Chapter 3
Transfer Learning Techniques

Karl Weiss, Taghi M. Khoshgoftaar and DingDing Wang

Introduction

The field of data mining and machine learning has been widely and successfully used in many applications where patterns from past information (training data) can be extracted in order to predict future outcomes [1]. Traditional machine learning is characterized by training data and testing data having the same input feature space and the same data distribution. When there is a difference in data distribution between the training data and test data, the results of a predictive learner can be degraded [2]. In certain scenarios, obtaining training data that matches the feature space and predicted data distribution characteristics of the test data can be difficult and expensive. Therefore, there is a need to create a high-performance learner for a target domain trained from a related source domain. This is the motivation for transfer learning.

Transfer learning is used to improve a learner from one domain by transferring information from a related domain. We can draw from real-world non-technical experiences to understand why transfer learning is possible. Consider an example of two people who want to learn to play the piano. One person has no previous experience playing music, and the other person has extensive music knowledge through playing the guitar. The person with an extensive music background will be able to learn the piano in a more efficient manner by transferring previously learned music knowledge to the task of learning to play the piano [3]. One person is able to take information from a previously learned task and use it in a beneficial way to learn a related task.

Looking at a concrete example from the domain of machine learning, consider the task of predicting text sentiment of product reviews where there exists an abundance of labeled data from digital camera reviews. If the training data and the

This chapter has been adopted from the Journal of Big Data, Borko Furht and Taghi Khoshgoftar, Editors-in-Chief. Springer, June 2016.

B. Furht and F. Villanustre, *Big Data Technologies and Applications*,
DOI 10.1007/978-3-319-44550-2_3

target data are both derived from digital camera reviews, then traditional machine learning techniques are used to achieve good prediction results. However, in the case where the training data is from digital camera reviews and the target data is from food reviews, then the prediction results are likely to degrade due to the differences in domain data. Digital camera reviews and food reviews still have a number of characteristics in common, if not exactly the same. They both are written in textual form using the same language, and they both express views about a purchased product. Because these two domains are related, transfer learning can be used to potentially improve the results of a target learner [3]. An alternative way to view the data domains in a transfer learning environment is that the training data and the target data exist in different sub-domains linked by a high-level common domain. For example, a piano player and a guitar player are subdomains of a musician domain. Further, a digital camera review and a food review are subdomains of a review domain. The high-level common domain determines how the subdomains are related.

As previously mentioned, the need for transfer learning occurs when there is a limited supply of target training data. This could be due to the data being rare, the data being expensive to collect and label, or the data being inaccessible. With big data repositories becoming more prevalent, using existing datasets that are related to, but not exactly the same as, a target domain of interest makes transfer learning solutions an attractive approach. There are many machine learning applications that transfer learning has been successfully applied to including text sentiment classification [4], image classification [5–7], human activity classification [8], software defect classification [9], and multi-language text classification [10–12].

This survey paper aims to provide a researcher interested in transfer learning with an overview of related works, examples of applications that are addressed by transfer learning, and issues and solutions that are relevant to the field of transfer learning. This survey paper provides an overview of current methods being used in the field of transfer learning as it pertains to data mining tasks for classification, regression, and clustering problems; however, it does not focus on transfer learning for reinforcement learning (for more information on reinforcement learning see Taylor and Stone [13]. Information pertaining to the history and taxonomy of transfer learning is not provided in this survey paper, but can be found in the paper by Pan and Yang [3]. Since the publication of the transfer learning survey paper by Pan and Yang [3] in 2010, there have been over 700 academic papers written addressing advancements and innovations on the subject of transfer learning. These works broadly cover the areas of new algorithm development, improvements to existing transfer learning algorithms, and algorithm deployment in new application domains. The selected surveyed works in this paper are meant to be diverse and representative of transfer learning solutions in the past five years. Most of the surveyed papers provide a generic transfer learning solution; however, some surveyed papers provide solutions that are specific to individual applications. This paper is written with the assumption the reader has a working knowledge of machine learning. For more information on machine learning see Witten and Frank [1]. The surveyed works in this paper are intended to present a high-level description of proposed solutions with unique and

salient points being highlighted. Experiments from the surveyed papers are described with respect to applied applications, other competing solutions tested, and overall relative results of the experiments. This survey paper provides a section on heterogeneous transfer learning which, to the best of our knowledge, is unique. Additionally, a list of software downloads for various surveyed papers is provided, which is unique to this paper.

The remainder of this paper is organized as follows. In ``Definitions of Transfer Learning'' section provides definitions and notations of transfer learning. In ``Homogeneous Transfer Learning'' and ``Heterogeneous Transfer Learning'' sections provide solutions on homogeneous and heterogeneous transfer learning, respectively. In Negative Transfer section provides information on negative transfer as it pertains to transfer learning. In Transfer Learning Applications section provides examples of transfer learning applications. In Conclusion and Discussion section summarizes and discusses potential future research work. The Appendix provides information on software downloads for transfer learning.

Definitions of Transfer Learning

The following section lists the notation and definitions used for the remainder of this paper. The notation and definitions in this section match those from the survey paper by Pan and Yang [3], if present in both papers, to maintain consistency across both surveys. To provide illustrative examples of the definitions listed below, a machine learning application of software module defect classification is used where a learner is trained to predict whether a software module is defect prone or not.

A domain D is defined by two parts, a feature space X and a marginal probability distribution P(X), where $X = \{x_1, \ldots, x_n\} \in X$. For example, if the machine learning application is software module defect classification and each software metric is taken as a feature, then x_i is the i-th feature vector (instance) corresponding to the i-th software module, n is the number of feature vectors in X, X is the space of all possible feature vectors, and X is a particular learning sample. For a given domain D, a task T is defined by two parts, a label space Y, and a predictive function f(•), which is learned from the feature vector and label pairs $\{x_i, y_i\}$ where $x_i \in X$ and $y_i \in Y$. Referring to the software module defect classification application, Y is the set of labels and in this case contains true and false, y_i takes on a value of true or false, and f (x) is the learner that predicts the label value for the software module x. From the definitions above, a domain $D = \{X, P(X)\}$ and a task $T = \{Y, f(\bullet)\}$. Now, D_S is defined as the source domain data where $D_S = \{ (x_{S1}, y_{S1}) \ldots, (x_{Sn}, y_{Sn})\}$, where $x_{Si} \in X_S$ is the i-th data instance of D_S and $y_{Si} \in Y_S$ is the corresponding class label for x_{Si}. In the same way, D_T is defined as the target domain data where $D_T = \{ (x_{T1}, y_{T1}) \ldots, (x_{Tn}, y_{Tn})\}$, where $x_{Ti}, \in X_T$ is the i-th data instance of D_T and $y_{Ti}, \in Y_T$ is the corresponding class label for x_{Ti}. Further, the source task is notated as T_S, the target task as T_T, the source predictive function as $f_S(\bullet)$, and the target predictive function as $f_T(\bullet)$.

Transfer Learning is now formally defined. Given a source domain D_S with a corresponding source task T_S and a target domain D_T with a corresponding task T_T, transfer learning is the process of improving the target predictive function $f_T(\cdot)$ by using the related information from D_S and T_S, where $D_S \neq D_T$ or $T_S \neq T_T$. The single source domain defined here can be extended to multiple source domains. Given the definition of transfer learning, since $D_S = \{X_S, P(X_S)\}$ and $D_T = \{X_T, P(X_T)\}$, the condition where $D_S \neq D_T$ means that $X_S \neq X_T$ and/or $P(X_S) \neq P(X_T)$. The case where $X_S \neq X_T$ with respect to transfer learning is defined as heterogeneous transfer learning. The case where $X_S = X_T$ with respect to transfer learning is defined as homogeneous transfer learning. Going back to the example of software module defect classification, heterogeneous transfer learning is the case where the source software project has different metrics (features) than the target software project. Alternatively, homogeneous transfer learning is when the software metrics are the same for both the source and the target software projects. Continuing with the definition of transfer learning, the case where $P(X_S) \neq P(X_T)$ means the marginal distributions in the input spaces are different between the source and the target domains. Shimodaira [2] demonstrated that a learner trained with a given source domain will not perform optimally on a target domain when the marginal distributions of the input domains are different. Referring to the software module defect classification application, an example of marginal distribution differences is when the source software program is written for a user interface system and the target software program is written for DSP signaling decoder algorithm. Another possible condition of transfer learning (from the definition above) is $T_S \neq T_T$, and it was stated that $T = \{Y, f(\cdot)\}$ or to rewrite this, $T = \{Y, P(Y|X)\}$. Therefore, in a transfer learning environment, it is possible that $Y_S \neq Y_T$ and/or $P(Y_S|X_S) \neq P(Y_T|X_T)$. The case where $P(Y_S|X_S) \neq P(Y_T|X_T)$ means the conditional probability distributions between the source and target domains are different. An example of a conditional distribution mismatch is when a particular software module yields different fault prone results in the source and target domains. The case of $Y_S \neq Y_T$ refers to a mismatch in the class space. An example of this case is when the source software project has a binary label space of true for defect prone and false for not defect prone, and the target domain has a label space that defines five levels of fault prone modules. Another case that can cause discriminative classifier degradation is when $P(Y_S) \neq P(Y_T)$, which is caused by an unbalanced labeled data set between the source and target domains. The case of traditional machine learning is $D_S = D_T$ and $T_S = T_T$. The common notation used in this paper is summarized in Table 3.1.

Table 3.1 Summary of commonly used notation

Notation	Description	Notation	Description
X	Input feature space	P(X)	Marginal distribution
Y	Label space	P(Y|X)	Conditional distribution
T	Predictive learning task	P(Y)	Label distribution
Subscript S	Denotes source	D_S	Source domain data
Subscript T	Denotes target	D_T	Target domain data

To elaborate on the distribution issues that can occur between the source and target domains, the application of natural language processing is used to illustrate. In natural language processing, text instances are often modeled as a bag-of-words where a unique word represents a feature. Consider the example of review text where the source covers movie reviews and the target covers book reviews. Words that are generic and domain independent should occur at a similar rate in both domains. However, words that are domain specific are used more frequently in one domain because of the strong relationship with that domain topic. This is referred to as frequency feature bias and will cause the marginal distribution between the source and target domains to be different ($P(X_S) \neq P(X_T)$). Another form of bias is referred to as context feature bias and this will cause the conditional distributions to be different between the source and target domains ($P(Y_S|X_S) \neq P(Y_T|X_T)$). An example of context feature bias is when a word can have different meanings in two domains. A specific example is the word "monitor" where in one domain it is used as a noun and in another domain it is used as a verb. Another example of context feature bias is with sentiment classification when a word has a positive meaning in one domain and a negative meaning in another domain. The word "small" can have a good meaning if describing a cell phone but a bad meaning if describing a hotel room. A further example of context feature bias is demonstrated in the case of document sentiment classification of reviews where the source domain contains reviews of one product written in German and the target domain contains reviews of a different product written in English. The translated words from the source document may not accurately represent the actual words used in the target documents. An example is the case of the German word "betonen", which translates to the English word "emphasize" by Google translator. However, in the target documents the corresponding English word used is "highlight" [12].

Negative transfer, with regards to transfer learning, occurs when the information learned from a source domain has a detrimental effect on a target learner. More formally, given a source domain D_S, a source task T_S, a target domain D_T, a target task T_T, a predictive learner $f_{T1}(\bullet)$ trained only with D_T, and a predictive learner $f_{T2}(\bullet)$ trained with a transfer learning process combining D_T and D_S, negative transfer occurs when the performance of $f_{T1}(\bullet)$ is greater than the performance of $f_{T2}(\bullet)$. The topic of negative transfer addresses the need to quantify the amount of relatedness between the source domain and the target domain and whether an attempt to transfer knowledge from the source domain should be made. Extending the definition above, positive transfer occurs when the performance of $f_{T2}(\bullet)$ is greater than the performance of $f_{T1}(\bullet)$.

Throughout the literature on transfer learning, there are a number of terminology inconsistencies. Phrases such as transfer learning and domain adaptation are used to refer to similar processes. The following definitions will be used in this paper. Domain adaptation, as it pertains to transfer learning, is the process of adapting one or more source domains for the means of transferring information to improve the performance of a target learner. The domain adaptation process attempts to alter a source domain in an attempt to bring the distribution of the source closer to that of the target. Another area of literature inconsistencies is in characterizing the transfer

learning process with respect to the availability of labeled and unlabeled data. For example, Daumé [14] and Chattopadhyay et al. [15] define supervised transfer learning as the case of having abundant labeled source data and limited labeled target data, and semi-supervised transfer learning as the case of abundant labeled source data and no labeled target data. In Gong et al. [16] and Blitzer et al. [17], semi-supervised transfer learning is the case of having abundant labeled source data and limited labeled target data, and unsupervised transfer learning is the case of abundant labeled source data and no labeled target data. Cook et al. [18] and Feuz and Cook [19] provide a different variation where the definition of supervised or unsupervised refers to the presence or absence of labeled data in the source domain and informed or uninformed refers to the presence or absence of labeled data in the target domain. With this definition, a labeled source and limited labeled target domain is referred to as informed supervised transfer learning. Pan and Yang [3] refers to inductive transfer learning as the case of having available labeled target domain data, transductive transfer learning as the case of having labeled source and no labeled target domain data, and unsupervised transfer learning as the case of having no labeled source and no labeled target domain data. This paper will explicitly state when labeled and unlabeled data are being used in the source and target domains.

There are different strategies and implementations for solving a transfer learning problem. The majority of the homogeneous transfer learning solutions employ one of three general strategies which include trying to correct for the marginal distribution difference in the source, trying to correct for the conditional distribution difference in the source, or trying to correct both the marginal and conditional distribution differences in the source. The majority of the heterogeneous transfer learning solutions are focused on aligning the input spaces of the source and target domains with the assumption that the domain distributions are the same. If the domain distributions are not equal, then further domain adaptation steps are needed. Another important aspect of a transfer learning solution is the form of information transfer (or what is being transferred). The form of information transfer is categorized into four general Transfer Categories [3]. The first Transfer Category is transfer learning through instances. A common method used in this case is for instances from the source domain to be reweighted in an attempt to correct for marginal distribution differences. These reweighted instances are then directly used in the target domain for training (examples in [20, 21]). These reweighting algorithms work best when the conditional distribution is the same in both domains. The second Transfer Category is transfer learning through features. Feature-based transfer learning approaches are categorized in two ways. The first approach transforms the features of the source through reweighting to more closely match the target domain (e.g. Pan et al. [22]). This is referred to as asymmetric feature transformation and is depicted in Fig. 3.1b. The second approach discovers underlying meaningful structures between the domains to find a common latent feature space that has predictive qualities while reducing the marginal distribution between the domains (e.g. Blitzer et al. [17]). This is referred to as symmetric feature transformation and is depicted in Fig. 3.1a. The third Transfer Category is to

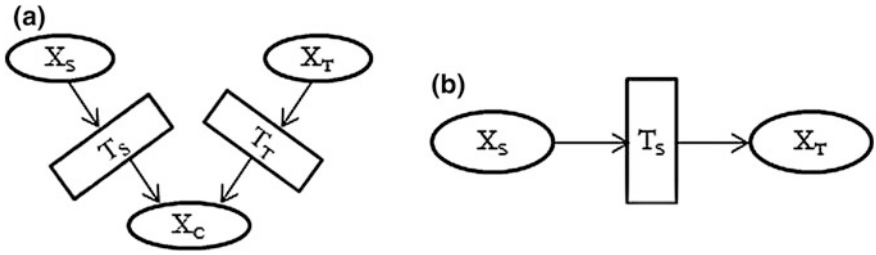

Fig. 3.1 **a** Shows the symmetric transformation mapping (T_S and T_T) of the source (X_S) and target (X_T) domains into a common latent feature space. **b** Shows the asymmetric transformation (T_T) of the source domain (X_S) to the target domain (X_T)

transfer knowledge through shared parameters of source and target domain learner models or by creating multiple source learner models and optimally combining the reweighted learners (ensemble learners) to form an improved target learner (examples in [23, 24, 25]). The last Transfer Category (and the least used approach) is to transfer knowledge based on some defined relationship between the source and target domains (examples in [26, 27].

