



Heterogeneous transfer learning techniques for machine learning

Muhammad Shahid Iqbal¹ · Bin Luo¹ · Tamoor Khan² · Rashid Mehmood³ · Muhammad Sadiq⁴

Received: 8 October 2017 / Accepted: 17 December 2017 / Published online: 9 January 2018
© Springer International Publishing AG, part of Springer Nature 2018

Abstract

The main assumption in machine learning and data mining is, training the data, and the future data have the same distribution and same features. However, in many applications, in the real world, such assumptions may not be retained. For example, sometimes, we have the task of classification in the one domain of interest, but when the same data is used in another domain, it needed enough training to work in the other domain of interest. In the field of heterogeneous transfer learning, train the data in one domain and test with other domain. In this case, knowledge is transfer; if there is a successful transfer, it can significantly improve performance by avoiding the learning in the label information which is more expensive. Over the past few years, the transfer learning has become a new learning framework to address this issue and heterogeneous transfer learning is the most active research area in the recent years. In this study, we are discussing the relationship between heterogeneous transfer learning and the other machine learning methods, including the field of adaptation, learning and multitasking learning and sample selection bias, as well as the associates of variables. We also reconnoiter some main challenges for the future issue in heterogeneous transfer learning.

Keywords Transfer learning · Heterogeneous · Symmetric transfer · Asymmetric transfer

1 Introduction

In heterogeneous transfer learning, the source and target domains represented different features spaces; there are so many applications, where the heterogeneous transfer learning is applicable and useful, like in the following areas, image recognition [1–6], the multi languages text classification [1,6–10], single languages text classification [11], drugs classification [4], the human activity classification [12], and software defect classification [13].

It is also applicable in the big data area and as the repository of big data is more available, these repositories (abundant resources) is used for learning task for the machine learning,

which save time and the potential cost for new data (data collection). In the target domain, a set of data is available and it has different feature spaces from the target dataset. Heterogeneous transfer learning is a bridge between different feature spaces and it built a predictive model; the main work of predictive model is to predict the target domain.

Heterogeneous transfer learning is still a relatively new area of research, most of the work published on this area was in the last 6–7 years. There are two type of heterogeneous transfer learning and the first approach is symmetric; in the first approach, it transfers the source domain separately and target domain in the common features space, which unifies the input space domain. The second approach is called asymmetric transformation, it aligns the input feature spaces of the source domain to the target domain, then it transfers the features. It is the best approach, as the source and target class have the same instances and it transformed without bias function.

The assumption about heterogeneous transfer learning is that the source and target domain instance will be drawn from the same domain. When the differences between the function spaces are resolved and they have no need for further domain adaptation, if in the case of homogenous transfer learning, the source and target domains contain the labeled data drive

✉ Muhammad Shahid Iqbal
nawabishahid@yahoo.com

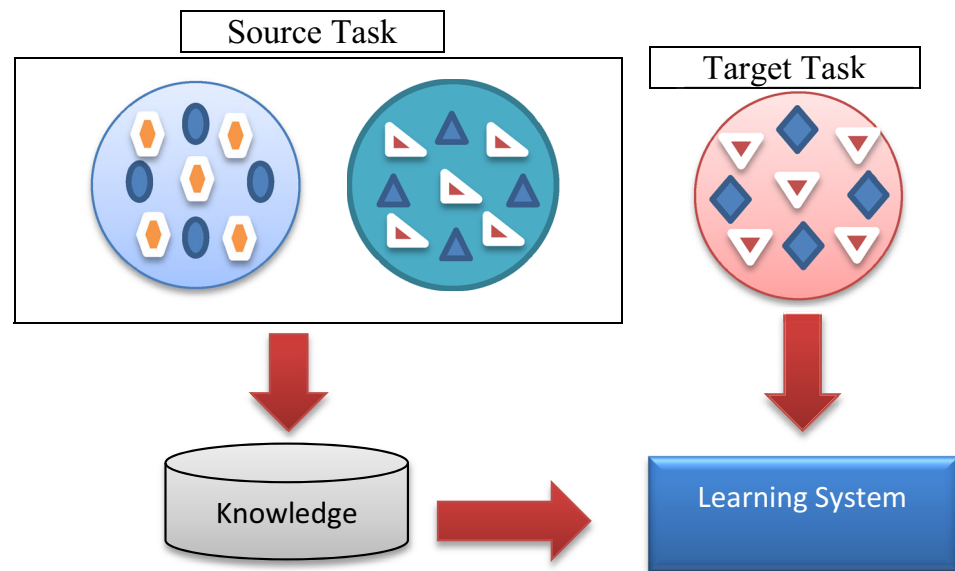
¹ School of Computer Science and Technology, Anhui University, Hefei, People's Republic of China

² School of Economic Information Engineering, Southwestern University of Finance and Economics, Chengdu, People's Republic of China

³ College of Information Science and Technology, Beijing Normal University, Beijing 100875, China

⁴ CRIDS (Center for Research in Distributed and Supercomputing) RIU, Islamabad, Pakistan

Fig. 1 Heterogeneous transfer learning (source and target domains have different feature spaces)



solution, and the solution is used for heterogeneous transfer learning and the available labeled data are a good basis for the transfer learning application. In this article, we have different types of data, labeled and unlabeled; in the form of labeled data, the source and target domain are related, and it is more feasible, and therefore the heterogeneous solution requires a clear correlation between the source and target. For example, the solution definition for Prettenhofer and Stein [8] and Wei and Pal [14] needs to be defined in accordance with the manual source and target. Figure 1 shows the heterogeneous transfer learning.

The rest of this paper is organized as follows: the second section is about heterogeneous transfer learning applications, Sect. 3 gives the heterogeneous transfer learning techniques, Sect. 4 describes the research challenges, and conclusions are drawn in Sect. 5.

2 Heterogeneous transfer learning applications

The first application of heterogeneous transfer learning is image classification and video recognition. Yang et al. [15] has proposed a learning algorithm for heterogeneous transfer, the image clustering lever annotated images auxiliary, their goal is to obtain images auxiliary annotated for the target image classification. Dai et al. [16] used learning translated, text label data to help classify the images, while in this work the data text is not auxiliary and labeled. Dai's work also focuses on the multimedia field, in particular, the works using the text and the image together, for example, the exploitation of the content of the image for the web [17], and both share the same consensus that verified that finding the correlation between the images and the text is essential to the under-

standing of the images. However, Wu et al., method is a new method, were in they used image and text from different sources and their work is also linked to the marked images, e.g., Wu et al. [18]. The image classification and video recognition have further application types, image segmentations and clustering.

