD, Khoshgoftaar TM, Huang H. The pairwise attribute noise detection algorithm. Knowl Inf Syst. 2007;11(2):171–90.
107. Bauder RA, Khoshgoftaar TM. A probabilistic programming approach for outlier detection in healthcare claims. In: 2016 15th IEEE international conference on machine learning and applications (ICMLA). New York: IEEE; 2016. p. 347–54.
108. Weiss KR, Khoshgoftaar TM. An investigation of transfer learning and traditional machine learning algorithms. In: 2016 IEEE 28th international conference on tools with artificial intelligence (ICTAI). New York: IEEE; 2016. p. 283–90.
109. Weiss KR, Khoshgoftaar TM. Investigating transfer learners for robustness to domain class imbalance. In: 2016 15th IEEE international conference on machine learning and applications (ICMLA). New York: IEEE; 2016. p. 207–13.
110. Weiss KR, Khoshgoftaar TM. Analysis of transfer learning performance measures. In: IEEE international conference on information reuse and integration (IRI). 2017. p. 338–44.
111. Weiss KR, Khoshgoftaar TM, Rehman O. Designing a testing framework for transfer learning algorithms (Application Paper). In: 2016 IEEE 17th international conference on information reuse and integration (IRI). New York: IEEE; 2016. p. 152–9.
112. Weiss KR, Khoshgoftaar TM. Analysis of a transfer learning application using the transfer learning test framework. In: 23rd ISSAT international conference on reliability and quality in design. 2017. p. 151–6.