Detailed information on specific transfer learning solutions are presented in ``Homogeneous Transfer Learning'', ``Heterogeneous Transfer Learning'', and ``Negative Transfer'' sections. These sections represent the majority of the works surveyed in this paper. In ``Homogeneous Transfer Learning'', ``Heterogeneous Transfer Learning'', and ``Negative Transfer'' sections cover homogeneous transfer learning solutions, heterogeneous transfer learning solutions, and solutions addressing negative transfer, respectively. The section covering transfer learning applications focuses on the general applications that transfer learning is applied to, but does not describe the solution details.

Homogeneous Transfer Learning

This section presents surveyed papers covering homogeneous transfer learning solutions and is divided into subsections that correspond to the Transfer Categories of instance-based, feature-based (both asymmetric and symmetric), parameter-based, and relational-based. Recall that homogeneous transfer learning is the case where $X_S = X_T$. The algorithms surveyed are summarized in Table 3.2 at the end of this section.

The methodology of homogeneous transfer learning is directly applicable to a big data environment. As repositories of big data become more available, there is a desire to use this abundant resource for machine learning tasks, avoiding the timely and potentially costly collection of new data. If there is an available dataset that is drawn from a domain that is related to, but does not an exactly match a target domain of interest, then homogeneous transfer learning can be used to build a predictive model for the target domain as long as the input feature space is the same.

Table 3.2 Homogeneous transfer learning approaches surveyed in Sect. 3 listing different characteristics of each approach

Approach	Transfer category	Source data	Target data	Multiple sources	Generic solution	Negative transfer
CP-MDA [15]	Parameter	Labeled	Limited labels	✓	✓	
2SW-MDA [15]	Instance	Labeled	Unlabeled	✓	✓	
FAM [14]	Asymmetric feature	Labeled	Limited labels	✓	✓	
DTMKL [28]	Asymmetric feature	Labeled	Unlabeled		✓	
JDA [29]	Asymmetric feature	Labeled	Unlabeled		✓	
ARTL [30]	Asymmetric feature	Labeled	Unlabeled		✓	
TCA [31]	Symmetric feature	Labeled	Unlabeled		✓	
SFA [32]	Symmetric feature	Labeled	Limited labels	✓	✓	
SDA [33]	Symmetric feature	Labeled	Unlabeled		✓	
GFK [16]	Symmetric feature	Labeled	Unlabeled		✓	✓
DCP [34]	Symmetric feature	Labeled	Unlabeled		✓	
TCNN [35]	Symmetric feature	Labeled	Limited labels		✓	
MMKT [36]	Parameter	Labeled	Limited labels	✓	✓	✓
DSM [37]	Parameter	Labeled	Unlabeled	✓		✓
MsTrAdaBoost [38]	Instance	Labeled	Limited labels	✓	✓	✓
TaskTrAdaBoost [38]	Parameter	Labeled	Limited labels	✓	✓	✓
RAP [27]	Relational	Labeled	Unlabeled			
SSFE [39]	Hybrid (instance and feature)	Labeled	Limited labels			

Instance-Based Transfer Learning

The paper by Chattopadhyay et al. [15] proposes two separate solutions both using multiple labeled source domains. The first solution is the Conditional Probability based Multi-source Domain Adaptation (CP-MDA) approach, which is a domain adaptation process based on correcting the conditional distribution differences

between the source and target domains. The CP-MDA approach assumes a limited amount of labeled target data is available. The main idea is to use a combination of source domain classifiers to label the unlabeled target data. This is accomplished by first building a classifier for each separate source domain. Then a weight value is found for each classifier as a function of the closeness in conditional distribution between each source and the target domain. The weighted source classifiers are summed together to create a learning task that will find the pseudo labels (estimated labels later used for training) for the unlabeled target data. Finally, the target learner is built from the labeled and pseudo labeled target data. The second proposed solution is the Two Stage Weighting framework for Multi-source Domain Adaptation (2SW-MDA) which addresses both marginal and conditional distribution differences between the source and target domains. Labeled target data is not required for the 2SW-MDA approach; however, it can be used if available. In this approach, a weight for each source domain is computed based on the marginal distribution differences between the source and target domains. In the second step, the source domain weights are modified as a function of the difference in the conditional distribution as performed in the CP-MDA approach previously described. Finally, a target classifier is learned based on the reweighted source instances and any labeled target instances that are available. The work presented in Chattopadhyay et al. [15] is an extension of Duan et al. [40] where the novelty is in calculating the source weights as a function of conditional probability. Note, the 2SW-MDA approach is an example of an instance-based Transfer Category, but the CP-MDA approach is more appropriately classified as a parameter-based Transfer Category (see ``Heterogeneous Transfer Learning'' section). Experiments are performed for muscle fatigue classification using surface electromyography data where classification accuracy is measured as the performance metric. Each source domain represents one person's surface electromyography measurements. A baseline approach is constructed using a Support Vector Machine (SVM) classifier trained on the combination of seven sources used for this test. The transfer learning approaches that are tested against include an approach proposed by Huang et al. [20], Pan et al. [31], Zhong et al. [41], Gao et al. [23], and Duan et al. [40]. The order of performance from best to worst is 2SW-MDA, CP-MDA, Duan et al. [40], Zhong et al. [41], Gao et al. [23], Pan et al. [31], Huang et al. [20], and the baseline approach. All the transfer learning approaches performed better than the baseline approach.

Asymmetric Feature-Based Transfer Learning

In an early and often cited work, Daumé [14] proposes a simple domain adaptation algorithm, referred to as the Feature Augmentation Method (FAM), requiring only 10 lines of Perl script that uses labeled source data and limited labeled target data. In a transfer learning environment, there are scenarios where a feature in the source domain may have a different meaning in the target domain. The issue is referred to as context feature bias, which causes the conditional distributions between the

source and target domains to be different. To resolve context feature bias, a method to augment the source and target feature space with three duplicate copies of the original feature set is proposed. More specifically, the three duplicate copies of the original feature set in the augmented source feature space represent a common feature set, a source specific feature set, and a target specific feature set which is always set to zero. In a similar way, the three duplicate copies of the original feature set in the augmented target feature space represent a common feature set, a source specific feature set which is always set to zero, and a target specific feature set. By performing this feature augmentation, the feature space is duplicated three times. From the feature augmentation structure, a classifier learns the individual feature weights for the augmented feature set, which will help correct for any feature bias issues. Using a text document example where features are modeled as a bag-of-words, a common word like "the" would be assigned (through the learning process) a high weight for the common feature set, and a word that is different between the source and target like "monitor" would be assigned a high weight for the corresponding domain feature set. The duplication of features creates feature separation between the source and target domains, and allows the final classifier to learn the optimal feature weights. For the experiments, a number of different natural language processing applications are tested and in each case the classification error rate is measured as the performance metric. An SVM learner is used to implement the Daumé [14] approach. A number of baseline approaches with no transfer learning techniques are measured along with a method by Chelba and Acero [42]. The test results show the Daumé [14] method is able to outperform the other methods tested. However, when the source and target domains are very similar, the Daumé [14] approach tends to underperform. The reason for the underperformance is the duplication of feature sets represents irrelevant and noisy information when the source and target domains are very similar.

Multiple kernel learning is a technique used in traditional machine learning algorithms as demonstrated in the works of Wu et al. [43] and Vedaldi et al. [44]. Multiple kernel learning allows for an optimal kernel function to be learned in a computationally efficient manner. The paper by Duan et al. [28] proposes to implement a multiple kernel learning framework for a transfer learning environment called the Domain Transfer Multiple Kernel Learning (DTMKL). Instead of learning one kernel, multiple kernel learning assumes the kernel is comprised of a linear combination of multiple predefined base kernels. The final classifier and the kernel function are learned simultaneously which has the advantage of using labeled data during the kernel learning process. This is an improvement over Pan et al. [22] and Huang et al. [20] where a two-stage approach is used. The final classifier learning process minimizes the structural risk functional [45] and the marginal distribution between domains using the Maximum Mean Discrepancy measure [46]. Pseudo labels are found for the unlabeled target data to take advantage of this information during the learning process. The pseudo labels are found as a weighted combination of base classifiers (one for each feature) trained from the labeled source data. A regularization term is added to the optimization problem to ensure the predicted values from the final target classifier and the base

classifiers are similar for the unlabeled target data. Experiments are performed on the applications of video concept detection, text classification, and email spam detection. The methods tested against include a baseline approach using an SVM classifier trained on the labeled source data, the feature replication method from Daumé [14], an adaptive SVM method from Yang et al. [47], a cross-domain SVM method proposed by Jiang et al. [48], and a kernel mean matching method by Huang et al. [20]. The DTMKL approach uses an SVM learner for the experiments. Average precision and classification accuracy are measured as the performance metrics. The DTMKL method performed the best for all applications, and the baseline approach is consistently the worst performing. The other methods showed better performance over the baseline which demonstrated a positive transfer learning effect.

The work by Long et al. [29] is a Joint Domain Adaptation (JDA) solution that aims to simultaneously correct for the marginal and conditional distribution differences between the labeled source domain and the unlabeled target domain. Principal Component Analysis (PCA) is used for optimization and dimensionality reduction. To address the difference in marginal distribution between the domains, the Maximum Mean Discrepancy distance measure [46] is used to compute the marginal distribution differences and is integrated into the PCA optimization algorithm. The next part of the solution requires a process to correct the conditional distribution differences, which requires labeled target data. Since the target data is unlabeled, pseudo labels (estimated target labels) are found by learning a classifier from the labeled source data. The Maximum Mean Discrepancy distance measure is modified to measure the distance between the conditional distributions and is integrated into the PCA optimization algorithm to minimize the conditional distributions. Finally, the features identified by the modified PCA algorithm are used to train the final target classifier. Experiments are performed for the application of image recognition and classification accuracy is measured as the performance metric. Two baseline approaches of a 1-nearest neighbor classifier and a PCA approach trained on the source data are tested. Transfer learning approaches tested for this experiment include the approach by Pan [31], Gong et al. [16], and Si et al. [49]. These transfer learning approaches only attempt to correct for marginal distribution differences between domains. The Long et al. [29] approach is the best performing, followed by the Pan [31] and Si et al. [49] approaches (a tie), then the Gong et al. [16] approach, and finally the baseline approaches. All transfer learning approaches perform better than the baseline approaches. The possible reason behind the underperformance of the Gong et al. [16] approach is the data smoothness assumption that is made for the Gong et al. [16] solution may not be intact for the data sets tested.

The paper by Long et al. [30] proposes an Adaptation Regularization based Transfer Learning (ARTL) framework for scenarios of labeled source data and unlabeled target data. This transfer learning framework proposes to correct the difference in marginal distribution between the source and target domains, correct the difference in conditional distribution between the domains, and improve classification performance through a manifold regularization [50] process (which

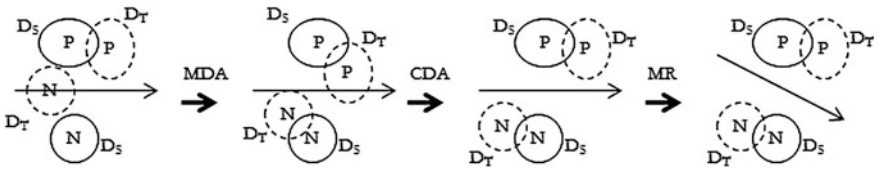

Fig. 3.2 ARTL overview showing marginal distribution adaptation (MDA), conditional distribution adaptation (CDA), and manifold regularization (MR). Diagram adapted from Long [30]

optimally shifts the hyperplane of an SVM learner). This complete framework process is depicted in Fig. 3.2. The proposed ARTL framework will learn a classifier by simultaneously performing structural risk minimization [45], reducing the marginal and conditional distributions between the domains, and optimizing the manifold consistency of the marginal distribution. To resolve the conditional distribution differences, pseudo labels are found for the target data in the same way as proposed by Long et al. [29]. A difference between the ARTL approach and Long et al. [29] is ARTL learns the final classifier simultaneously while minimizing the domain distribution differences, which is claimed by Long et al. [30] to be a more optimal solution. Unfortunately, the solution by Long et al. [29] is not included in the experiments. Experiments are performed on the applications of text classification and image classification where classification accuracy is measured as the performance metric. There are three baseline methods tested where different classifiers are trained with the labeled source data. There are five transfer learning methods tested against, which include methods by Ling et al. [51], Pan et al. [32], Pan et al. [31], Quanz and Huan [52], and Xiao and Guo [53]. The order of performance from best to worst is ARTL, Xiao and Guo [53], Pan et al. [31], Pan et al. [32], Quanz and Huan [52] and Ling et al. [51] (tie), and the baseline approaches. The baseline methods underperformed all other transfer learning approaches tested.

Symmetric Feature-Based Transfer Learning

The paper by Pan et al. [31] proposes a feature transformation approach for domain adaptation called Transfer Component Analysis (TCA), which does not require labeled target data. The goal is to discover common latent features that have the same marginal distribution across the source and target domains while maintaining the intrinsic structure of the original domain data. The latent features are learned between the source and target domains in a Reproducing Kernel Hilbert Space [54] using the Maximum Mean Discrepancy [46] as a marginal distribution measurement criteria. Once the latent features are found, traditional machine learning is used to train the final target classifier. The TCA approach extends the work of Pan et al. [22] by improving computational efficiency. Experiments are conducted for

the application of WiFi localization where the location of a particular device is being predicted. The source domain is comprised of data measured from different room and building topologies. The performance metric measured is the average error distance of the position of a device. The transfer learning methods tested against are from Blitzer et al. [17] and Huang et al. [20]. The TCA method performed the best followed by the Huang et al. [20] approach and the Blitzer et al. [17] approach. For the Blitzer et al. [17] approach, the manual definition of the pivot functions (functions that define the correspondence) is important to performance and specific to the end application. There is no mention as to how the pivot functions are defined for WiFi localization.