The second application is human activity classification; Sargano et al. recognized the human activity on the base of pre-train deep CNN model, feature extraction and representation followed by a hybrid support vector machine (SVM) and K -nearest neighbor (KNN) classifier for activity recognition [19].

The third application is software defect classification; Nam and Kim propose heterogeneous defect prediction (HDP) to predict the defects across projects with heterogeneous metric set [13]. The fourth application of heterogeneous transfer learning is cardiac arrhythmia classification; the fifth application is multi language text classification; authors assume back of word (BoW) document representations x and linear classifiers w [8], and the sixth application of heterogeneous transfer learning is drug efficacy classification, Fig. 4 shows the applications of heterogeneous transfer learning.

Wu et al. combined the multi-source domains, the source are related to target domain and labeled data are used, for increased the performance in target domain, used combine data, so the experiment is conducted on real data set [20]. Yang et al. used transfer weight to evaluate the relatedness among the domain, in this method to compute the principal components of each feature space, in the each feature co-occur data represent in principal component and then applied Markov chain Monte Carlo method for cyclic network, and also the edge weighted of the network was employed, as a

Fig. 2 Symmetric transfer (features base transfer)

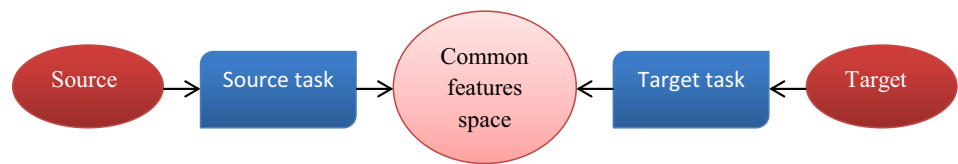


Fig. 3 Asymmetric transfer (features base transfer)

transfer weight from source domain to the target domain, in many existing heterogeneous methods, the weighted values are used as prior setting parameters, this parameter is used to control the transferring parameter [21]. Li et al. analyzed the remote sensing images, and used heterogeneous space for transfer learning problem, the author proposed the iterative reweighting heterogeneous transfer learning method, it learns the common space from source and target data, to conducted reweighted policy, and it used two projection function (SVM) to map the common space, the source data are reweighted on the base of common subspace and reused for transferring [22].

Tong et al., proposed the method, canonical correlation analysis and restricted Boltzmann machine (MCR), which address the data of different companies; the first main thing in this method is a unified metric, which estimated the effort of heterogeneous data and the second main thing is combining the MCR with estimated effort in heterogeneous cross company effort [23]. Pan et al used the heterogeneous one class collaborative filtering, this novel algorithm is called transfer via joint similarity learning (TJSL), this method jointly learn the similarity of candidate and preferred item [24]. Xue et al., proposed a method task selection machine (TSM), it used features mapping method and it is also useful for the unsupervised multi task learning area [25].

3 Heterogeneous transfer learning techniques

Heterogeneous transfer learning has two main techniques and the first technique is symmetric transfer learning and the second is asymmetric transfer learning, the Figs. 2 and 3 show both types of techniques.

3.1 Symmetric transfer learning

Prettenhofer and Stein proposed transfer learning [8] which deals with the heterogeneity of a scenario containing source

domain of labeled data and the target is not labeled. Blitzer et al. proposed learning techniques for structural correspondence, which is applied on this problem [26]. It learns the structure correspondence of source and target domain, for good prediction quality, used pivot function and the pivot function used for features identification, occurs frequently in two domains, in source and target data, the each pivot function transformed using linear classifier. It learns the characteristic between the correspondence elements and finally, it used latent function space for the final classifier. Prettenhofer and Stein [8] uses this solution to resolve the problem of classification of text, where the source is in one language and the target is in a different language.

This implementation is called language correspondence (structural learning) and the pivot function is defined as two different pairs, one is objective and the other is source, which translate words from one language to others. The experiments are carried out on document classification, for source documents, English language is used and the other languages are used for target documents. It trains a learner from the wording of the source documents, on basis of the source, it translates the target document in the source language and finally, it tests the translated version of the document. Through this, a top method is established to learn the labeled target documents and test them. The classification is measured as the measure of performance. The results shows that the Prettenhofer and Stein [8] method has a better result compared to other baseline methods. The pivot function has one problem in correspondence structure, it is difficult to make generalized, and it must be unique and manually defined for specific applications. It is difficult to import on other applications or portables.

The article by Shi et al. [4], that is called the spectral mapping heterogeneous (HeMap), addresses the specific scenario of transfer learning, where the entry of the feature space is different between the source and the target $X_S \neq X_T$, the marginal distribution is different between the source and the target ($P(X) \neq P(X_T)$), and the space of output is different between the source and the target ($Y_S \neq Y_T$). This solution uses the source data as label and target domain have limited labeled data in the first step which is to find a common space between the inputs latent source and target domains using a technique of spectral mapping. The method of spectral mapping is to optimization objectives and to retain its original structure of data, while reducing to minimum differences in the two areas. The next step is to use this way to the

selected sample-based clustering, select the relevant bodies such as the training of new information on the marginal distribution and also address the different input spaces latent. For this, a Bayesian method is used to find the relationships and resolve the issues between different output areas such as Image classification and drug anticipation. The performance is measured in term error rate. It gives a better performance as compared to a baseline approach, even in this article the basic approach are not documents, also not tested the learning solution of the test of transfer.

Wang and Mahadevan [11] algorithm is called filed alignment manifold adaptation, which used alignment manifold [27]. This method is used for symmetric transformation for entry space of the domain. Several areas in this solution, the source called limited target area and the area K for output of the space, it share the same label, this approach created function for separate matching, in each domain to transfer input space, in common entry space, but it preserve each area structure and each domain modeled is consider as collector. For input to create a space latent, the entry of all the European Union areas created a large matrix model and this matrix captures the representatives of European Union area entries. In this model, a Laplace matrix represents each domain, which captures instances, which share the same labels for and, force to be neighbor, while the different labels are separating.

A reduction of the dimensionality step is performed by a process of decomposition to the values own generalized to eliminate the redundancy of function. The last student is constructed two stages. The first one is called “linear regression model formed on the source data and it used the latent function of the space”. The second step is also a linear regression model which is summarized in the first step and the second step used a regularization of the manifold [28]. The processes are used to confirm that the errors of prediction are reduced, when it used the label data for target. The first step is training the data which is used as data source and the second stage compensates for differences in the field produced by the first step to achieve better target forecasts.