The work by Pan et al. [32] proposes a Spectral Feature Alignment (SFA) transfer learning algorithm that discovers a new feature representation for the source and target domain to resolve the marginal distribution differences. The SFA method assumes an abundance of labeled source data and a limited amount of labeled target data. The SFA approach identifies domain-specific and domain-independent features and uses the domain-independent features as a bridge to build a bipartite graph modeling the co-occurrence relationship between the domain-independent and domain-specific features. If the graph shows two domain-specific features having connections to common domain-independent feature, then there is a higher chance the domain-specific features are aligned. A spectral clustering algorithm based on graph spectral theory [55] is used on the bipartite graph to align domain-specific features and domain-independent features into a set of clusters representing new features. These clusters are used to reduce the difference between domain-specific features in the source and the target domains. All the data instances are projected into this new feature space and a final target classifier is trained using the new feature representation. The SFA algorithm is a type of correspondence learning where the domain-independent features act as pivot features (see Blitzer et al. [17] and Prettenhofer and Stein [11] for further information on correspondence learning). The SFA method is well-suited for the application of text document classification where a bag-of-words model is used to define features. For this application there are domain-independent words that will appear often in both domains and domain-specific words that will appear often only in a specific domain. This is referred to as frequency feature bias, which causes marginal distribution differences between the domains. An example of domain-specific features being combined is the word “sharp” appearing often in the source domain but not in the target domain, and the word “hooked” appearing often in the target but not in the source domain. These words are both connected to the same domain-independent words (for example “good” and “exciting”). Further, when the words “sharp” or “hooked” appear in text instances, the labels are the same. The idea is to combine (or align) these two features (in this case “sharp” and “hooked”) to form a new single invariant feature. The experiments are performed on sentiment classification where classification accuracy is measured as the performance metric. A baseline approach is tested where a classifier is trained only on source data. An upper limit approach is also tested where a classifier is trained on a large amount of labeled target data. The competing transfer learning approach tested against is by Blitzer et al. [17]. The order of performance

for the tests from best to worst is the upper limit approach, SFA, Blitzer et al. [17], and baseline approach. Not only does the SFA approach demonstrate better performance than Blitzer et al. [17], the SFA approach does not need to manually define pivot functions as in the Blitzer et al. [17] approach. The SFA approach only addresses the issue of marginal distribution differences and does not address any context feature bias issues, which would represent conditional distribution differences.

The work by Glorot et al. [33] proposes a deep learning algorithm for transfer learning called a Stacked Denoising Autoencoder (SDA) to resolve the marginal distribution differences between a labeled source domain and an unlabeled target domain. Deep learning algorithms learn intermediate invariant concepts between two data sources, which are used to find a common latent feature set. The first step in this process is to train the Stacked Denoising Autoencoders [56] with unlabeled data from the source and target domains. This transforms the input space to discover the common invariant latent feature space. The next step is to train a classifier using the transformed latent features with the labeled source data. Experiments are performed on text review sentiment classification where transfer loss is measured as the performance metric. Transfer loss is defined as the classification error rate using a learner only trained on the source domain and tested on the target minus the classification error rate using a learner only trained on the target domain and tested on the target. There are 12 different source and target domain pairs that are created from four unique review topics. A baseline method is tested where an SVM classifier is trained on the source domain. The transfer learning approaches that are tested include an approach by Blitzer et al. [17], Li and Zong [57], and Pan et al. [32]. The Glorot et al. [33] approach performed the best with the Blitzer et al. [17], Li and Zong [57], and Pan et al. [32] methods all having similar performance and all outperforming the baseline approach.

In the paper by Gong et al. [16], a domain adaptation technique called the Geodesic Flow Kernel (GFK) is proposed that finds a low-dimensional feature space, which reduces the marginal distribution differences between the labeled source and unlabeled target domains. To accomplish this, a geodesic flow kernel is constructed using the source and target input feature data, which projects a large number of subspaces that lie on the geodesic flow curve. The geodesic flow curve represents incremental differences in geometric and statistical properties between the source and target domain spaces. A classifier is then learned from the geodesic flow kernel by selecting the features from the geodesic flow curve that are domain invariant. The work of Gong et al. [16] directly enhances the work of Gopalan et al. [58] by eliminating tuning parameters and improving computational efficiency. In addition, a Rank of Domain (ROD) metric is developed to evaluate which of many source domains is the best match for the target domain. The ROD metric is a function of the geometric alignment between the domains and the Kullback–Leibler divergence in data distributions between the projected source and target subspaces. Experiments are performed for the application of image classification where classification accuracy is measured as the performance metric. The tests use pairs of source and target data sets from four available data sets. A baseline approach is

defined that does not use transfer learning, along with the approach defined by Gopalan et al. [58]. Additionally, the Gong et al. [16] approach uses a 1-nearest neighbor classifier. The results in order from best to worst performance are Gong et al. [16], Gopalan et al. [58], and the baseline approach. The ROD measurements between the different source and target domain pairs tested have a high correlation to the actual test results, meaning the domains that are found to be more related with respect to the ROD measurement had higher classification accuracies.

The solution by Shi and Sha [34], referred to as the Discriminative Clustering Process (DCP), proposes to equalize the marginal distribution of the labeled source and unlabeled target domains. A discriminative clustering process is used to discover a common latent feature space that is domain invariant while simultaneously learning the final target classifier. The motivating assumptions for this solution are the data in both domains form well-defined clusters which correspond to unique class labels, and the clusters from the source domain are geometrically close to the target clusters if they share the same label. Through clustering, the source domain labels can be used to estimate the target labels. A one-stage solution is formulated that minimizes the marginal distribution differences while minimizing the predicted classification error in the target domain using a nearest neighbor classifier. Experiments are performed for object recognition and sentiment classification where classification accuracy is measured as the performance metric. The approach described above is tested against a baseline approach taken from Weinberger and Saul [59] with no transfer learning. Other transfer learning approaches tested include an approach from Pan et al. et al. [31], Blitzer et al. [17], and Gopalan et al. [58]. The Blitzer et al. [17] approach is not tested for the object recognition application because the pivot functions are not easily defined for this application. For the object recognition tests, the Shi and Sha [34] method is best in five out of six comparison tests. For the text classification tests, the Shi and Sha [34] approach is the best performing overall, with the Blitzer et al. [17] approach a close second. An important point to note is the baseline method outperformed the Pan et al. [31] and Gopalan et al. [58] methods in both tests. Both the Pan et al. [31] and Gopalan et al. [58] methods are two-stage domain adaptation processes where the first stage reduces the marginal distributions between the domains and the second stage trains a classifier with the adapted domain data. This paper offers a hypothesis that two-stage processes are actually detrimental to transfer learning (causes negative transfer). The one-stage learning process is a novel idea presented by this paper. The hypothesis that the two-stage transfer learning process creates low performing learners does not agree with the results presented in the individual papers by Gopalan et al. [58] and Pan et al. [31] and other previously surveyed works.

Convolutional Neural Networks (CNN) have been successfully used in traditional data mining environments [60]. However, a CNN requires a large amount of labeled training data to be effective, which may not be available. The paper by Oquab et al. [35] proposes a transfer learning method of training a CNN with available labeled source data (a source learner) and then extracting the CNN internal layers (which represent a generic mid-level feature representation) to a target CNN learner. This method is referred to as the Transfer Convolutional Neural

Network (TCNN). To correct for any further distribution differences between the source and the target domains, an adaptation layer is added to the target CNN learner, which is trained from the limited labeled target data. The experiments are run on the application of object image classification where average precision is measured as the performance metric. The Oquab et al. [35] method is tested against a method proposed by Marszalek et al. [61] and a method proposed by Song et al. [62]. Both the Marszalek et al. [61] and Song et al. [62] approaches are not transfer learning approaches and are trained on the limited labeled target data. The first experiment is performed using the Pascal VOC 2007 data set as the target and ImageNet 2012 as the source. The Oquab et al. [35] method outperformed both Song et al. [62] and Marszalek et al. [61] approaches for this test. The second experiment is performed using the Pascal VOC 2012 data set as the target and ImageNet 2012 as the source. In the second test, the Oquab et al. [35] method marginally outperformed the Song et al. [62] method (the Marszalek et al. [61] method was not tested for the second test). The tests successfully demonstrated the ability to transfer information from one CNN learner to another.

Parameter-Based Transfer Learning

The paper by Tommasi et al. [36] addresses the transfer learning environment characterized by limited labeled target data and multiple labeled source domains where each source corresponds to a particular class. In this case, each source is able to build a binary learner to predict that class. The objective is to build a target binary learner for a new class using minimal labeled target data and knowledge transferred from the multiple source learners. An algorithm is proposed to transfer the SVM hyperplane information of each of the source learners to the new target learner. To minimize the effects of negative transfer, the information transferred from each source to the target will be weighted such that the most related source domains receive the highest weighting. The weights are determined through a leave out one process as defined by Cawley [63]. The Tommasi et al. [36] approach, called the Multi-Model Knowledge Transfer (MMKT) method, extends the method proposed by Tommasi and Caputo [64] that only transfers a single source domain. Experiments are performed on the application of image recognition where classification accuracy is measured as the performance metric. Transfer learning methods tested include an average weight approach (same as Tommasi et al. [36] but all source weights are equal), and the Tommasi and Caputo [64] approach. A baseline approach is tested, which is trained on the limited labeled target data. The best performing method is Tommasi et al. [36], followed by the average weight, Tommasi and Caputo [64], and the baseline approach. As the number of labeled target instances goes up, the Tommasi et al. [36] and average weight methods converge to the same performance. This is because the adverse effects of negative transfer are lessened as the labeled target data increases. This result demonstrates

the Tommasi et al. [36] approach is able to lessen the effects of negative transfer from unrelated sources.

The transfer learning approach presented in the paper by Duan et al. [37], referred to as the Domain Selection Machine (DSM), is tightly coupled to the application of event recognition in consumer videos. Event recognition in videos is the process of predicting the occurrence of a particular event or topic (e.g. "show" or "performance") in a given video. In this scenario, the target domain is unlabeled and the source information is obtained from annotated images found via web searches. For example, a text query of the event "show" for images on Photosig.com represents one source and the same query on Flickr.com represents another separate source. The Domain Selection Machine proposed in this paper is realized as follows. For each individual source, an SVM classifier is created using SIFT [65] image features. The final target classifier is made up of two parts. The first part is a weighted sum of the source classifier outputs whose input is the SIFT features from key frames of the input video. The second part is a learning function whose inputs are Space-Time features [66] from the input video and is trained from target data where the target labels are estimated (pseudo labels) from the weighted sum of the source classifiers. To combat the effects of negative transfer from unrelated sources, the most relevant source domains are selected by using an alternating optimization algorithm that iteratively solves the target decision function and the domain selection vector. Experiments are performed in the application of event recognition in videos as described above where the mean average precision is measured as the performance metric. A baseline method is created by training a separate SVM classifier on each source domain and then equally combining the classifiers. The other transfer learning approaches tested include the approach by Bruzzone and Marconcini [67], Schweikert et al. [68], Duan et al. [40], and Chattopadhyay et al. [15]. The Duan et al. [40] approach outperforms all the other approaches tested. The other approaches all have similar results, meaning the transfer learning methods did not outperform the baseline approach. The possible reason for this result is the existence of unrelated sources in the experiment. The other transfer learning approaches tested had no mechanism to guard against negative transfer from unrelated sources.

The paper by Yao and Doretto [38] first presents an instance-based transfer learning approach followed by a separate parameter-based transfer learning approach. In the transfer learning process, if the source and target domains are not related enough, negative transfer can occur. Since it is difficult to measure the relatedness between any particular source and target domain, Yao and Doretto [38] proposes to transfer knowledge from multiple source domains using a boosting method in an attempt to minimize the effects of negative transfer from a single unrelated source domain. The boosting process requires some amount of labeled target data. Yao and Doretto [38] effectively extends the work of Dai et al. [69] (TrAdaBoost) by expanding the transfer boosting algorithm to multiple source domains. In the TrAdaBoost algorithm, during every boosting iteration, a so-called weak classifier is built using weighted instance data from the previous iteration. Then, the misclassified source instances are lowered in importance and the

misclassified target instances are raised in importance. In the multi-source TrAdaBoost algorithm (called MsTrAdaBoost), each iteration step first finds a weak classifier for each source and target combination, and then the final weak classifier is selected for that iteration by finding the one that minimizes the target classification error. The instance reweighting step remains the same as in the TrAdaBoost. An alternative multi-source boosting method (TaskTrAdaBoost) is proposed that transfers internal learner parameter information from the source to the target. The TaskTrAdaBoost algorithm first finds candidate weak classifiers from each individual source by performing an AdaBoost process on each source domain. Then an AdaBoost process is performed on the labeled target data, and at every boosting iteration, the weak classifier used is selected from the candidate weak source classifiers (found in the previous step) that has the lowest classification error using the labeled target data. Experiments are performed for the application of object category recognition where the area under the curve (AUC) is measured as the performance metric. An AdaBoost baseline approach using only the limited labeled target data is measured along with a TrAdaBoost approach using a single source (the multiple sources are combined to one) and the limited labeled target data. Linear SVM learners are used as the base classifiers in all approaches. Both the MsTrAdaBoost and TaskTrAdaBoost approaches outperform the baseline approach and TrAdaBoost approach. The MsTrAdaBoost and TaskTrAdaBoost demonstrated similar performance.

Relational-Based Transfer Learning

The specific application addressed in the paper by Li et al. [27] is to classify words from a text document into one of three classes (e.g. sentiments, topics, or neither). In this scenario, there exists a labeled text source domain on one particular subject matter and an unlabeled text target domain on a different subject matter. The main idea is that sentiment words remain constant between the source and target domains. By learning the grammatical and sentence structure patterns of the source, a relational pattern is found between the source and target domains, which is used to predict the topic words in the target. The sentiment words act as a common linkage or bridge between the source and target domains. A bipartite word graph is used to represent and score the sentence structure patterns. A bootstrapping algorithm is used to iteratively build a target classifier from the two domains. The bootstrapping process starts with defining seeds which are instances from the source that match frequent patterns in the target. A cross domain classifier is then trained with the seed information and extracted target information (there is no target information in the first iteration). The classifier is used to predict the target labels and the top confidence rated target instances are selected to reconstruct the bipartite word graph. The bipartite word graph is now used to select new target instances that are added to the seed list. This bootstrapping process continues over a selected number of iterations, and the cross domain classifier learned in the bootstrapping process is now available

to predict target samples. This method is referred to as the Relational Adaptive bootstraPping (RAP) approach. The experiments tested the Li et al. [27] approach against an upper bound method where a standard classifier is trained with a large amount of target data. Other transfer learning methods tested include an approach by Hu and Liu [70], Qiu et al. [71], Jakob and Gurevych [72], and Dai et al. [69]. The application tested is word classification as described above where the F1 score is measured as the performance metric. The two domains tested are related to movie reviews and product reviews. The Li et al. [27] method performed better than the other transfer learning methods, but fell short of the upper bound method as expected. In its current form, this algorithm is tightly coupled with its underlying text application, which makes it difficult to use for other non-text applications.