In this experiment mainly focused on the classification of text documents and the classification accuracy is measured in term of performance. The techniques were tested against: an approach to the analysis of canonical correlations and the approach for the regularization of the collector, to find out which is considered as the reference technique. The base technique uses the data from the target domain called limited and it does not use this information in the source domain. The method presented in this article is much more of the canonical analysis of correlations and the base line; however, these methods are not directly mentioned, so it is problematic to cognize the significance of the results of the test. A sole aspect of this study is the demonstrating the source of multiple areas of the heterogeneous solution. The main reason is that the large amount of unlabeled heterogeneous data is eas-

ily available, and for a particular target, it is used to predict the performance of the target and improve the performance.

The article of Zhu et al. [3], which presents a technique for the classification of image heterogeneous transfer learning for image classification (HTLIC); in this scenario, the main assumption is that target data is sufficiently important. The main objective is used large amount of source data and the source data created a common latent space, in target domain it will improve the performance target classifier. The solution recommended by Zhu et al. [3] is closely coupled to the application of the classification of image and it is described as follows, the images with the reference categories (like cake, dog, star, etc.) which are available in target domain, for basic data, search on web, which is carried out by Flickr and the image data is available with the category like the word dog is in doggie. The idea Flickr used the image annotated for source and it was proposed for the first time by Yang et al. [15]. By Flickr concept, each image has several associated word tags, when the image is retrieved.

For full text document search using google search, these word tagged images are used. Then, two-part (bipartite graph) graph in two layers is built and the first layer of graph is the represented links between the tags and source image. The second layer of the graph is the represented links between the text document and images. If marker of image appears in text documents, then a link is created, otherwise no link created. In two images like the source and the target, which are initially represented by a set of entry features derives information with the help of SIFT descriptor pixels [29]. Using the initial source as image and the graphic representation only bipartite graph derived the source of the images and the text data, the tags for a latent semantic is learned by employing the analysis latent semantic [30]. A learner is now formed with the label transformed. Target experiments are carried out on the proposed approach, where 19 different categories of images are selected. The binary classification is performed to test different pairs of category of the image. Once a reference technique is done with the help of a classifier SVM, then it is trained with the help of label target data.

The methods of Raina et al. [31] and Wang et al. [32] were tested. The Zhu proposed method is [3] and it is best followed of Raina et al. [31] and Wang et al. [32] and also from the basic method. A very attractive principle to improve the performance for the used abundant source of data and which is not available for performing the search on internet. The image classification through website in this method is very specific like the Flickr method, and the Flickr method covers the image data. In other applications, it is difficult to import this method. Qi et al. [5] proposed a method for transfer learning, in this method, it deals with the image classification. Qi et al. [5] the image classification is more difficult compared to text classification, because it is not directly inherent semantic class of label. Image features derived from the pixel informa-

tion that is not semantically linked to class of the labels is by opposition to the floor functions of semantic interpretation to the class of labels. In addition, the image data labeled is rarer by report to the mention of data like (text data). So, the image classification using transfer learning, which required the abundance of data for source in the form of text, which was used to improve the learning of the learner and it used a little amount of image data which is called target data. In this method, as a result of text classification, which are identified through a search of web document like (Wikipedia), the labels class.

When transferring the knowledge from text to image, from source to target and it makes a bridge between the text and image, the bridge is the matrix of concurrence, the matrix is used like a door between the text and image. Every image has their correspondences which contain the co-occurrence matrix (text, correspondence image). This method, based on the text to image (TTI), is alike to Zhu et al. [3]. But, Zhu et al. [3] for the improvement of the transfer of knowledge, does not use tagged source data, which will outcome in a degradation of enactment, when labeled target data is so little.

The experiments are carried out with the methods proposed by Qi et al. [5,16], Zhu et al. [3], so the basic method using a classifier SVM which are trained limited label target data. From Wikipedia, the full test documents are collected, so the performance is measured as term of rate of error classification. These results demonstrate that the Zhu et al. [3] method performs 15% better, then perform the best in 15% of these experimental judgments, the Dai et al. [16] method performed 10% better in experimental judgments and the Qi et al. [5] technique is prominent in 75% of the experimental judgments. For example, in Zhu's case [3], this method is very specific for the image classification and it is difficult to post other different fields or application. The area of high level of adaptation is shown in Fig. 2. The experiments are carried out for three applications which include the image. The Table 1 shows the analysis of symmetric transfer learning.

Shi et al., in this equation, used T as target data matrix and S as the source data matrix, B_S is source data of optimization objective, and B_T projection target data [4]. Prettenhofer and Stein, W^* is a vector, L is loss function and non-negative regularization parameter that penalizes the model complexity is $W^T W$ [8]. Wang and Mahadevan, in this paper, define W_s similarity matrix, W_d dissimilarity matrix, and the other matrices are define L , Z and Laplacian matrix L_s , L_d [11]. Zhu et al., $R(U, V$ and $W)$ is the regularization function, which control the complexity of the latent and V , U and W [3]. Yang et al., equation shows the f as image feature spaces and v as the image data set and w is text feature spaces [15]. Dai et al. equation shows where $P(y_t/x_t)$ is which is estimated using the feature extractor in the target feature space Y_t [16]. Wang et al., in this equation, selected randomly N pairs of images,

so S_i is i -th and f_j is a j -th pair, so the authors calculated the distance of i -th pair with f_j [32]. Daumé, in this equation, mapped the source and target, respectively, Φ^s and Φ^t ; so the x and y are input and output space [29]. Raina et al., in this equation, $x_u^{(i)}$ is unlabeled data, b is a base vector and a is the activation of b_j for the input of $x_u^{(i)}$ [31]. Qi et al., in this equation, λ and γ are balancing parameters, f_T is discriminant function and $l(\cdot)$ is loss function, so the second part of equation is sum taken C weighted co-occurrence; $X(\cdot)$ is decreasing function; when Z is larger than the output, it becomes $X(z)$ [5].

3.2 Asymmetric transfer learning

The exertion of Kulis et al. [2] that is called the cross-domain asymmetric regularized (ARC-t transformation) put forward a procedure of asymmetric transformation to resolve the heterogeneous function amongst the zones of the interplanetary. For this situation, there is a profusion of data source and target. Labeled data an inadequate objective function is mainly well-defined for the learning of the transformation matrix. The objective function contains a regularizer and long term function and therefore the cost of long-term function applies on inter-areas of each pair of instances and learned transformation matrix. The building of the objective function is accountable for the invariant of arena transformation process. The optimization of the objective function aims to minimize the cost function and regularize it. The transformation matrix is learned in a non-linear in the kernel space RBF Gaussian. The untaken technique is named as the cross-domain regularized asymmetrical transformation (CDRAT). Twofold experimentations are carried out through the help of this method for the classification of images wherever the accurateness of the classification is measured as the measure of performance.

There are 31 classes of the image well-defined on behalf of these experiments. In test 1, the first experience is when the cases of all the thirty-one classes of the image are encompassed in the basis and target the training data. In another experiment (test two), only 16 classes of the image are signified in the training target of data (all the 31 are represented as a source). To test contradiction of other basic approaches, a technique is essential to put the source and target data together. A preprocessing phase the kernel named is canonical correlation analysis which proposed by Shawe Taylor [33] is used for prominent the source and target domains in a common domain by means of the transformation of space symmetrical. The basic methods tested include the K -nearest neighbors, metric learning through SVM proposed by Davis et al. [34], to upsurge the function proposed by Daumé [29], and a technique of learning of the metrics area proposed by Saenko et al. [35].

Table 1 Analysis of symmetric transfer
Analysis of symmetric transfer learning

References	Data type Data set	Result		Data		Method Mechanism	Equations Equation	Description Explanation
		Accuracy		Source	Target			
Shi et al. [4]	Drug efficacy prediction data, image data set	This method reduced error rate 50%	Labeled	Unlabeled	Bayesian-based, HeMap	$\min_{B_T, B_S} l(B_T, T) + l(B_S, S) + \beta \cdot D(B_T, B_S)$	Used possible high link to produce better result Unify the source and target features spaces Simple selection method is applied on the relative source	
Blitzer et al. [26]	WSJ MEDLINE	88.9 72.0	Labeled	Unlabeled	SCL	$\hat{W}_l = \operatorname{argmin}_w (\sum_j L(w \cdot x_j, p_j)(x_j)) + \lambda \ w\ ^2$	Training and testing data from the same distribution Introducing correspondence between features The method is applied for on speech tagging and achieved better performance	
Prettenhofer and Stein [8]	Books Dvds Music	79.50 (± 0.33) 78.49 (± 0.03) 73.09 (± 0.07) 76.92 (± 0.07) 78.80 (± 0.01) 71.07 (± 0.02) 77.79 (± 0.02) 77.92 (± 0.03) 75 (± 0.06)	Unlabeled	Unlabeled	Cross-language text classification (CL-SCL)	$W^* = \operatorname{argmin}_{w \in R^{ \mathcal{V} }} \sum_{x, y \in D_S} L(y, w^T x) + \frac{\lambda}{2} \ w\ ^2$	Cross languages classification using structural learning Introduce the cross-semantic classification The analysis is based on the English as a base and other three languages as target (French, German and Japanese) cover performance	
Wang and Mahadevan [11]	Text	Average of 90 average precision scores	Labeled	Limited labels	DAMA manifold alignment for heterogeneous transfer	$Z(\mu L + L_S) Z^T x = \lambda Z L_d Z^T x$	For knowledge transfer, it constructs a map for linking different feature spaces across domain	

Table 1 continued

References		Data type		Result		Data		Method		Equations		Description	
	Data set			Accuracy		Source	Target		Mechanism	Equation		Explanation	
Analysis of symmetric transfer learning													
Categorization													
	TDT2												It used source data from different domains
	TREC												It combined the existing domain adaptation approach for learning the common features
Zhu et al. [3]	Unlabeled auxiliary text, tagged image data from a social web	71.5493	Unlabeled	Abundant labels				HTLIC		$\min_{U, V, W} \lambda \ G - UV^T\ _F^2 + (1 - \lambda) \ F - WV^T\ _2^2 + R(U, V, W)$ $R(U, V, W) = \gamma_1 \ U\ _F^2 + \gamma_2 \ V\ _F^2 + \gamma_3 \ W\ _F^2$	Transfer knowledge between text and image	It used source data from different domains	
Yang et al. [15]	Caltech-256	0.741 ± 0.018	Labeled	Unlabeled				aPLSA		$P(f w) = \sum_{z \in Z} P(f z)P(z w)$	Image from web and text, bridge between both, then transfer to text For efficient transfer, use matrix factorization Improve the performance	It used source data from different domains	
Dai et al. [16]	Caltech-256	91 ± 0.045, 0.122 ± 0.043, 0.253 ± 0.062, 0.247 ± 0.059, 0.183 ± 0.072	Labeled	Little labeled				TLRisk (translated learning via risk minimization)		$P(y_i \hat{\theta}_{x_i}) = \int_{x_i} P(y_i x_i)P(x_i \hat{\theta}_{x_i})dx_i'$	No correspondence between data instances Improve the transfer learning using auxiliary data Using label data of one source and learning the other domain	It used source data from different domains	
													Link between the two domains (make bridge between the domains)

Table 1 continued

Analysis of symmetric transfer learning						
References	Data type	Result	Data		Method	Description
			Source	Target		
Wang et al. [32]	PASCAL VOC 2006, PASCAL VOC 2007	Accuracy	Labeled	Unlabeled	TBIF (text-based image feature)	Features built using auxiliary image data with tags Text features obtain unannotated image form tag of KNN in this auxiliary collection
		0.8370				
		0.8173				
		Equation	$\sum_i \left(e^{-\sum_j w_j d_i^j - s_i} \right)^2$			
Daumé [29]	ACE-NER, CoNLL, PubMed, CNN, Tree bank-Chunk, Treebank-brown	Accuracy	Labeled	Limited label	AUGMENT FAM	Slight best enough target data to source
		0.8207				
		0.7436				
		0.8606				
Equation	$\Phi^s(x) = \langle x, x, 0 \rangle,$ $\Phi^t(x) = \langle x, 0, x \rangle$					
Raina et al. [31]	Handwritten Characters Font Characters Webpages UseNet	Accuracy	Unlabeled	Labeled	(STLSC) Self-taught learning through sparse coding	Method domain adaptation Trivial multi domain adaptation Self-taught learning Used sparse coding to construct high level features For classification used SVM Fisher kernel learns this representation
		41.8%				
		62.6%				
		Equation	$\text{minimize}_{b,a} \sum_i \ x_i^{(i)} - \sum_j a_j^{(i)} b_j\ _2^2 + \beta \ a^{(i)}\ _1$			
Qi et al. [5]	Flickr Wikipedia	Accuracy	Labeled	Limited labels	Proximal gradient solver (PGS) TTI	Transfer text to image Bridge between the two different domains Semantically connect two different domains
		0.2133 ± 0.0008				
		0.2435 ± 0.0009				
		Equation	$\min_T \gamma \sum_{j=1}^m l(y_j^i, f_T(x_j^i)) + \lambda \sum_c x^{(c,k,j)} \cdot T(x_k^{(s)}, x_j^{(t)})$			