Hybrid-Based (Instance and Parameter) Transfer Learning

The paper by Xia et al. [39] proposes a two step approach to address marginal distribution differences and conditional distribution differences between the source and target domains called the Sample Selection and Feature Ensemble (SSFE) method. A sample selection process, using a modified version of Principal Component Analysis, is employed to select labeled source domain samples such that the source and target marginal distributions are equalized. Next, a feature ensemble step attempts to resolve the conditional distribution differences between the source and target domains. Four individual classifiers are defined corresponding to parts of speech of noun, verb, adverb/adjective, and other. The four classifiers are trained using only the features that correspond to that part of speech. The training data is the limited labeled target and the labeled source selected in the previous sample selection step. The four classifiers are weighted as a function of minimizing the classification error using the limited labeled target data. The weighted output of the four classifiers is used as the final target classifier. This work by Xia et al. [39] extends the earlier work of Xia and Zong [73]. The experiments are performed for the application of review sentiment classification using four different review categories, where each category is combined to create 12 different source and target pairs. Classification accuracy is measured as the performance metric. A baseline approach using all the training data from the source is constructed, along with a sample selection approach (only using the first step defined above), a feature ensemble approach (only using the second step defined above) and the complete approach outlined above. The complete approach is the best performing, followed by sample selection and feature ensemble approaches, and the baseline approach. The sample selection and feature ensemble approaches perform equally as well in head-to-head tests. The weighting of the four classifiers (defined by the corresponding parts of speech) in the procedure above gives limited resolution in attempting to adjust for context feature bias issues. A method of having more classifiers in the ensemble step could yield better performance at the expense of higher complexity.

Discussion of Homogeneous Transfer Learning

The previous surveyed homogeneous transfer learning works (summarized in Table 3.2) demonstrate many different characteristics and attributes. Which homogeneous transfer learning solution is best for a particular application? An important characteristic to evaluate in the selection process is what type of differences exist between a given source and target domain. The previous solutions surveyed address domain adaptation by correcting for marginal distribution differences, correcting for conditional distribution differences, or correcting for both marginal and conditional distribution differences. The surveyed works of Duan et al. [28], Gong et al. [16], Pan et al. [31], Li et al. [27], Shi and Sha [34], Oquab et al. [35], Glorot et al. [33], and Pan et al. [32] are focused on solving the differences in marginal distribution between the source and target domains. The surveyed works of Daumé [14], Yao and Doretto [38], Tommasi et al. [36] are focused on solving the differences in conditional distribution between the source and target domains. Lastly, the surveyed works of Long et al. [30], Xia et al. [39], Chattopadhyay et al. [15], Duan et al. [37], and Long et al. [29] correct the differences in both the marginal and conditional distributions. Correcting for the conditional distribution differences between the source and target domain can be problematic as the nature of a transfer learning environment is to have minimal labeled target data. To compensate for the limited labeled target data, many of the recent transfer learning solutions create pseudo labels for the unlabeled target data to facilitate the conditional distribution correction process between the source and target domains. To further help determine which solution is best for a given transfer learning application, the information in Table 3.2 should be used to match the characteristics of the solution to that of the desired application environment. If the application domain contains multiple sources where the sources are not mutually uniformly distributed, a solution that guards against negative transfer may be of greater benefit. A recent trend in the development of transfer learning solutions is for solutions to address both marginal and conditional distribution differences between the source and target domains. Another emerging solution trend is the implementation of a one-stage process as compared to a two-stage process. In the recent works of Long et al. [30], Duan et al. [28], Shi and Sha [34], and Xia et al. [39], a one-stage process is employed that simultaneously performs the domain adaptation process while learning the final classifier. A two-stage solution first performs the domain adaptation process and then independently learns the final classifier. The claim by Long et al. [30] is a one-stage solution achieves enhanced performance because the simultaneous solving of domain adaptation and the classifier establishes mutual reinforcement. The surveyed homogeneous transfer learning works are not specifically applied to big data solutions; however, there is nothing to preclude their use in a big data environment.

Heterogeneous Transfer Learning

Heterogeneous transfer learning is the scenario where the source and target domains are represented in different feature spaces. There are many applications where heterogeneous transfer learning is beneficial. Heterogeneous transfer learning applications that are covered in this section include image recognition [5–7, 74–76], multi-language text classification [5, 10–12, 76], single language text classification [4], drug efficacy classification [74], human activity classification [8], and software defect classification [9]. Heterogeneous transfer learning is also directly applicable to a big data environment. As repositories of big data become more available, there is a desire to use this abundant resource for machine learning tasks, avoiding the timely and potentially costly collection of new data. If there is an available dataset drawn from a target domain of interest that has a different feature space from another target dataset (also drawn from the same target domain), then heterogeneous transfer learning can be used to bridge the difference in the feature spaces and build a predictive model for that target domain. Heterogeneous transfer learning is still a relatively new area of study as the majority of the works covering this topic have been published in the last five years. From a high-level view, there are two main approaches to solving the heterogeneous feature space difference. The first approach, referred to as symmetric transformation shown in Fig. 3.1a, separately transforms the source and target domains into a common latent feature space in an attempt to unify the input spaces of the domains. The second approach, referred to as asymmetric transformation as shown in Fig. 3.1b, transforms the source feature space to the target feature space to align the input feature spaces. The asymmetrical transformation approach is best used when the same class instances in the source and target can be transformed without context feature bias. Many of the heterogeneous transfer learning solutions surveyed make the implicit or explicit assumption that the source and the target domain instances are drawn from the same domain space. With this assumption there should be no significant distribution differences between the domains. Therefore, once the differences in input feature spaces are resolved, no further domain adaptation needs to be performed.

As is the case with homogeneous transfer learning solutions, whether the source and target domains contain labeled data drives the solution formulation for heterogeneous approaches. Data label availability is a function of the underlying application. The solutions surveyed in this paper have different labeled data requirements. For transfer learning to be feasible, the source and the target domains must be related in some way. Some heterogeneous solutions require an explicit mapping of the relationship or correspondence between the source and target domains. For example, the solutions defined for Prettenhofer and Stein [11] and Wei and Pal [77] require manual definitions of source and target correspondence.

Symmetric Feature-Based Transfer Learning

The transfer learning approach proposed by Prettenhofer and Stein [11] addresses the heterogeneous scenario of a source domain containing labeled and unlabeled data, and a target domain containing unlabeled data. The structural correspondence learning technique from Blitzer et al. [17] is applied to this problem. Structural correspondence learning depends on the manual definition of pivot functions that capture correspondence between the source and target domains. Effective pivot functions should use features that occur frequently in both domains and have good predictive qualities. Each pivot function is turned into a linear classifier using data from the source and target domains. From these pivot classifiers, correspondences between features are discovered and a latent feature space is learned. The latent feature space is used to train the final target classifier. The paper by Prettenhofer and Stein [11] uses this solution to solve the problem of text classification where the source is written in one language and the target is written in a different language. In this specific implementation referred to as Cross-Language Structural Correspondence Learning (CLSCL), the pivot functions are defined by pairs of words, one from the target and one from the source, that represent direct word translations from one language to the other. The experiments are performed on the applications of document sentiment classification and document topic classification. English documents are used in the source and other language documents are used in the target. The baseline method used in this test trains a learner on the labeled source documents, then translates the target documents to the source language and tests the translated version. An upper bound method is established by training a learner with the labeled target documents and testing with the target documents. Average classification accuracy is measured as the performance metric. The average results show the upper bound method performing the best and the Prettenhofer and Stein [11] method performing better than the baseline method. An issue with using structural correspondence learning is the difficulty in generalizing the pivot functions. For this solution, the pivot functions need to be manually and uniquely defined for a specific application, which makes it very difficult to port to other applications.

The paper by Shi et al. [74], referred to as Heterogeneous Spectral Mapping (HeMap), addresses the specific transfer learning scenario where the input feature space is different between the source and target ($X_S \neq X_T$), the marginal distribution is different between the source and the target ($P(X_S) \neq P(X_T)$), and the output space is different between the source and the target ($Y_S \neq Y_T$.). This solution uses labeled source data that is related to the target domain and limited labeled target data. The first step is to find a common latent input space between the source and target domains using a spectral mapping technique. The spectral mapping technique is modeled as an optimization objective that maintains the original structure of the data while minimizing the difference between the two domains. The next step is to apply a clustering based sample selection method to select related instances as new training data, which resolves the marginal distribution differences in the latent input space. Finally, a Bayesian based method is used to find the

relationship and resolve the differences in the output space. Experiments are performed for the applications of image classification and drug efficacy prediction. Classification error rate is measured as the performance metric. This solution demonstrated better performance as compared to a baseline approach; however, details on the baseline approach are not documented in the paper and no other transfer learning solutions are tested.

The algorithm by Wang and Mahadevan [4], referred to as the Domain Adaptation Manifold Alignment (DAMA) algorithm, proposes using a manifold alignment [78] process to perform a symmetric transformation of the domain input spaces. In this solution, there are multiple labeled source domains and a limited labeled target domain for a total of K domains where all K domains share the same output label space. The approach is to create a separate mapping function for each domain to transform the heterogeneous input space to a common latent input space while preserving the underlying structure of each domain. Each domain is modeled as a manifold. To create the latent input space, a larger matrix model is created that represents and captures the joint manifold union of all input domains. In this manifold model, each domain is represented by a Laplacian matrix that captures the closeness to other instances sharing the same label. The instances with the same labels are forced to be neighbors while separating the instances with different labels. A dimensionality reduction step is performed through a generalized eigenvalue decomposition process to eliminate feature redundancy. The final learner is built in two stages. The first stage is a linear regression model trained on the source data using the latent feature space. The second stage is also a linear regression model that is summed with the first stage. The second stage uses a manifold regularization [50] process to ensure the prediction error is minimized when using the labeled target data. The first stage is trained only using the source data and the second stage compensates for the domain differences caused by the first stage to achieve enhanced target predictions. The experiments are focused on the application of document text classification where classification accuracy is measured as the performance metric. The methods tested against include a Canonical Correlation Analysis approach and a Manifold Regularization approach, which is considered the baseline method. The baseline method uses the limited labeled target domain data and does not use source domain information. The approach presented in this paper substantially outperforms the Canonical Correlation Analysis and baseline approach; however, these approaches are not directly referenced so it is difficult to understand the significance of the test results. A unique aspect of this paper is the modeling of multiple source domains in a heterogeneous solution.

There are scenarios where a large amount of unlabeled heterogeneous source data is readily available that could be used to improve the predictive performance of a particular target learner. The paper by Zhu et al. [7], which presents the method called the Heterogeneous Transfer Learning Image Classification (HTLIC), addresses this scenario with the assumption of having access to a sufficiently large amount of labeled target data. The objective is to use the large supply of available unlabeled source data to create a common latent feature input space that will improve prediction performance in the target classifier. The solution proposed by

Zhu et al. [7] is tightly coupled to the application of image classification and is described as follows. Images with labeled categories (e.g. dog, cake, starfish, etc.) are available in the target domain. To obtain the source data, a web search is performed from Flickr for images that "relate" to the labeled categories. For example, for the category of dog, the words dog, doggy, and greyhound may be used in the Flickr search. As a reference point, the idea of using annotated images from Flickr as unlabeled source data was first proposed by Yang et al. [79]. The retrieved images from Flickr have one or more word tags associated with each image. These tagged image words are then used to search for text documents using Google search. Next, a two-layer bipartite graph is constructed where the first layer represents linkages between the source images and the image tags. The second layer represents linkages between the image tags and the text documents. If an image tag appears in a text document, then a link is created, otherwise there is no link. Images in both the source and the target are initially represented by an input feature set that is derived from the pixel information using SIFT descriptors [65]. Using the initial source image features and the bipartite graph representation derived only from the source image tags and text data, a common latent semantic feature set is learned by employing Latent Semantic Analysis [80]. A learner is now trained with the transformed labeled target instances. Experiments are performed on the proposed approach where 19 different image categories are selected. Binary classification is performed testing different image category pairs. A baseline method is tested using an SVM classifier trained only with the labeled target data. Methods by Raina et al. [81] and by Wang et al. [82] are also tested. The approach proposed by Zhu et al. [7] performed the best overall followed by Raina et al. [81], Wang et al. [82], and baseline approach. The idea of using an abundant source of unlabeled data available through an internet search to improve prediction performance is a very alluring premise. However, this method is very specific to image classification and is enabled by having a web site like Flickr, which essentially provides unlimited labeled image data. This method is difficult to port to other applications.

The transfer learning solution proposed by Qi et al. [75] is another example of an approach that specifically addresses the application of image classification. In the paper by Qi et al. [75], the author claims the application of image classification is inherently more difficult than text classification because image features are not directly related to semantic concepts inherent in class labels. Image features are derived from pixel information, which is not semantically related to class labels, as opposed to word features that have semantic interpretability to class labels. Further, labeled image data is more scarce as compared to labeled text data. Therefore, a transfer learning environment for image classification is desired where an abundance of labeled text data (source) is used to enhance a learner trained on limited labeled image data (target). In this solution, text documents are identified by performing a web search (from Wikipedia for example) on class labels. In order to perform the knowledge transfer from the text documents (source) to the image (target) domain, a bridge in the form of a co-occurrence matrix is used that relates the text and image information. The co-occurrence matrix contains text instances with the corresponding image instances that are found in that particular text

document. The co-occurrence matrix can be programmatically built by crawling web pages and extracting the relevant text and image feature information. Using the co-occurrence matrix, a common latent feature space is found between the text and image features, which is used to learn the final target classifier. This approach, called the Text To Image (TTI) method, is similar to Zhu et al. [7]. However, Zhu et al. [7] does not use labeled source data to enhance the knowledge transfer, which will result in degraded performance when there is limited labeled target data. Experiments are performed with the methods proposed by Qi et al. [75], Dai et al. [83], Zhu et al. [7], and a baseline approach using a standard SVM classifier trained on the limited labeled target data. The text documents are collected from Wikipedia, and classification error rate is measured as the performance metric. The results show the Zhu et al. [7] method performing the best in 15 % of the trials, the Dai et al. [83] method being the best in 10 % of the trials, and the Qi et al. [75] method leading in 75 % of the trials. As with the case of Zhu et al. [7], this method is very specific to the application of image classification and is difficult to port to other applications.