For the test one, the Kulis et al. [2] performs a method slightly better than the other approaches verified. For the test two, the Kulis method [2] performs best in relation to the technique of the k -nearest neighbors. The Kulis et al. [2] method is better suitable for situations where all the classes are not signified in the target workout data as shown in the test two. The problem well-defined by Harel and Mannor [12] is of little amount of data that is labeled in target and source data are labeled when there is need for asymmetric transformation. The first step in this procedure is to regulate the functions of the source and target areas, and then the instances of group by class in the area's source and target. For each class, the functions are set to zero mean. Then, each source class group is paired with the group target class consistent, and then a process of singular values decomposition (SVD) is performed to find the transformation matrix that is specific to this class of grouping.

When the processing is carried out, the features are transported to reverse the previous step and the concluding objective classifier is molded using the transformed data. Find the transformation matrix through SVD which permits you to marginal distributions in the class of clusters to be allied, while maintaining the structure of the data. This method is chosen under the name of the algorithm of mapping multiple outlooks. The experiments uses data from sensors. There are five different actions defined for the knowledge: run, walk, move down, up, and long. The source domain contains the same (but dissimilar) statements of the sensor in relation to the target.

The technique proposed by Harel and Mannor [12] is equated against a reference point technique that trains a classifier by the inadequate labeled target data and an upper bound technique that uses a meaningfully higher set of labeled objective data to train a classifier. A support vector machine (SVM) learner is used by way of the base classifier then a balanced inaccuracy rate (imbalance in the test data) stands measured by way of the performance metric. Harel and Mannor [12] method outperforms the baseline method in each test, then falls short of the upper bound technique in each trial with respect towards the balanced error rate. The heterogeneous transfer learning situation addressed by Zhou et al. [7] needs plenty of labeled source data and little amount of labeled target data. Then, an asymmetric transformation function is proposed which maps the source features to the target features. While learning transformation matrix, Ando and Zhang [36] adopted multi task learning method.

The result, mentioned as the sparse heterogeneous feature representation (SHFR) is implemented by making a binary classifier aimed at apiece class in the source and the target domains distinctly. To each binary classifier is allocated a weight term wherever the weight terms are learned by merging the weighted classifier outputs, whereas minimizing the classification error of each domain. The weight rappers are

now used toward finding the transformation matrix by minimizing the difference between the target weights and the transformed source weights. At the end, the final target classifier is trained by original data and transformed source data.

These experiments are carrying out on text classification; the target documents in one language and source documents is in different languages. A baseline scheme using a linear support vector machine (SVM) classifier trained on the labeled target is established along with testing against the methods proposed by Wang and Mahadevan [11], Kulis et al. [2], and Duan et al. [1]. The method proposed by Zhou et al. [7] 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. [1] method performed either second or third best. Asymmetric base transfer learning is shown in Fig. 3. Figure 4 shows some related domain of heterogeneous transfer learning, its application and current research trends. Figure 3 shows the asymmetric transfer learning and Table 2 shows the analysis of the asymmetric transfer learning.

Zhou and Dai, in this equation, C denote the candidate images and d is the dimension feature vector of each image and z denotes the point in feature spaces [17]. Wei and Pal, in this equation, x_s, x_t and x_s, x_t are the pivot features [10]. Kulis et al., in this equation point, have two points (x, y) and x is subset of A and y is subset of B , in the first part of equation, the x, y belong to same category and in second part of equation, the x, y belong to different categories [2]. Harel and Mannor, in this equation, $D_i^{(j)}$ is defined as the utilization matrix and $D_i^{(1)}$ and $D_i^{(2)}$ are concatenated matrix and h is the largest eigenvalue [12]. Wei and Pal, in this equation, H is hidden layer, and it must share the mapping H to Y ; so H is hidden layer and X_s and Y_t are the features, $P(x/y)$ marginal conditional probability [14]. Duan et al., in this equation, $L\epsilon(\cdot)$ is extensive loss function; f^T is decision vector; λ, C, θ are regularization parameters and $f(x)$ is target classifier [37]. Duan et al., in this equation, P and Q are projection matrices and W is weighted vector, C is regularization parameter [1]. Zhou et al., in this equation, D_T is target unlabeled data; D_S source label data; D_C cross-domain parallel label data; H_T, H_S high level features space, λ is regularization parameter and G_K transformation features [9]. Zhou et al., in this equation, b_i is concatenated row vector and n_C for all task [7]. Nam and Kim, in this equation, $P_{ij}(n)$ is comparison function percentiles and i -th is source and j -th is target matrix [13]. Liu et al., in this equation, the loss is define as triplet and the first part of equation is positive exemplar and second part of equation is negative exemplar, and there is constant distance amongst x_i^a and x_i^p . The positive is x_i^p and negative is x_i^n [39].

Table 2 Analysis of asymmetric transfer

References	Data type	Result	Data		Method	Equations	Description
			Source	Target			
Analysis of asymmetric transfer learning							
Zhou and Dai [17]	Brin and Page op ten Web images	Bigger improvement in 75% cases and the improve is 25%	Labeled	Little labeled	WEBSEIC	$\text{Density}(z) = \sum_{i=1}^{ c } e^{-\sum_{j=1}^{ c } z_j - x_{ij} ^2}$	Web searching and search the image content, which helps to improve the performance
Wei and Pal [10]	Labeled Chinese, labeled English, unlabeled Chinese	0.831, 0.837, 0.825, 0.828, 0.851	Labeled	Unlabeled	SCL	$(x_i, x_{ip}) \text{ and } (x_{sp}, x_t) \text{ where } x_{(s t)p}$	Candidate image and associated image Analyze the density of visual information and generate the base rank of web pages, then combine the keyword and produce the final result The annotated English to other languages
Kulis et al. [2]	Webcam, webcam-800, webcam-surf, dslr, dslr 600, dslr shift	0.790, 0.833, 0.817, 0.823, 0.857	Labeled	Limited labeled	ARC-t	$C_i(X^T W Y) = (\max(0, I - x^T W y))^2$ $C_i(X^T W Y) = (\max(0, x^T W y - u))^2$	To minimize the noise using translation P stands for pivot features Transferring of object to visual model
		37.4					45.0 24.8 32.5 34.8 20.6

Table 2 continued

References		Data type	Result	Data	Method	Equations	Description
Analysis of asymmetric transfer learning		Data set	Accuracy	Source	Mechanism	Equation	Explanation
Harel and Mannor [12]	Accelerometer, phototransistors, barometric, GPS	Labeled	P value < 0.05	Labeled	MOMAP	$\{R_i\}$ $= \operatorname{argmin}_{\{R\}} \sum_{i=1}^c \ R_i D_i^{(2)} - D_i^j\ _F^2$	Data will be different feature spaces, learned in the single task
Wei and Pal [14]	UIUC	Labeled	83.2% 0.13%	Labeled	RBM SCL	$R = \alpha \sum \log(p(\{y\} x_{vi})) + (1 - \alpha) \sum \log(p(\{y\} x_{tk}))$	Feature space is called outlook It takes data from all the outlooks and classified into a better outlook The target and source task are disjoint RBM (restricted Boltzmann machine) This model is probabilistic graphical From consumer video event recognition
Duan et al. [37]	NUS-WIDE	Labeled	83.0%, 0.10%	Labeled	DSM (domain selection machine)	$\min_{d, w, b, \beta, f^T} \frac{1}{2} (\ w\ ^2 + \lambda \ \beta\ ^2) + c \sum_{i=1}^m l_{\in}(f^T(x_i^T) - f(x_i^T)) + \frac{\theta}{2} \sum_{s=1}^d d_s \sum_{i=1}^m (f^T(x_i^T) - f^s(x_i^T))^2$	Consumer videos are not labeled Used SVM to find SHIFT features from image Then most relevant domains are selected A new regularization is defined SVR (support vector regularization)
Duan et al. [1]	Amazon Dslr Webcam	Unlabeled	55.4 ± 2.8 54.3 ± 3.7	Unlabeled	HFA	$\min_{P, Q} \min_{w, \xi, \xi'} \frac{1}{2} \ w\ + c \left(\sum_{s=1}^{d_{AS}} \xi + \sum_{t=1}^{d_{AT}} \xi' \right)$	Used two projection matrices From two domains, transformed data in the common subspace Proposed new matching feature functions