The scenario addressed in the paper by Duan et al. [5] is focused on heterogeneous domain adaptation with a single labeled source domain and a target domain with limited labeled samples. The solution proposed is called Heterogeneous Feature Augmentation (HFA). A transformation matrix P is defined for the source and a transformation matrix Q is defined for the target to project the feature spaces to a common latent space. The latent feature space is augmented with the original source and target feature set and zeros where appropriate. This means the source input data projection has the common latent features, the original source features, and zeros for the original target features. The target input data projection has the common latent features, zeros for the original source features, and the original target features. This feature augmentation method was first introduced by Daumé [14] and is used to correct for conditional distribution differences between the domains. For computational simplification, the P and Q matrices are not directly found but combined and represented by an H matrix. An optimization problem is defined by minimizing the structural risk functional [45] of SVM as a function of the H matrix. The final target prediction function is found using an alternating optimization algorithm to simultaneously solve the dual problem of SVM and the optimal transformation H matrix. The experiments are performed for the applications of image classification and text classification. The source contains labeled image data and the target contains limited labeled image data. For the image features, SURF [84] features are extracted from the pixel information and then clustered into different dimension feature spaces creating the heterogeneous source and target environment. For the text classification experiments, the target contains Spanish language documents and the source contains documents in four different languages. The experiments test against a baseline method, which is constructed by training an SVM learner on the limited labeled target data. Other heterogeneous adaptation methods that are tested include the method by Wang and Mahadevan [4], Shi et al. [74], and Kulis et al. [6]. For the image classification test, the HFA method outperforms all the methods tested by an average of one standard deviation with respect to classification accuracy. The Kulis et al. [6] method has comparable

results to the baseline method (possibly due to some uniqueness in the data set) and the Wang and Mahadevan [4] method slightly outperforms the baseline method (possibly due to a weak manifold structure in the data set). For the text classification test, the HFA method outperforms all methods tested by an average of 1.5 standard deviation. For this test, the Kulis et al. [6] method is second in performance, followed by Wang and Mahadevan [4], and then the baseline method. The Shi et al. [74] method performed worse than the baseline method in both tests. A possible reason for this result is the Shi et al. [74] method does not specifically use the labeled information from the target when performing the symmetric transformation, which will result in degraded classification performance [76].

The work of Li et al. [76], called the Semi-supervised Heterogeneous Feature Augmentation (SHFA) approach, addresses the heterogeneous scenario of an abundance of labeled source data and limited target data, and directly extends the work of Duan et al. [5]. In this work, the H transformation matrix, which is described above by Duan et al. [5], is decomposed into a linear combination of a set of rank-one positive semi-definite matrices that allow for Multiple Kernel Learning solvers (defined by Kloft et al. [85]) to be used to find a solution. In the process of learning the H transformation matrix, the labels for the unlabeled target data are estimated (pseudo labels created) and used while learning the final target classifier. The pseudo labels for the unlabeled target data are found from an SVM classifier trained on the limited labeled target data. The high-level domain adaptation is shown in Fig. 3.3. Experiments are performed for three applications which include image classification (where 31 unique classes are defined), multi-language text document classification (where six unique classes are defined), and multi-language text sentiment classification. Classification accuracy is measured as the performance metric. The method by Li et al. [76] is tested against a baseline method using an SVM learner and trained on the limited labeled target data. Further, other heterogeneous methods tested include Wang and Mahadevan [4], Duan et al. [5], Kulis et al. [6], Shi et al. [74]. By averaging the three different application test results, the order of performance from best to worst is Li et al. [76], Duan et al. [5], Wang and Mahadevan [4], baseline and Kulis et al. [6] (tie), and Shi et al. [74].

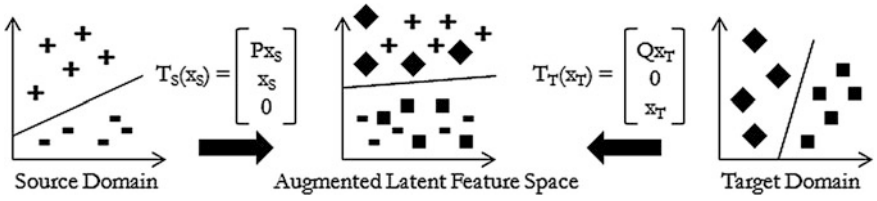

Fig. 3.3 Depicts algorithm approach by Li [76] where the heterogeneous source and target features are transformed to an augmented latent feature space. T_S and T_T are transformation functions. P and Q are projection matrices as described in Duan [5]. Diagram adapted from Li [76]

Asymmetric Feature-Based Transfer Learning

The work of Kulis et al. [6], referred to as the Asymmetric Regularized Cross-domain Transformation (ARC-t), proposes an asymmetric transformation algorithm to resolve the heterogeneous feature space between domains. For this scenario, there is an abundance of labeled source data and limited labeled target data. An objective function is first defined for learning the transformation matrix. The objective function contains a regularizer term and a cost function term that is applied to each pair of cross-domain instances and the learned transformation matrix. The construction of the objective function is responsible for the domain invariant transformation process. The optimization of the objective function aims to minimize the regularizer and the cost function terms. The transformation matrix is learned in a non-linear Gaussian RBF kernel space. The method presented is referred to as the Asymmetric Regularized Cross-domain transformation. Two experiments using this approach are performed for image classification where classification accuracy is measured as the performance metric. There are 31 image classes defined for these experiments. The first experiment (test 1) is where instances of all 31 image classes are included in the source and target training data. In the second experiment (test 2), only 16 image classes are represented in the target training data (all 31 are represented in the source). To test against other baseline approaches, a method is needed to bring the source and target input domains together. A preprocessing step called Kernel Canonical Correlation Analysis (proposed by Shawe-Taylor and Cristianini [86]) is used to project the source and target domains into a common domain space using symmetric transformation. Baseline approaches tested include k-nearest neighbors, SVM, metric learning proposed by Davis et al. [87], feature augmentation proposed by Daumé [14], and a cross domain metric learning method proposed by Saenko et al. [88]. For test 1, the Kulis et al. [6] approach performs marginally better than the other methods tested. For test 2, the Kulis et al. [6] approach performs significantly better compared to the k-nearest neighbors approach (note the other methods cannot be tested against as they require all 31 classes to be represented in the target training data). The Kulis et al. [6] approach is best suited for scenarios where all of the classes are not represented in the target training data as demonstrated in test 2.

The problem domain defined by Harel and Mannor [8] is of limited labeled target data and multiple labeled data sources where an asymmetric transformation is desired for each source to resolve the mismatch in feature space. The first step in the process is to normalize the features in the source and target domains, then group the instances by class in the source and target domains. For each class grouping, the features are mean adjusted to zero. Next, each individual source class group is paired with the corresponding target class group, and a singular value decomposition process is performed to find the specific transformation matrix for that class grouping. Once the transformation is performed, the features are mean shifted back reversing the previous step, and the final target classifier is trained using the transformed data. Finding the transformation matrix using the singular value

decomposition process allows for the marginal distributions within the class groupings to be aligned while maintaining the structure of the data. This approach is referred to as the Multiple Outlook MAPping algorithm (MOMAP). The experiments use data taken from wearable sensors for the application of activity classification. There are five different activities defined for the experiment which include walking, running, going upstairs, going downstairs, and lingering. The source domain contains similar (but different) sensor readings as compared to the target. The method proposed by Harel and Mannor [8] is compared against a baseline method that trains a classifier with the limited labeled target data and an upper bound method that uses a significantly larger set of labeled target data to train a classifier. An SVM learner is used as the base classifier and a balanced error rate (due to an imbalance in the test data) is measured as the performance metric. The Harel and Mannor [8] approach outperforms the baseline method in every test and falls short of the upper bound method in every test with respect to the balanced error rate.

The heterogeneous transfer learning scenario addressed by Zhou et al. [10] requires an abundance of labeled source data and limited labeled target data. An asymmetric transformation function is proposed to map the source features to the target features. To learn the transformation matrix, a multi-task learning method based on Ando and Zhang [89] is adopted. The solution, referred to as the Sparse Heterogeneous Feature Representation (SHFR), is implemented by creating a binary classifier for each class in the source and the target domains separately. Each binary classifier is assigned a weight term where the weight terms are learned by combining the weighted classifier outputs, while minimizing the classification error of each domain. The weight terms are now used to find the transformation matrix by minimizing the difference between the target weights and the transformed source weights. The final target classifier is trained using the transformed source data and original target data. Experiments are performed for text document classification where the target domain contains documents written in one language and the source domain contains documents written in different languages. A baseline method using a linear SVM classifier trained on the labeled target is established along with testing against the methods proposed by Wang and Mahadevan [4], Kulis et al. [6], and Duan et al. [5]. The method proposed by Zhou et al. [10] performed the best for all tests with respect to classification accuracy. The results of the other approaches are mixed as a function of the data sets used where the Duan et al. [5] method performed either second or third best.

The application of software module defect prediction is usually addressed by training a classifier with labeled data taken from the software project of interest. The environment described in Nam and Kim [9] for software module defect prediction attempts to use labeled source data from one software project to train a classifier to predict unlabeled target data from another project. The source and target software projects collect different metrics making the source and target feature spaces heterogeneous. The proposed solution, referred to as the Heterogeneous Defect Prediction (HDP) approach, is to first select the important features from the source domain using a feature selection method to eliminate redundant and irrelevant features. Feature selection methods used include gain ratio, chi-square, relief-F, and

significance attribute evaluation (see Gao et al. [90] and Shivaji et al. [91]). The next step is to statistically match the selected source domain features to ones in the target using a Kolmogorov-Smirnov test that measures the closeness of the empirical distribution between the two sources. A learner is trained with the source features that exhibit a close statistical match to the corresponding target features. The target data is tested with the trained classifier using the corresponding matched features of the target. Even though the approach by Nam and Kim [9] is applied directly to the application of software module defect prediction, this method can be used for other applications. Experiments are performed using five different software defect data sets with heterogeneous features. The proposed method by Nam and Kim [9] uses Logistic Regression as the base learner. The other approaches tested include a within project defect prediction (WPDP) approach where the learner is trained on labeled target data, a cross project defect prediction (CPDP-CM) approach where the source and target represent different software projects but have homogeneous features, and a cross project defect prediction approach with heterogeneous features (CPDP-IFS) as proposed by He et al. [92]. The results of the experiment show the Nam and Kim [9] method significantly outperformed all other approaches with respect to area under the curve measurement. The WPDP approach is next best followed by the CPDP-CM approach and the CPDP-IFS approach. These results can be misleading as the Nam and Kim [9] approach could only match at least one or more input features between the source and target domains in 37 % of the tests. Therefore, in 63 % of the cases, the Nam and Kim [9] method could not be used and these cases are not counted. The WPDP method represents an upper bound and it is an unexpected result that the Nam and Kim [9] approach would outperform the WPDP method.

The paper by Zhou et al. [12] claims that previous heterogeneous solutions assume the instance correspondence between the source and target domains are statistically representative (distributions are equal), which may not always be the case. An example of this claim is in the application of text sentiment classification where the word bias problem previously discussed causes distribution differences between the source and target domains. The paper by Zhou et al. [12] proposes a solution called the Hybrid Heterogeneous Transfer Learning (HHTL) method for a heterogeneous environment with abundant labeled source data and abundant unlabeled target data. The idea is to first learn an asymmetric transformation from the target to the source domain, which reduces the problem to a homogeneous domain adaptation issue. The next step is to discover a common latent feature space using the transformed data (from the previous step) to reduce the distribution bias between the transformed unlabeled target domain and the labeled source domain. Finally, a classifier is trained using the common latent feature space from the labeled source data. This solution is realized using a deep learning method employing a Marginalized Stacked Denoised Autoencoder as proposed by Chen et al. [93] to learn the asymmetric transformation and the mapping to a common latent feature space. The previous surveyed paper by Glorot et al. [33] demonstrated a deep learning approach finding a common latent feature space for homogeneous source and target feature set. The experiments focused on multiple language text

sentiment classification where English is used in the source and three other languages are separately used in the target. Classification accuracy is measured as the performance metric. Other methods tested include a heterogeneous spectral mapping approach proposed by Shi et al. [74], a method proposed by Vinokourov et al. [94], and a multimodal deep learning approach proposed by Ngiam et al. [95]. An SVM learner is used as the base classifier for all methods. The results of the experiment from best to worst performance are Zhou et al. [12], Ngiam et al. [95], Vinokourov et al. [94], and Shi et al. [74].

Improvements to Heterogeneous Solutions

The paper by Yang et al. [96] proposes to quantify the amount of knowledge that can be transferred between domains in a heterogeneous transfer learning environment. In other words, it attempts to measure the "relatedness" of the domains. This is accomplished by first building a co-occurrence matrix for each domain. The co-occurrence matrix contains the set of instances represented in every domain. For example, if one particular text document is an instance in the co-occurrence matrix, that text document is required to be represented in every domain. Next, Principal Component Analysis is used to select the most important features in each domain and assign the principal component coefficient to those features. The principal component coefficients are used to form a directed cyclic network (DCN) where each node represents a domain (either source or target) and each node connection (edge weight) is the conditional dependence from one domain to another. The DCN is built using a Markov Chain Monte Carlo method. The edge weights represent the potential amount of knowledge that can be transferred between domains where a higher value means higher knowledge transfer. These edge weights are then used as tuning parameters in different heterogeneous transfer learning solutions, which include works from Yang et al. [79], Ng et al. [97], and Zhu et al. [7] (the weights are calculated first using Yang et al. [96] and then applied as tuning values in the other solutions). Note, that integrating the edge weight values into a particular approach is specific to the implementation of the solution and cannot be generically applied. The experiments are run on the three different learning solutions comparing the original solution against the solution using the weighted edges of the DCN as the tuned parameters. In all three solutions, the classification accuracy is improved using the DCN tuned parameters. One potential issue with this approach is the construction of the co-occurrence matrix. The co-occurrence matrix contains many instances; however, each instance must be represented in each domain. This may be an unrealistic constraint in many real-world applications.

Experiment Results

In reviewing the experiment results of the previous surveyed papers, there are instances where one solution can show varying results over a range of different experiments. There are many reasons why this can happen which include varying test environments, different test implementations, different applications being tested, and different data sets being used. An interesting area of future work is to evaluate the solutions presented to determine the best performing solutions as a function of specific datasets. To facilitate that goal, a repository of open-source software containing the software implementations for solutions used in each paper would be extremely beneficial. Table 3.3 lists a compilation of head-to-head results for the most commonly tested solutions contained in the Heterogeneous Transfer Learning section. The results listed in Table 3.3 represent a win, loss, and tie performance record of the head-to-head solution comparisons. Note, these results are compiled directly from the surveyed papers. It is difficult to draw exact conclusions from this information because of the reasons just outlined; however, it provides some interesting insight into the comparative performances of the solutions.

Discussion of Heterogeneous Solutions

The previous surveyed heterogeneous transfer learning works demonstrate many different characteristics and attributes. Which heterogeneous transfer learning solution is best for a particular application? The heterogeneous transfer learning solutions use either a symmetric transformation or an asymmetric transformation process in an attempt to resolve the differences between the input feature space (as shown in Fig. 3.1). The asymmetrical transformation approach is best used when the same class instances in the source and target domains can be transformed

Table 3.3 Lists the head-to-head results of experiments performed in the heterogeneous transfer learning works surveyed

Methods	HeMap	ARC-t	DAMA	HFA	SHFR	SHFA
HeMap [74]	–	0-5-0	0-5-0	0-5-0	0-0-0	0-3-0
ARC-t [6]	5-0-0	–	4-2-0	1-7-0	0-3-0	0-3-0
DAMA [4]	5-0-0	2-4-0	–	0-8-0	0-3-0	0-3-0
HFA [5]	5-0-0	7-1-0	8-0-0	–	0-3-0	0-3-0
SHFR [10]	0-0-0	3-0-0	3-0-0	3-0-0	–	0-0-0
SHFA [76]	3-0-0	3-0-0	3-0-0	3-0-0	0-0-0	–

The numbers (x-y-z) in the table indicate the far left column method outperforms the top row method x times, underperforms y times, and has similar performance z times

Table 3.4 Heterogeneous transfer learning approaches surveyed in Sect. 4 listing various characteristics of each approach

Approach	Transfer category	Source data	Target data	Multiple sources	Generic solution	Negative transfer
CLSCL [11]	Symmetric feature	Labeled	Unlabeled			
HeMap [74]	Symmetric feature	Labeled	Limited labels		✓	
DAMA [4]	Symmetric feature	Labeled	Limited labels	✓	✓	
HTLIC [7]	Symmetric feature	Unlabeled	Abundant labels			
TTI [75]	Symmetric feature	Labeled	Limited labels			
HFA [5]	Symmetric feature	Labeled	Limited labels		✓	
SHFA [76]	Symmetric feature	Labeled	Limited labels		✓	
ARC-t [6]	Asymmetric feature	Labeled	Limited labels		✓	
MOMAP [8]	Asymmetric feature	Labeled	Limited labels	✓		
SHFR [10]	Asymmetric feature	Labeled	Limited labels		✓	
HDP [9]	Asymmetric feature	Labeled	Unlabeled		✓	
HHTL [12]	Asymmetric feature	Labeled	Unlabeled		✓	

without context feature bias. Many of the surveyed heterogeneous transfer learning solutions only address the issue of the input feature space being different between the source and target domains and do not address other domain adaptation steps needed for marginal and/or conditional distribution differences. If further domain adaptation needs to be performed after the input feature spaces are aligned, then an appropriate homogeneous solution should be used. To further help determine which solution is best for a given transfer learning application, the information in Table 3.4 should be used to match the characteristics of the solution to that of the desired application environment. None of the surveyed heterogeneous transfer learning solutions have a means to guard against negative transfer effects. However, the paper by Yang et al. [96] demonstrates that negative transfer guards can benefit heterogeneous transfer learning solutions. It seems likely that future heterogeneous transfer learning works will integrate means for negative transfer protection. Many of the same heterogeneous transfer learning solutions are tested in the surveyed solution experiments. These head-to-head comparisons are summarized in Table 3.3 and can be used as a starting point to understand the relative performance between the solutions. As observed as a trend in the previous homogeneous

solutions, the recent heterogeneous solution by Duan et al. [5] employs a one-stage solution that simultaneously performs the feature input space alignment process while learning the final classifier. As is the case for the surveyed homogeneous transfer learning works, the surveyed heterogeneous transfer learning works are not specifically applied to big data solutions; however, there is nothing to preclude their use in a big data environment.

Negative Transfer

The high-level concept of transfer learning is to improve a target learner by using data from a related source domain. But what happens if the source domain is not well-related to the target? In this case, the target learner can be negatively impacted by this weak relation, which is referred to as negative transfer. In a big data environment, there may be a large dataset where only a portion of the data is related to a target domain of interest. For this case, there is a need to divide the dataset into multiple sources and employ negative transfer methods when using transfer learning algorithm. In the scenario where multiple datasets are available that initially appear to be related to the target domain of interest, it is desired to select the datasets that provide the best information transfer and avoid the datasets that cause negative transfer. This allows for the best use of the available large datasets. How related do the source and target domains need to be for transfer learning to be advantageous? The area of negative transfer has not been widely researched, but the following papers begin to address this issue.

An early paper by Rosenstein et al. [98] discusses the concept of negative transfer in transfer learning and claims that the source domain needs to be sufficiently related to the target domain; otherwise, the attempt to transfer knowledge from the source can have a negative impact on the target learner. Cases of negative transfer are demonstrated by Rosenstein et al. [98] in experiments using a hierarchical Naive Bayes classifier. The author also demonstrates the chance of negative transfer goes down as the number of labeled target training samples goes up.

The paper by Eaton et al. [99] proposes to build a target learner based on a transferability measure from multiple related source domains. The approach first builds a Logistic Regression learner for each source domain. Next, a model transfer graph is constructed to represent the transferability between each source learner. In this case, transferability from a first learner to a second learner is defined as the performance of the second learner with learning from the first learner minus the performance of the second learner without learning from the first learner. Next, the model transfer graph is modified by adding the transferability measures between the target learner and all the source learners. Using spectral graph theory [55] on the model transfer graph, a transfer function is derived that maintains the geometry of the model transfer graph and is used in the final target learner to determine the level

of transfer from each source. Experiments are performed in the applications of document classification and alphabet classification. Source domains are identified that are either related or unrelated to the target domain. The method by Eaton et al. [99] is tested along with a handpicked method where the source domains are manually selected to be related to the target, an average method that uses all sources available, and a baseline method that does not use transfer learning. Classification accuracy is the performance metric measured in the experiments. The source and target domains are represented by a homogeneous feature input space. The results of the experiments are mixed. Overall, the Eaton et al. [99] approach performs the best; however, there are certain instances where Eaton et al. [99] performed worse than the handpicked, average, and baseline methods. In the implementation of the algorithm, the transferability measure between two sources is required to be the same; however, the transferability from source 1 to source 2 is not always equal to the transferability from source 2 to source 1. A suggestion for future improvement is to use directed graphs to specify the bidirectional nature of the transferability measure between two sources.

The paper by Ge et al. [100] claims that knowledge transfer can be inhibited due to the existence of unrelated or irrelevant source domains. Further, current transfer learning solutions are focused on transferring knowledge from source domains to a target domain, but are not concerned about different source domains that could potentially be irrelevant and cause negative transfer. In the model presented by Ge et al. [100], there is a single target domain with limited labeled data and multiple labeled source domains for knowledge transfer. To reduce negative transfer effects from unrelated source domains, each source is assigned a weight (called the Supervised Local Weight) corresponding to how related the source is with the target (the higher the weight the more it is related). The Supervised Local Weight is found by first using a spectral clustering algorithm [55] on the unlabeled target information and propagating labels to the clusters from the labeled target information. Next, each source is separately clustered and labels assigned to the clusters from the labeled source. The Supervised Local Weight of each source cluster is computed by comparing the source and target clusters. This solution further addresses the issue of imbalanced class distribution in source domains by preventing a high-weight class assignment in the case of high-accuracy predictions in a minority target class. The final target learner uses the Supervised Local Weights to attenuate the effects of negative transfer. Experiments are performed in three application areas including Cardiac Arrhythmia Detection, Spam Email Filtering, and Intrusion Detection. Area under the curve is measured as the performance metric. The source and target domains are represented by a homogeneous feature input space. The method presented in this paper is compared against methods by Luo et al. [101], by Gao et al. [102], by Chattopadhyay et al. [15], and by Gao et al. [23]. The Luo et al. [101] and Gao et al. [102] methods are the worst performing, most likely due to the fact that these solutions do not attempt to combat negative transfer effects. The Chattopadhyay et al. [15] and Gao et al. [23] methods are the next best performing,

which have means in place to reduce the effects of negative transfer from the source domains. The Chattopadhyay et al. [15] and Gao et al. [23] methods do address the negative transfer problem but do not address the imbalanced distribution issue. The Ge et al. [100] method does exhibit the best overall performance due to the handling of negative transfer and imbalanced class distribution.

The paper by Seah et al. [103] claims the root cause of negative transfer is mainly due to conditional distribution differences between source domains (P_{S1} (y|x) $\neq$ P_{S2} (y|x)) and a difference in class distribution (class imbalance) between the source and target ($P_S(y) \neq P_T(y)$). Because the target domain usually contains a small number of labeled instances, it is difficult to find the true class distribution of the target domain. A Predictive Distribution Matching (PDM) framework is proposed to align the conditional distributions of the source domains and target domain in an attempt to minimize negative transfer effects. A positive transferability measure is defined that measures the transferability of instance pairs with the same label from the source and target domains. The first step in the PDM framework is to assign pseudo labels to the unlabeled target data. This is accomplished by an iterative process that forces source and target instances which are similar (as defined by the positive transferability measure) to have the same label. Next, irrelevant source data are removed by identifying data that does not align with the conditional distribution of the pseudo labeled target data for each class. Both Logistic Regression and SVM classifiers are implemented using the PDM framework. Experiments are performed on document classification using the PDM method described in this paper, the approach from Daumé [14], the approach from Huang et al. [20], and the approach from Bruzzone and Marconcini [67]. Classification accuracy is measured as the performance metric. The source and target domains are represented by a homogeneous feature input space. The PDM approach demonstrates better performance as compared to the other approaches tested as these solutions do not attempt to account for negative transfer effects.

A select number of previously surveyed papers contain solutions addressing negative transfer. The paper by Yang et al. [96] addresses the negative transfer issue, which is presented in the Heterogeneous Transfer Learning section. The homogeneous solution by Gong et al. [16] defines an ROD value that measures the relatedness between a source and target domain. The work presented in Chattopadhyay et al. [15] is a multiple source transfer learning approach that calculates the source weights as a function of conditional probability differences between the source and target domains attempting to give the most related sources the highest weights. Duan et al. [37] proposes a transfer learning approach that only uses source domains that are deemed relevant and test data demonstrates better performance compared to methods with no negative transfer protection.

The previous papers attempt to measure how related source data is to the target data in a transfer learning environment and then selectively transfer the information that is highly related. The experiments in the above papers demonstrate that accounting for negative transfer effects from source domain data can improve target

learner performance. However, most transfer learning solutions do not attempt to account for negative transfer effects. Robust negative transfer measurements are difficult to define. Since the target domain typically has limited labeled data, it is inherently difficult to find a true measure of the relatedness between the source and target domains. Further, by selectively transferring information that seems related to the limited labeled target domain, a risk of overfitting in the target learner is a concern. The topic of negative transfer is a fertile area for further research.

Transfer Learning Applications

The surveyed works in this paper demonstrate that transfer learning has been applied to many real-world applications. There are a number of application examples pertaining to natural language processing, more specifically in the areas of sentiment classification, text classification, spam email detection, and multiple language text classification. Other well-represented transfer learning applications include image classification and video concept classification. Applications that are more selectively addressed in the previous papers include WiFi localization classification, muscle fatigue classification, drug efficacy classification, human activity classification, software defect classification, and cardiac arrhythmia classification.

The majority of the solutions surveyed are generic, meaning the solution can be easily applied to applications other than the ones implemented and tested in the papers. The application-specific solutions tend to be related to the field of natural language processing and image processing. In the literature, there are a number of transfer learning solutions that are specific to the application of recommendation systems. Recommendation systems provide users with recommendations or ratings for a particular domain (e.g. movies, books, etc.), which are based on historical information. However, when the system does not have sufficient historical information (referred to as the data sparsity issue presented in [104], then the recommendations are not reliable. In the cases where the system does not have sufficient domain data to make reliable predictions (for example when a movie is just released), there is a need to use previously collected information from a different domain (using books for example). The aforementioned problem has been directly addressed using transfer learning methodologies and captured in papers by Moreno et al. [104], Cao et al. [105], Li et al. [106, 107], Pan et al. [108, 110], Zhang et al. [109], Roy et al. [111], Jiang et al. [112], and Zhao et al. [113].

Transfer learning solutions continue to be applied to a diverse number of real-world applications, and in some cases the applications are quite obscure. The application of head pose classification finds a learner trained with previously captured labeled head positions to predict a new head position. Head pose classification

is used for determining the attentiveness of drivers, analyzing social behavior, and human interaction with robots. Head positions captured in source training data will have different head tilt ranges and angles than that of the predicted target. The paper by Rajagopal et al. [114] addresses the head pose classification issues using transfer learning solutions.

Other transfer learning applications include the paper by Ma et al. [115] that uses transfer learning for atmospheric dust aerosol particle classification to enhance global climate models. Here the TrAdaBoost algorithm proposed by Dai et al. [69] is used in conjunction with an SVM classifier to improve on classification results. Being able to identify areas of low income in developing countries is important for disaster relief efforts, food security, and achieving sustainable growth. To better predict poverty mapping, Xie et al. [116] proposes an approach similar to Oquab et al. [35] that uses a convolution neural network model. The first prediction model is trained to predict night time light intensity from source image data. The final target prediction model predicts the poverty mapping from source night time light intensity data. In the paper by Ogoe et al. [117], transfer learning in used to enhance disease prediction. In this solution, a rule-based learning approach is formulated to use abstract source domain data to perform modeling of multiple types of gene expression data. Online display web advertising is a growing industry where transfer learning is used to optimally predict targeted ads. In the paper by Perlich et al. [118], a transfer learning approach is employed that uses the weighted outputs of multiple source classifiers to enhance a target classifier trained to predict targeted online display advertising results. The paper by Kan et al. [119] addresses the field of facial recognition and is able to use face image information from one ethnic group to improve the learning of a classifier for a different ethnic group. The paper by Farhadi et al. [120] is focused on the application of sign language recognition where the model is able to learn from different people signing at various angles. Transfer learning is applied to the field of biology in the paper by Widmer and Ratsch [121]. Specifically, a multi-task learning approach is used in the prediction of splice sites in genome biology. Predicting if patients will contract a particular bacteria when admitted to a hospital is addressed in the paper by Wiens et al. [122]. Information taken from different hospitals is used to predict the infection rate for a different hospital. In the paper by Romera-Paredes et al. [123], a multi-task transfer learning approach is used to predict pain levels from an individual's facial expression by using labeled source facial images from other individuals. The paper by Deng et al. [124] applies transfer learning to the application of speech emotion recognition where information is transferred from multiple labeled speech sources. The application of wine quality classification is implemented in Zhang and Yeung [125] using a multi-task transfer learning approach. As a reference, the survey paper by Cook et al. [18] covers transfer learning for the application of activity recognition and the survey papers by Patel et al. [126] and Shao et al. [127] address transfer learning in the domain of image recognition.

Conclusion and Discussion

The subject of transfer learning is a well-researched area as evidenced with more than 700 academic papers addressing the topic in the last five years. This survey paper presents solutions from the literature representing current trends in transfer learning. Homogeneous transfer learning papers are surveyed that demonstrate instance-based, feature-based, parameter-based, and relational-based information transfer techniques. Solutions having various requirements for labeled and unlabeled data are also presented as a key attribute. The relatively new area of heterogeneous transfer learning is surveyed showing the two dominant approaches for domain adaptation being asymmetric and symmetric transformations. Many real-world applications that transfer learning is applied to are listed and discussed in this survey paper. In some cases, the proposed transfer learning solutions are very specific to the underlying application and cannot be generically used for other applications. A list of software downloads implementing a portion of the solutions surveyed is presented in the appendix of this paper. A great benefit to researchers is to have software available from previous solutions so experiments can be performed more efficiently and more reliably. A single open-source software repository for published transfer learning solutions would be a great asset to the research community.