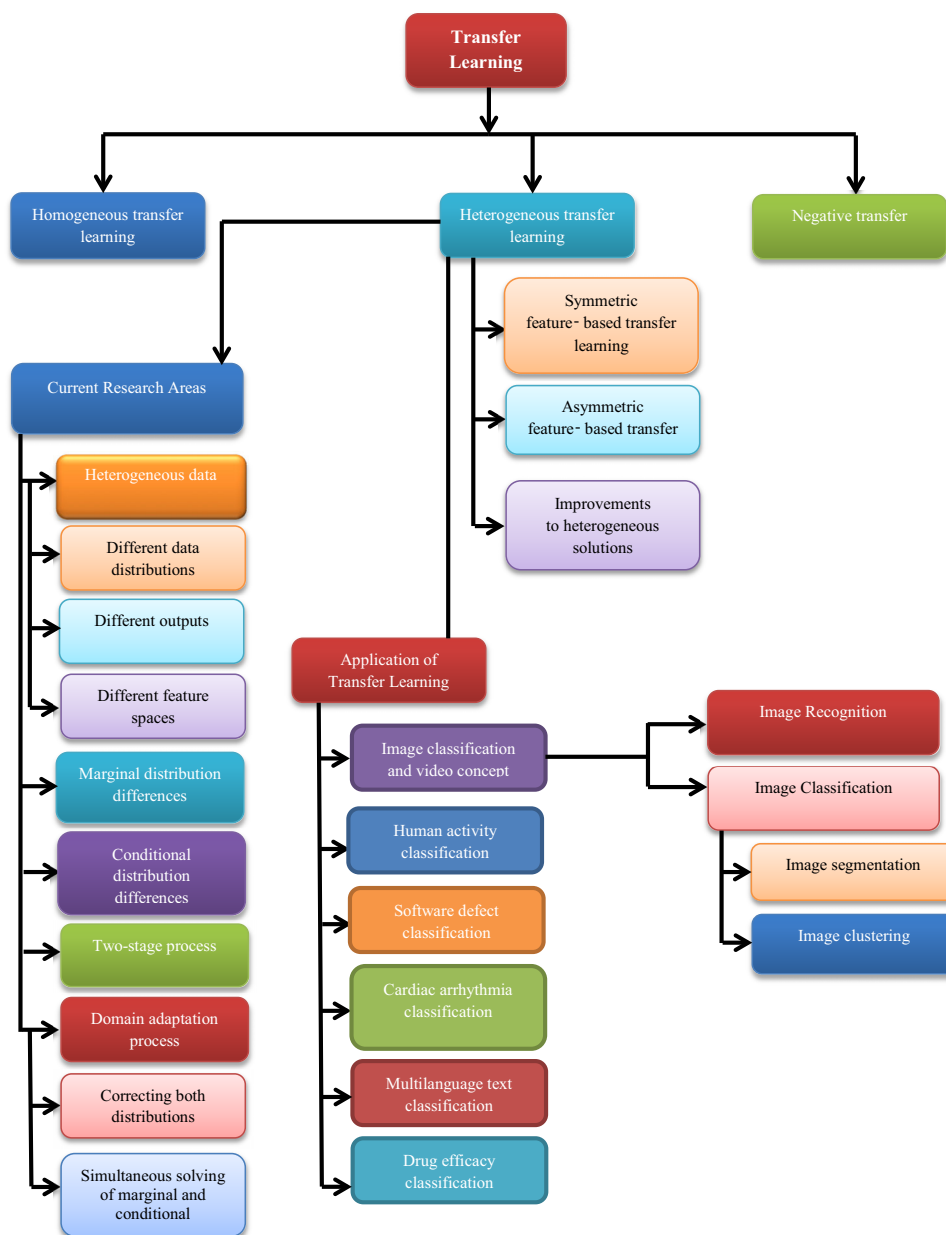
Table 2 continued

References		Data type	Result	Data	Method	Equations	Description
Data set		Source	Accuracy	Target	Mechanism	Equation	Explanation
Zhou et al. [9]	Prettenhofer and Stein Amazon product	Labeled	82.50	Unlabeled	HHTL (hybrid heterogeneous transfer learning)	$\min_{G_k} \ H_{S,k}^{(c)} - G_k H_{T,k}^{(c)}\ _F^2 + \lambda \ G_k\ _F^2$	It allows the cross-domain and used deep learning approach for matching features
			82.70				
			75.56				
Zhou et al. [7]	Multilingual Reuters collection	Labeled	72.79 ± 1.10	Limited labeled	SHFR	$\min_{g \geq 0} \sum_i^{d^r} (\frac{1}{n_c} \ b_i - Dg_i\ _2^2 + \lambda_i \ g_i\ _1)$	MCHDA (multi-class heterogeneous domain adaptation)
			73.82 ± 1.1				Reconstruct the sparse matrix for weighted vector map
			274.15 ± 1.14				Used sensor measurement, which is based on sensing theory
			73.35 ± 1.31				Used many binary classifiers for error correcting and output coding
	BBC		90.45 ± 1.00.				
			91.82 ± 0.84				
	Cross-lingual sentiment		62.09 ± 2.15				
			65.22 ± 2.03				
			62.05 ± 3.16				
Li et al. [38]	ImageNet	Limited number of labeled	0.1196	Unlabeled	HFA	$\min_{H, \alpha} \max_{\alpha} I_{ns+nt} - \frac{1}{2} (\alpha \alpha y)'$ $K_H (\alpha \alpha y)$	Features of different dimensions transferred in common subspace
			0.1321				Used bag-based reranking method to compute the similarity and rank query result

Table 2 continued

References	Data type	Result	Data		Method	Equations	Description
			Source	Target			
	NUS-WIDE						
	Text	0.2233					
		0.2561					
Nam and Kim [13]	EQ, JDT, LC, ML, PDE, Apache, Safe, ZXing, ant-1.3, arc, camel-1.0, poi-1.5, redactor, skarbonka, Tomcat, velocity-1.4, xalan-2.4, xerces-1.2 cm1, mw1, pc1, pc3, pc4, ar1, ar3, ar4, ar5, ar6	0.724*	Labeled	Unlabeled	HDP Heterogeneous defect prediction	$M_{i,j} = \frac{\sum_{k=1}^9 P_{i,j}(10 \times k)}{9}$	Used heterogeneous metric prediction and the prediction is based on metric selection and metric matching
Liu et al. [39]	NIR-VIS Face Databases	95.74 ± 0.52	Labeled	Unlabeled	TRIVET	$\text{Loss} = \sum_i^N [\ f(a_i^q) - f(a_i^p)\ _2^2 - \ f(a_i^q) - f(a_i^p)\ _2^2 + \alpha]$	Used VIS face image and for discriminative model used deep convolution neural network Then, transfer model to NIR-VIS, fine-tune the triplet losses It reduces the intera class and fine-tune possible small data set

Fig. 4 Heterogeneous transfer learning



4 Research challenges

Heterogeneous transfer learning is a relatively new research area and this area has the following research challenges (research area), the first challenge is ‘heterogeneous data’, and it has the following three main types, the first one is ‘different data distributions’ and the second is ‘different outputs’ and third is ‘different feature spaces’. The second main challenge is ‘marginal distribution differences’ and the third challenge is ‘conditional distribution differences’ and the fourth challenge is ‘two-stage process’ and the fifth challenge is ‘domain adaptation process’, I has two sub-types, the first is ‘correcting both distributions’ and the second is

‘simultaneous solving of marginal and conditional’. These are the main challenges in heterogeneous transfer learning, which are shown in figure.

5 Conclusion

Heterogeneous transfer learning is about transferring knowledge from one domain to the other domain and the source and target domain have different feature spaces. In this review paper, we have reviewed several articles related to heterogeneous transfer learning and current trends in this field. Heterogeneous transfer learning is classified in two main categories, symmetric and asymmetric. In this article, we also

discussed the application of heterogeneous transfer learning and the future direction is also given in the Fig. 4 and what is the current trend in this area. Most of previous work done in this field is symmetric and asymmetric transfer learning and related to the classification like (image, text, drugs and video etc.).

Acknowledgements The authors are grateful to the School of Computer Sciences, Anhui University Hefei, China for their support and cooperation.