In many transfer learning solutions, the domain adaptation process performed is focused either on correcting the marginal distribution differences or the conditional distribution differences between the source and target domains. Correcting the conditional distribution differences is a challenging problem due to the lack of labeled target data. To address the lack of labeled target data, some solutions estimate the labels for the target data (called pseudo labels), which are then used to correct the conditional distribution differences. This method is problematic because the conditional distribution corrections are being made with the aid of pseudo labels. Improved methods for correcting the conditional distribution differences is a potential area of future research. A number of more recent works attempt to correct both the marginal distribution differences and the conditional distribution differences during the domain adaptation process. An area of future work is to quantify the advantage of correcting both distributions and in what scenarios it is most effective. Further, Long et al. [30] states that the simultaneous solving of marginal and conditional distribution differences is preferred over serial alignment as it reduces the risk of overfitting. Another area of future work is to quantify any performance gains for simultaneously solving both distribution differences. In addition to solving for distribution differences in the domain adaptation process, exploring possible data preprocessing steps using heuristic knowledge of the domain features can be used as a method to improve the target learner performance. The heuristic knowledge would represent a set of complex rules or relations that standard transfer learning techniques cannot account for. In most cases, this heuristic knowledge would be specific to each domain, which would not lead to a

generic solution. However, if such a preprocessing step leads to improved target learner performance, it is likely worth the effort.

A trend observed in the formulation of transfer learning solutions is in the implementation of a one-stage process as opposed to a two-stage process. A two-stage solution first performs the domain adaptation process and then independently learns the final classifier. A one-stage process simultaneously performs the domain adaptation process while learning the final classifier. Recent solutions employing a one-stage solution include Long et al. [30], Duan et al. [28], Shi and Sha [34], Xia et al. [39], and Duan et al. [5]. With respect to the one-stage solution, Long et al. [30] claims the simultaneous solving of domain adaptation and the classifier establishes mutual reinforcement for enhanced performance. An area of future work is to better quantify the effects of a one-stage approach over a two-stage approach.

This paper surveys a number of works addressing the topic of negative transfer. The subject of negative transfer is still a lightly researched area. The expanded integration of negative transfer techniques into transfer learning solutions is a natural extension for future research. Solutions supporting multiple source domains enabling the splitting of larger source domains into smaller domains to more easily discriminate against unrelated source data are a logical area for continued research. Additionally, optimal transfer is another fertile area for future research. Negative transfer is defined as a source domain having a negative impact on a target learner. The concept of optimal transfer is when select information from a source domain is transferred to achieve the highest possible performance in a target learner. There is overlap between the concepts of negative transfer and optimal transfer; however, optimal transfer attempts to find the best performing target learner, which goes well beyond the negative transfer concept.

With the recent proliferation of sensors being deployed in cell phones, vehicles, buildings, roadways, and computers, larger and more diverse information is being collected. The diversity in data collection makes heterogeneous transfer learning solutions more important moving forward. Larger data collection sizes highlight the potential for big data solutions being deployed concurrent with current transfer learning solutions. How the diversity and large size of sensor data integrates into transfer learning solutions is an interesting topic of future research. Another area of future work pertains to the scenario where the output label space is different between domains. With new data sets being captured and being made available, this topic could be a needed area of focus for the future. Lastly, the literature has very few transfer learning solutions addressing the scenario of unlabeled source and unlabeled target data, which is certainly an area for expanded research.

Appendix

The majority of transfer learning solutions surveyed are complex and implemented with non-trivial software. It is a great advantage for a researcher to have access to software implementations of transfer learning solutions so comparisons with competing solutions are facilitated more quickly and fairly. Table 3.5 provides a list of available software downloads for a number of the solutions surveyed in this paper. Table 3.6 provides a resource for useful links that point to transfer learning tutorials and other interesting articles on the topic of transfer learning.

Table 3.5 Software downloads for various transfer learning solutions

Approach	Location
Prettenhofer and Stein [11]	https://github.com/pprett/bolt [128]
Zhu et al. [7]	http://www.cse.ust.hk/~yinz/ [129]
Dai et al. [69]	https://github.com/BoChen90/machine-learning-matlab/blob/master/TrAdaBoost.m [130]
Daumé [14]	http://hal3.name/easyadapt.pl.gz [131]
Duan et al. [5]	https://sites.google.com/site/xyzliwen/publications/HFA_release_0315.rar [132]
Kulis et al. [6]	http://vision.cs.uml.edu/adaptation.html [133]
Qi et al. [75]	http://www.eecs.ucf.edu/~gqi/publications.html [134]
Li et al. [76]	http://www.lxduan.info/#sourcecode_hfa [135]
Gong [16]	http://www-scf.usc.edu/~boqinggo/ [136]
Long et al. [30]	http://ise.thss.tsinghua.edu.cn/~mlong/ [137]
Oquab et al. [35]	http://leon.bottou.org/papers/oquab-2014 [138]
Long et al. [29]	http://ise.thss.tsinghua.edu.cn/~mlong/ [137]
Other transfer learning code	http://www.cse.ust.hk/TL/ [139]

Table 3.6 Useful links for transfer learning information

Item	Location
Slides for Nam and Kim [9]	http://www.slideshare.net/hunkim/heterogeneous-defect-prediction-esecfse-2015 [140]
Code for SVMLIB	http://www.csie.ntu.edu.tw/~cjlin/libsvm [141]
Slide for Kulis et al. [6]	https://www.eecs.berkeley.edu/~jhoffman/domainadapt/ [142]
Tutorial on transfer learning	http://tommasit.wix.com/datl14tutorial [143]
Tutorial on transfer learning	http://sifaka.cs.uiuc.edu/jiang4/domain_adaptation/survey/da_survey.html [144]
Overview of Duan et al. [37]	http://lxduan.info/papers/DuanCVPR2012_poster.pdf [145]

References

1. Witten IH, Frank E. Data mining, practical machine learning tools and techniques. 3rd ed. San Francisco, CA: Morgan Kaufmann Publishers; 2011.
2. Shimodaira H. Improving predictive inference under covariate shift by weighting the log-likelihood function. J Stat Plan Inference. 2000;90(2):227–44.
3. Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng. 2010;22 (10):1345–59.
4. Wang C, Mahadevan S. Heterogeneous domain adaptation using manifold alignment. In: Proceedings of the twenty-second international joint conference on artificial intelligence, vol. 2; 2011. p. 1541–6.
5. Duan L, Xu D, Tsang IW. Learning with augmented features for heterogeneous domain adaptation. IEEE Trans Pattern Anal Mach Intell. 2012;36(6):1134–48.
6. Kulis B, Saenko K, Darrell T. What you saw is not what you get: domain adaptation using asymmetric kernel transforms. In: IEEE 2011 conference on computer vision and pattern recognition; 2011. p. 1785–92.
7. Zhu Y, Chen Y, Lu Z, Pan S, Xue G, Yu Y, Yang Q. Heterogeneous transfer learning for image classification. Proc Nat Conf Artif Intell. 2011;2:1304–9.
8. Harel M, Mannor S. Learning from multiple outlooks. In: Proceedings of the 28th international conference on machine learning; 2011. p. 401–408.
9. Nam J, Kim S. Heterogeneous defect prediction. In: Proceedings of the 2015 10th joint meeting on foundations of software engineering; 2015. p. 508–19.
10. Zhou JT, Tsang IW, Pan SJ Tan M. Heterogeneous domain adaptation for multiple classes. In: International conference on artificial intelligence and statistics; 2014. p. 1095–103.
11. 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; 2010. p. 1118–27.
12. Zhou JT, Pan S, Tsang IW, Yan Y. Hybrid heterogeneous transfer learning through deep learning. Proc Nat Conf Artif Intell. 2014;3:2213–20.
13. Taylor ME, Stone P. Transfer learning for reinforcement learning domains: a survey. JMLR. 2009;10:1633–85.
14. Daumé H III. Frustratingly easy domain adaptation. In: Proceedings of 2007 ACL; 2007. p. 256–63.
15. Chattopadhyay R, Ye J, Panchanathan S, Fan W, Davidson I. Multi-source domain adaptation and its application to early detection of fatigue. ACM Trans Knowl Discov Data. 2011;6(4):18.
16. Gong B, Shi Y, Sha F, Grauman K. Geodesic flow kernel for unsupervised domain adaptation. In: Proceedings of the 2012 IEEE conference on computer vision and pattern recognition; 2012. p. 2066–73.
17. 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; 2006. p. 120–28.
18. Cook DJ, Feuz KD, Krishnan NC. Transfer learning for activity recognition: a survey. Knowl Inf Syst. 2012;36(3):537–56.
19. Feuz KD, Cook DJ. Transfer learning across feature-rich heterogeneous feature spaces via feature-space remapping (FSR). J ACM Trans Intell Syst Technol. 2014;6(1):1–27.
20. Huang J, Smola A, Gretton A, Borgwardt KM, Schölkopf B. Correcting sample selection bias by unlabeled data. In: Proceedings of the 2006 conference in advances in neural information processing systems; 2006. p. 601–8.
21. Jiang J, Zhai C. Instance weighting for domain adaptation in NLP. In: Proceedings of the 45th annual meeting of the association of computational linguistics; 2007. p. 264–271.
22. Pan SJ, Kwok JT, Yang Q. Transfer learning via dimensionality reduction. In: Proceedings of the 23rd national conference on artificial intelligence, vol. 2; 2008. p. 677–82.

23. Gao J, Fan W, Jiang J, Han J. Knowledge transfer via multiple model local structure mapping. In: Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining; 2008. p. 283–91.
24. Bonilla E, Chai KM, Williams C. Multi-task Gaussian process prediction. In: Proceedings of the 20th annual conference of neural information processing systems; 2008. p. 153–60.
25. Evgeniou T, Pontil M. Regularized multi-task learning. In: Proceedings of the 10th ACM SIGKDD international conference on knowledge discovery and data mining; 2004. p. 109–17.
26. Mihalkova L, Mooney RJ. Transfer learning by mapping with minimal target data. In: Proceedings of the association for the advancement of artificial intelligence workshop transfer learning for complex tasks; 2008. p. 31–6.
27. Li F, Pan SJ, Jin O, Yang Q, Zhu X. Cross-domain co-extraction of sentiment and topic lexicons. In: Proceedings of the 50th annual meeting of the association for computational linguistics long papers, vol. 1; 2012. p. 410–9.
28. Duan L, Tsang IW, Xu D. Domain transfer multiple kernel learning. IEEE Trans Pattern Anal Mach Intell. 2012;34(3):465–79.
29. Long M, Wang J, Ding G, Sun J, Yu PS. Transfer feature learning with joint distribution adaptation. In: Proceedings of the 2013 IEEE international conference on computer vision; 2013. p. 2200–7.
30. Long M, Wang J, Ding G, Pan SJ, Yu PS. Adaptation regularization: a general framework for transfer learning. IEEE Trans Knowl Data Eng. 2014;26(5):1076–89.
31. Pan SJ, Tsang IW, Kwok JT, Yang Q. Domain adaptation via transfer component analysis. IEEE Trans Neural Netw. 2009;22(2):199–210.
32. Pan SJ, Ni X, Sun JT, Yang Q, Chen Z. Cross-domain sentiment classification via spectral feature alignment. In: Proceedings of the 19th international conference on world wide web; 2010. p. 751–60.
33. Glorot X, Bordes A, Bengio Y. Domain adaptation for large-scale sentiment classification: a deep learning approach. In: Proceedings of the twenty-eight international conference on machine learning, vol. 27; 2011. p. 97–110.
34. Shi Y, Sha F. Information-theoretical learning of discriminative clusters for unsupervised domain adaptation. In: Proceedings of the 29th international conference on machine learning; 2012. p. 1–8.
35. Oquab M, Bottou L, Laptev I, Sivic J. Learning and transferring mid-level image representations using convolutional neural networks. In: Proceedings of the 2014 IEEE conference on computer vision and pattern recognition; 2013. p. 1717–24.
36. Tommasi T, Orabona F, Caputo B. Safety in numbers: learning categories from few examples with multi model knowledge transfer. In: 2010 IEEE conference on computer vision and pattern recognition; 2010. p. 3081–8.
37. Duan L, Xu D, Chang SF. Exploiting web images for event recognition in consumer videos: a multiple source domain adaptation approach. In: IEEE 2012 conference on computer vision and pattern recognition; 2012. p. 1338–45.
38. Yao Y, Doretto G. Boosting for transfer learning with multiple sources. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition; 2010. p. 1855–62.
39. Xia R, Zong C, Hu X, Cambria E. Feature ensemble plus sample selection: Domain adaptation for sentiment classification. IEEE Intell Syst. 2013;28(3):10–8.
40. Duan L, Xu D, Tsang IW. Domain adaptation from multiple sources: a domain-dependent regularization approach. IEEE Trans Neural Netw Learn Syst. 2012;23(3):504–18.
41. Zhong E, Fan W, Peng J, Zhang K, Ren J, Turaga D, Verscheure O. Cross domain distribution adaptation via kernel mapping. In: Proceedings of the 15th ACM SIGKDD; 2009. p 1027–36.
42. Chelba C, Acero A. Adaptation of maximum entropy classifier: little data can help a lot. Comput Speech Lang. 2004;20(4):382–99.

43. Wu X, Xu D, Duan L, Luo J. Action recognition using context and appearance distribution features. In: IEEE 2011 conference on computer vision and pattern recognition; 2011. p. 489–96.
44. Vedaldi A, Gulshan V, Varma M, Zisserman A. Multiple kernels for object detection. In: 2009 IEEE 12th international conference on computer vision; 2009. p. 606–13.
45. Vapnik V. Principles of risk minimization for learning theory. Adv Neural Inf Process Syst. 1992;4:831–8.
46. Borgwardt KM, Gretton A, Rasch MJ, Kriegel HP, Schölkopf B, Smola AJ. Integrating structured biological data by kernel maximum mean discrepancy. Bioinformatics. 2006;22 (4):49–57.
47. Yang J, Yan R, Hauptmann AG. Cross-domain video concept detection using adaptive SVMs. In: Proceedings of the 15th ACM international conference on multimedia; 2007. p. 188–97.
48. Jiang W, Zavesky E, Chang SF, Loui A. Cross-domain learning methods for high-level visual concept classification. In: IEEE 2008 15th international conference on image processing; 2008. p. 161–4.
49. Si S, Tao D, Geng B. Bregman divergence-based regularization for transfer subspace learning. IEEE Trans Knowl Data Eng. 2010;22(7):929–42.
50. Belkin M, Niyogi P, Sindhwani V. Manifold regularization: a geometric framework for learning from examples. J Mach Learn Res Arch. 2006;7:2399–434.
51. Ling X, Dai W, Xue GR, Yang Q, Yu Y. Spectral domain-transfer learning. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining; 2008. p. 488–96.
52. Quanz B, Huan J. Large margin transductive transfer learning. In: Proceedings of the 18th ACM conference on information and knowledge management; 2009. p. 1327–36.
53. Xiao M, Guo Y. Semi-supervised kernel matching for domain adaptation. In: Proceedings of the twenty-sixth AAAI conference on artificial intelligence; 2012. p. 1183–89.
54. Steinwart I. On the influence of the kernel on the consistency of support vector machines. JMLR. 2001;2:67–93.
55. Chung FRK. Spectral graph theory. In: Number 92 in CBMS regional conference series in mathematics. American Mathematical Society, Published by AMS; 1994.
56. 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; 2008. p. 1096–103.
57. Li S, Zong C. Multi-domain adaptation for sentiment classification: using multiple classifier combining methods. In: Proceedings of the conference on natural language processing and knowledge engineering; 2008. p. 1–8.
58. Gopalan R, Li R, Chellappa R. Domain adaptation for object recognition: an unsupervised approach. Int Conf Comput Vis. 2011;2011:999–1006.
59. Weinberger KQ, Saul LK. Distance metric learning for large margin nearest neighbor classification. JMLR. 2009;10:207–44.
60. LeCun Y, Bottou L, HuangFu J. Learning methods for generic object recognition with invariance to pose and lighting. In: Proceedings of the 2004 IEEE computer society conference on computer vision and pattern recognition; 2004, vol. 2, p. 97–104.
61. Marszalek M, Schmid C, Harzallah H, Van de Weijer J. Learning object representations for visual object class recognition. In: Visual recognition challenge workshop ICCV; 2007. p. 1–10.
62. Song Z, Chen Q, Huang Z, Hua Y, Yan S. Contextualizing object detection and classification. IEEE Trans Pattern Anal Mach Intell. 2011;37(1):13–27.
63. Cawley G. Leave-one-out cross-validation based model selection criteria for weighted LS-SVMs. In: IEEE 2006 international joint conference on neural network proceedings; 2006. p. 1661–68.
64. Tommasi T, Caputo B. The more you know, the less you learn: from knowledge transfer to one-shot learning of object categories. In: BMVC; 2009. p. 1–11.