References

- Duan, L., Xu, D., Tsang, I.W.: Learning with augmented features for heterogeneous domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.* **36**(6), 1134–1148 (2012)
- 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*. pp. 1785–1792 (2011)
- Zhu, Y., Chen, Y., Lu, Z., Pan, S., Xue, G., Yu, Y., Yang, Q.: Heterogeneous transfer learning for image classification. In: *Proceedings of the National Conference on Artificial Intelligence*, vol. 2, pp. 1304–1309 (2011)
- Shi, X., Liu, Q., Fan, W., Yu, P.S., Zhu, R.: Transfer learning on heterogeneous feature spaces via spectral transformation. In: *2010 IEEE International Conference on Data Mining*, pp. 1049–1054 (2010)
- Qi, G.J., 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*, pp. 297–306
- Li, W., Duan, L., Xu, D., Tsang, I.W.: Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.* **36**(6), 1134–1148 (2014)
- Zhou, J.T., Tsang, I.W., Pan, S.J., Tan, M.: Heterogeneous domain adaptation for multiple classes. In: *International conference on Artificial Intelligence and Statistics*, pp. 1095–1103 (2014)
- Prettenhofer, P., Stein, B.: (2010) Cross-language text classification using structural correspondence learning. In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 1118–1127 (2010)
- Zhou, J.T., Pan, S., Tsang, I.W., Yan, Y.: Hybrid heterogeneous transfer learning through deep learning. In: *Proceedings of the National Conference on Artificial Intelligence*, vol. 3, pp. 2213–2220 (2014)
- Wei, B., Pal, C.: Cross-lingual adaptation: an experiment on sentiment classifications. In: *Proceedings of the ACL 2010 Conference Short Papers*, pp. 258–262 (2010)
- Wang, C., Mahadevan, S.: Heterogeneous domain adaptation using manifold alignment. In: *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, vol. 2, pp. 541–546 (2011)
- Harel, M., Mannor, S.: Learning from multiple outlooks. *arXiv preprint* (2010). [arXiv:1005.0027](https://arxiv.org/abs/1005.0027)
- Nam, J., Kim, S.: (2015) Heterogeneous defect prediction. In: *Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering*, pp. 508–519 (2015)
- Wei, B., Pal, C.: Heterogeneous transfer learning with RBMs. In: *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, pp. 531–536 (2011)
- Yang, Q., Chen, Y., Xue, G.R., 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*, vol. 1, pp. 1–9 (2009)
- Dai, W., Chen, Y., Xue, G.R., Yang, Q., Yu, Y.: Translated learning: transfer learning across different feature spaces. *Adv. Neural Inf. Process. Syst.* **21**, 353–360 (2008)
- Zhou, Z.-H., Dai, H.-B.: Exploiting image contents in web search. In: *IJCAI*, pp. 2922–2927 (2007)
- Wu, Lei, Hoi, Steven C.H., Jin, Rong, Zhu, Jianke, Yu, Nenghai: Distance metric learning from uncertain side information for automated photo tagging. *ACM TIST* **2**(2), 13 (2011)
- Sargano, A.B., Wang, X., Angelov, P., Habib, Z.: Human action recognition using transfer learning with deep representations. In: *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 463–469. *IEEE* (2017)
- Wu, Q., Wu, H., Zhou, X., Tan, M., Xu, Y., Yan, Y., Hao, T.: Online transfer learning with multiple homogeneous or heterogeneous sources. *IEEE Trans. Knowl. Data Eng.* **29**(7), 1494–1507 (2017)
- Yang, L., Jing, L., Yu, J., Ng, M.K.: Learning transferred weights from co-occurrence data for heterogeneous transfer learning. *IEEE Trans. Neural Netw. Learn. Syst.* **27**(11), 2187–2200 (2016)
- Li, X., Zhang, L., Du, B., Zhang, L., Shi, Q.: Iterative reweighting heterogeneous transfer learning framework for supervised remote sensing image classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **10**(5), 2022–2035 (2017)
- Tong, S., He, Q., Chen, Y., Yang, Y., Shen, B.: Heterogeneous cross-company effort estimation through transfer learning. In: *2016 23rd Asia-Pacific on Software Engineering Conference (APSEC)*, pp. 169–176. *IEEE* (2016)
- Pan, W., Liu, M., Ming, Z.: Transfer learning for heterogeneous one-class collaborative filtering. *IEEE Intell. Syst.* **31**(4), 43–49 (2016)
- Xue, S., Lu, J., Zhang, G., Xiong, L.: Heterogeneous feature space based task selection machine for unsupervised transfer learning. In: *2015 10th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*, pp. 46–51. *IEEE* (2015)
- 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*, pp. 120–128 (2006)
- Ham, J.H., Lee, D.D., Saul, L.K.: Learning high dimensional correspondences from low dimensional manifolds. In: *Proceedings of the Twentieth International Conference on Machine Learning*, pp. 1–8 (2003)
- Vapnik, V.: Principles of risk minimization for learning theory. *Adv. Neural Inf. Process. Syst.* **4**, 831–838 (1992)
- Daumé, H. III: Frustratingly easy domain adaptation. In: *Proceedings of ACL*, pp. 256–263 (2007)
- Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. *J. Am. Soc. Inf. Sci.* **41**, 391–407 (1990)
- Raina, R., Battle, A., Lee, H., Packer, B., Ng, A.Y.: Self-taught learning: transfer learning from unlabeled data. In: *Proceedings of the 24th International Conference on Machine Learning*, pp. 759–766 (2007)
- Wang, G., Hoiem, D., Forsyth, D.A.: Building text Features for object image classification. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1367–1374 (2009)
- Shawe-Taylor, J., Cristianini, N.: 2 Cambridge University Press (2004)
- Davis, J., Kulis, B., Jain, P., Sra, S., Dhillon, I.: Information theoretic metric learning. In: *Proceedings of the 24th International Conference on Machine Learning*, pp. 209–216 (2007)
- Saenko, K., Kulis, B., Fritz, M., Darrell, T.: Adapting visual category models to new domains. *Comput. Vis. ECCV* **6314**, 213–226 (2010)

36. Ando, R.K., Zhang, T.: A framework for learning predictive structures from multiple tasks and unlabeled data. *J. Mach. Learn. Res.* **6**, 1817–1853 (2005)
37. Duan, L., Xu, D., Chang, S.F.: 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*, pp. 1338–1345 (2012)
38. Li, Q., Han, Y., Dang, J.: Large-scale cross-media retrieval by heterogeneous feature augmentation. In: *2014 12th International Conference on Signal Processing (ICSP)*, pp. 977–980. IEEE (2014)
39. Liu, X., Song, L., Wu, X., et al.: Transferring deep representation for NIR-VIS heterogeneous face recognition. In: *2016 International Conference on Biometrics (ICB)*, pp. 1–8. IEEE (2016)
40. Bay, H., Tuytelaars, T., Gool, L.V.: Surf: speeded up robust features. *Comput. Vis. Image Underst.* **110**(3), 346–359 (2006)
41. Kloft, M., Brefeld, U., Sonnenburg, S., Zien, A.: Lp-norm multiple kernel learning. *J. Mach. Learn. Res.* **12**, 953–997 (2011)
42. Gao, K., Khoshgoftar, T.M., Wang, H., Seliya, N.: Choosing software metrics for defect prediction: an investigation on feature selection techniques. *J. Softw. Pract. Exp.* **41**(5), 579–606 (2011)
43. Shivaji, S., Whitehead, E.J., Akella, R., Kim, S.: Reducing features to improve code change-based bug prediction. *IEEE Trans. Softw. Eng.* **39**(4), 552–569 (2013)
44. He, P., Li, B., Ma, Y.: Towards cross-project defect prediction with imbalanced feature sets (2014). [arXiv:1411.4228](https://arxiv.org/abs/1411.4228)
45. Chen, M., Xu, Z.E., Weinberger, K.Q., Sha, F.: Marginalized denoising autoencoders for domain adaptation. *ICML*. (2012) (preprint) [arXiv:1206.4683](https://arxiv.org/abs/1206.4683)
46. Vinokourov, A., Shawe-Taylor, J., Cristianini, N.: Inferring a semantic representation of text via crosslanguage correlation analysis. *Adv. Neural Inf. Proces. Syst.* **15**, 1473–80 (2002)
47. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: *The 28th International Conference on Machine Learning*, pp. 689–696 (2011)
48. 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, pp. 97–110 (2011)