65. Lowe DG. Distinctive image features from scale-invariant keypoints. Int J Comput Vision. 2004;60(2):91–110.
66. Wang H, Klaser A, Schmid C, Liu CL. Action recognition by dense trajectories. In: IEEE 2011 conference on computer vision and pattern recognition; 2011. p. 3169–76.
67. 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.
68. Schweikert G, Widmer C, Schölkopf B, Rätsch G. An empirical analysis of domain adaptation algorithms for genomic sequence analysis. Adv Neural Inf Process Syst. 2009;21:1433–40.
69. Dai W, Yang Q, Xue GR, Yu Y. Boosting for transfer learning. In: Proceedings of the 24th international conference on Machine learning; 2007. p. 193–200.
70. Hu M, Liu B. Mining and summarizing customer reviews. In: Proceedings of the 10th ACM SIGKDD international conference on knowledge discovery and data mining; 2004. p. 168–77.
71. Qiu G, Liu B, Bu J, Chen C. Expanding domain sentiment lexicon through double propagation. In: Proceedings of the 21st international joint conference on artificial intelligence; 2009. p. 1199–204.
72. Jakob N, Gurevych I. Extracting opinion targets in a single and cross-domain setting with conditional random fields. In: Proceedings of the 2010 conference on empirical methods in NLP; 2010. p. 1035–45.
73. Xia R, Zong C. A POS-based ensemble model for cross-domain sentiment classification. In: Proceedings of the 5th international joint conference on natural language processing; 2011. p. 614–622.
74. Shi X, Liu Q, Fan W, Yu PS, Zhu R. Transfer learning on heterogeneous feature spaces via spectral transformation. IEEE Int Conf Data Mining. 2010;2010:1049–54.
75. 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; 2011. p. 297–306.
76. Li W, Duan L, Xu D, 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.
77. Wei B, Pal C. Heterogeneous transfer learning with RBMs. In: Proceedings of the twenty-fifth AAAI conference on artificial intelligence; 2011. p. 531–36.
78. Ham JH, Lee DD, Saul LK. Learning high dimensional correspondences from low dimensional manifolds. In: Proceedings of the twentieth international conference on machine learning; 2003. p. 1–8.
79. 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; 2009, vol. 1. p. 1–9.
80. Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R. Indexing by latent semantic analysis. J Am Soc Inf Sci. 1990;41:391–407.
81. Raina R, Battle A, Lee H, Packer B, Ng AY. Self-taught learning: transfer learning from unlabeled data. In: Proceedings of the 24th international conference on machine learning; 2007. p. 759–66.
82. Wang G, Hoiem D, Forsyth DA. Building text Features for object image classification. IEEE Conf Comput Vis Pattern Recognit. 2009;2009:1367–74.
83. Dai W, Chen Y, Xue GR, Yang Q, Yu Y. Translated learning: transfer learning across different feature spaces. Adv Neural Inf Process Syst. 2008;21:353–60.
84. Bay H, Tuytelaars T, Gool LV. Surf: speeded up robust features. Comput Vis Image Underst. 2006;110(3):346–59.
85. Kloft M, Brefeld U, Sonnenburg S, Zien A. Lp-norm multiple kernel learning. J Mach Learn Res. 2011;12:953–97.

86. Shawe-Taylor J, Cristianini N. Kernel methods for pattern analysis. Cambridge: Cambridge University Press; 2004.
87. Davis J, Kulis B, Jain P, Sra S, Dhillon I. Information theoretic metric learning. In: Proceedings of the 24th international conference on machine learning; 2007. p. 209–16.
88. Saenko K, Kulis B, Fritz M, Darrell T. Adapting visual category models to new domains. Comput Vis ECCV. 2010;6314:213–26.
89. Ando RK, Zhang T. A framework for learning predictive structures from multiple tasks and unlabeled data. J Mach Learn Res. 2005;6:1817–53.
90. Gao K, Khoshgoftaar TM, Wang H, Seliya N. Choosing software metrics for defect prediction: an investigation on feature selection techniques. J Softw Pract Exp. 2011;41 (5):579–606.
91. Shivaji S, Whitehead EJ, Akella R, Kim S. Reducing features to improve code change-based bug prediction. IEEE Trans Software Eng. 2013;39(4):552–69.
92. He P, Li B, Ma Y. Towards cross-project defect prediction with imbalanced feature sets; 2014. arXiv preprint arXiv:1411.4228.
93. Chen M, Xu ZE, Weinberger KQ, Sha F. Marginalized denoising autoencoders for domain adaptation. ICML; 2012. arXiv preprint arXiv:1206.4683.
94. Vinokourov A, Shawe-Taylor J, Cristianini N. Inferring a semantic representation of text via crosslanguage correlation analysis. Adv Neural Inf Process Syst. 2002;15:1473–80.
95. Ngiam J, Khosla A, Kim M, Nam J, Lee H, Ng AY. Multimodal deep learning. In: The 28th International conference on machine learning; 2011. p. 689–96.
96. Yang L, Jing L, Yu J, Ng MK. Learning transferred weights from co-occurrence data for heterogeneous transfer learning. In: IEEE transaction on neural networks and learning systems; 2015. p. 1–14.
97. 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; 2012. p. 1–9.
98. Rosenstein MT, Marx Z, Kaelbling LP, Dietterich TG. To transfer or not to transfer. In: Proceedings NIPS'05 workshop, inductive transfer, 10 years later; 2005. p. 1–4.
99. Eaton E, desJardins M, Lane R. Modeling transfer relationships between learning tasks for improved inductive transfer. Proc Mach Learn Knowl Discov Databases. 2008;5211:317–32.
100. Ge L, Gao J, Ngo H, Li K, Zhang A. On handling negative transfer and imbalanced distributions in multiple source transfer learning. In: Proceedings of the 2013 SIAM international conference on data mining; 2013. p. 254–71.
101. Luo P, Zhuang F, Xiong H, Xiong Y, He Q. Transfer learning from multiple source domains via consensus regularization. In: Proceedings of the 17th ACM conference on information and knowledge management; 2008. p. 103–12.
102. Gao J, Liang F, Fan W, Sun Y, Han J. Graph based consensus maximization among multiple supervised and unsupervised models. Adv Neural Inf Process Syst. 2009;22:1–9.
103. Seah CW, Ong YS, Tsang IW. Combating negative transfer from predictive distribution differences. IEEE Trans Cybern. 2013;43(4):1153–65.
104. Moreno O, Shapira B, Rokach L, Shani G. TALMUD—transfer learning for multiple domains. In: Proceedings of the 21st ACM international conference on Information and knowledge management; 2012. p. 425–34.
105. Cao B, Liu N, Yang Q. Transfer learning for collective link prediction in multiple heterogeneous domains. In: Proceedings of the 27th international conference on machine learning; 2010. p. 159–66.
106. Li B, Yang Q, Xue X. Can movies and books collaborate? Cross-domain collaborative filtering for sparsity reduction. In: Proceedings of the 21st international joint conference on artificial intelligence; 2009. p. 2052–57.
107. Li B, Yang Q, Xue X. Transfer learning for collaborative filtering via a rating-matrix generative model. In: Proceedings of the 26th annual international conference on machine learning; 2009. p. 617–24.

108. Pan W. Xiang EW, Liu NN, Yang Q. Transfer learning in collaborative filtering for sparsity reduction. In: Twenty-fourth AAAI conference on artificial intelligence, vol. 1; 2010. p. 230–5.
109. Zhang Y, Cao B, Yeung D. Multi-domain collaborative filtering. In: Proceedings of the 26th conference on uncertainty in artificial intelligence; 2010. p. 725–32.
110. Pan W, Liu NN, Xiang EW, Yang Q. Transfer learning to predict missing ratings via heterogeneous user feedbacks. In: Proceedings of the 22nd international joint conference on artificial intelligence; 2011. p. 2318–23.
111. Roy SD, Mei T, Zeng W, Li S. Social transfer: cross-domain transfer learning from social streams for media applications. In: Proceedings of the 20th ACM international conference on multimedia; 2012. p. 649–58.
112. Jiang M, Cui P, Wang F, Yang Q, Zhu W, Yang S. Social recommendation across multiple relational domains. In: Proceedings of the 21st ACM international conference on information and knowledge management; 2012. p. 1422–31.
113. Zhao L, Pan SJ, Xiang EW, Zhong E, Lu Z, Yang Q. Active transfer learning for cross-system recommendation. In: Proceedings of the 27th AAAI conference on artificial intelligence; 2013. p. 1205–11.
114. Rajagopal AN, Subramanian R, Ricci E, Vieriu RL, Lanz O, Ramakrishnan KR, Sebe N. Exploring transfer learning approaches for head pose classification from multi-view surveillance images. Int J Comput Vision. 2014;109(1–2):146–67.
115. Ma Y, Gong W, Mao F. Transfer learning used to analyze the dynamic evolution of the dust aerosol. J Quant Spectrosc Radiat Transfer. 2015;153:119–30.
116. Xie M, Jean N, Burke M, Lobell D, Ermon S. Transfer learning from deep features for remote sensing and poverty mapping. In: Proceedings 30th AAAI conference on artificial intelligence; 2015. p. 1–10.
117. Ogoe HA, Visweswaran S, Lu X, Gopalakrishnan V. Knowledge transfer via classification rules using functional mapping for integrative modeling of gene expression data. BMC Bioinformatics. 2015;16:1–15.
118. Perlich C, Dalessandro B, Raeder T, Stitelman O, Provost F. Machine learning for targeted display advertising: transfer learning in action. Mach Learn. 2014;95:103–27.
119. Kan M, Wu J, Shan S, Chen X. Domain adaptation for face recognition: targetize source domain bridged by common subspace. Int J Comput Vis. 2014;109(1–2):94–109.
120. Farhadi A, Forsyth D, White R. Transfer learning in sign language. In: IEEE 2007 conference on computer vision and pattern recognition; 2007. p. 1–8.
121. Widmer C, Ratsch G. Multitask learning in computational biology. JMLR. 2012;27:207–16.
122. Wiens J, Guttag J, Horvitz EJ. A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions. J Am Med Inform Assoc. 2013;21(4):699–706.
123. Romera-Paredes B, Aung MSH, Pontil M, Bianchi-Berthouze N, Williams AC de C, Watson P. Transfer learning to account for idiosyncrasy in face and body expressions. In: Proceedings of the 10th international conference on automatic face and gesture recognition (FG); 2013. p. 1–6.
124. Deng J, Zhang Z, Marchi E, Schuller B. Sparse autoencoder based feature transfer learning for speech emotion recognition. In: Humaine association conference on affective computing and intelligent interaction; 2013. p. 511–6.
125. Zhang Y, Yeung DY. Transfer metric learning by learning task relationships. In: Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining; 2010. p. 1199–208.
126. Patel VM, Gopalan R, Li R, Chellappa R. Visual domain adaptation: a survey of recent advances. IEEE Signal Process Mag. 2014;32(3):53–69.
127. Shao L, Zhu F, Li X. Transfer learning for visual categorization: a survey. IEEE Trans Neural Netw Learn Syst. 2014;26(5):1019–34.
128. Bolt Online Learning Toolbox. http://pprett.github.com/bolt/. Accessed 4 Mar 2016.
129. Zhu Y. http://www.cse.ust.hk/∼yinz/. Accessed 4 Mar 2016.

130. BoChen90 Update TrAdaBoost.m. https://github.com/BoChen90/machine-learning-matlab/blob/master/TrAdaBoost.m. Accessed 4 Mar 2016.
131. EasyAdapt.pl.gz (Download). http://hal3.name/easyadapt.pl.gz Accessed 4 Mar 2016.
132. HFA_release_0315.rar (Download). https://sites.google.com/site/xyzliwen/publications/HFA_release_0315.rar. Accessed 4 Mar 2016.
133. Computer Vision and Learning Group. http://vision.cs.uml.edu/adaptation.html. Accessed 4 Mar 2016.
134. Guo-Jun Qi's Publication List. http://www.eecs.ucf.edu/∼gqi/publications.html. Accessed 4 Mar 2016.
135. Duan L. http://www.lxduan.info/#sourcecode_hfa. Accessed 4 Mar 2016.
136. Gong B. http://www-scf.usc.edu/∼boqinggo/. Accessed 4 Mar 2016.
137. Long MS—Tsinghua University. http://ise.thss.tsinghua.edu.cn/∼mlong/. Accessed 4 Mar 2016.
138. Papers:oquab-2014. http://leon.bottou.org/papers/oquab-2014. Accessed 4 Mar 2016.
139. Transfer Learning Resources. http://www.cse.ust.hk/TL/. Accessed 4 Mar 2016.
140. Heterogeneous Defect Prediction. http://www.slideshare.net/hunkim/heterogeneous-defect-prediction-esecfse-2015. Accessed 4 Mar 2016.
141. LIBSVM—A library for support vector machines. http://www.csie.ntu.edu.tw/∼cjlin/libsvm. Accessed 4 Mar 2016.
142. Domain Adaptation Project. https://www.eecs.berkeley.edu/∼jhoffman/domainadapt/. Accessed 4 Mar 2016.
143. Tutorial on domain adaptation and transfer learning. http://tommasit.wix.com/datl14tutorial. Accessed 4 Mar 2016.
144. A literature survey on domain adaptation of statistical classifiers. http://sifaka.cs.uiuc.edu/jiang4/domain_adaptation/survey/da_survey.html. Accessed 4 Mar 2016.
145. Exploiting web images for event recognition in consumer videos: a multiple source domain adaptation approach. http://lxduan.info/papers/DuanCVPR2012_poster.pdf. Accessed 4 Mar 2016.