Profile of paper

Title: A Survey on Transfer Learning

Authors: Sinno Jialin Pan and Qiang Yang, Fellow, IEEE

Link: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=5288526&tag=1

Paper Structure

- Introduction
- Overview
    - A brief history of transfer learning
    - Notations and Definitions
    - A categorisation of transfer learning techniques
- Inductive Transfer Learning
    - Transferring knowledge of instances
    - Transferring knowledge of features representations
        - supervised feature construction
        - unsupervised feature construction
    - Transferring knowledge of parameters
    - Transferring relational knowledge
- Transductive Transfer Learning
    - Transferring knowledge of instances
    - Transferring knowledge of features representations
    - Unsupervised Transfer learning
    - Transfer bounds and negative transfer
- Applications of Transfer learning
- Conclusions

Note

1. A Brief History of Transfer learning

Problem Setting in general

We sometimes have a classification task in one domain of interest, but we only have sufficient training data in another domain of interest, where the latter data may be in a different feature space or follow a different data distribution. In such cases, **knowledge transfer**, if done successfully, would greatly improve the performance of learning by avoiding much expensive data-labelling efforts.

Problem in traditional machine learning algorithms

Traditional data mining and machine learning algorithms make predictions on the future data using statistical models that are trained on previously collected labelled or unlabelled training data.

[11] X. Yin, J. Han, J. Yang, and P.S. Yu, “Efficient Classification across Multiple Database Relations: A Crossmine Approach,” IEEE Trans. Knowledge and Data Eng., vol. 18, no. 6, pp. 770-783, June 2006. [12] L.I. Kuncheva and J.J. Rodrłguez, “Classifier Ensembles with a Random Linear Oracle,” IEEE Trans. Knowledge and Data Eng., vol. 19, no. 4, pp. 500-508, Apr. 2007. [13] E. Baralis, S. Chiusano, and P. Garza, “A Lazy Approach to Associative Classification,” IEEE Trans. Knowledge and Data Eng., vol. 20, no. 2, pp. 156-171, Feb. 2008.

Motivation for *transfer learning*

The study of Transfer learning is motivated by the fact that people can intelligently apply knowledge learned previously to solve new problems faster or with better solutions. The fundamental motivation for Transfer learning in the field of machine learning was discussed in a NIPS-95 workshop on “Learning to Learn”, which focused on the need for lifelong machine learning methods that retain and reuse previously learned knowledge. **p2**

```
====== NIPS-95 workshop on “Learning to Learn” ======
@book{thrun2012learning,
  title={Learning to learn},
  author={Thrun, Sebastian and Pratt, Lorien},
  year={2012},
  publisher={Springer Science \& Business Media}
}
```

Similarity and Dissimilarity: transfer learning vs multi-task learning

1. Similarity

Among these, a closely related learning technique to transfer learning is the multitask learning framework [21], which tries to learn multiple tasks simultaneously even when they are different. A typical approach for multitask learning is to uncover the common (latent) features that can benefit each individual task. **p2**

[21] R. Caruana, “Multitask Learning,” Machine Learning, vol. 28, no. 1, pp. 41-75, 1997.

2. Dissimilarity

In 2005, the Broad Agency Announcement (BAA) 05-29 of Defense Advanced Research Projects Agency (DARPA)’s Information Processing Technology Office (IPTO) gave a new mission of transfer learning: the ability of a system to recognise and apply knowledge and skills learned in previous tasks to novel tasks. In this definition, transfer learning aims to extract the knowledge from one or more source tasks and applies the knowledge to a target task. In contrast to multitask learning, rather than learning all of the source and target tasks simultaneously, transfer learning cares most about the target task. The roles of the source and target tasks are no longer symmetric in transfer learning. **p2**

2. Notations and Definitions

Domain and Task

- *a domain* consists of two components: a feature space χ and a marginal probability distribution $P(X)$, where $X = \{x_1, x_2, \ldots, x_n\} \in \chi$. In general, if two domains are different, then they may have different feature spaces or different marginal probability distributions. **p2-3**
- given a specific domain defined as before, $D = \{\chi, P(X)\}$, *a task* consists of two components as it is in the normal machine learning settings: a label space y and an objective predictive function $f(\cdot)$. Hence *a task* can be formulated as: $T = \{y, f(\cdot)\}$. In addition, we define the training data, which consists of pairs $\{x_i, y_i\}$, where $x_i \in X$ and $y_i \in Y$ **p3**

  Definition 1(Transfer Learning)

  Given a source domain D_S and learning task T_S, a target domain D_T and learning task T_T , transfer learning aims to help improve the learning of the target predictive function $f_T(\cdot)$ in D_T using the knowledge in D_S and T_S, where $D_S \neq D_T$, or $T_S \neq T_T$. **p3**

In this context, *different among domain(source and target)* means that either

1. a label space Y is different between domains
2. the conditional probability distributions between the domains are different

A categorisation of Transfer Learning Techniques

1. **What to transfer**

   which part of knowledge can be transferred across domains or tasks. Some knowledge is specific for individual domains or tasks, and some knowledge may be common between different domains such that they may help improve performance for the target domain or task. **p4**

2. **How to transfer**

   After discovering which knowledge can be transferred, learning algorithms need to be developed to transfer the knowledge, which corresponds to the “how to transfer”issue. **p4**

3. **When to transfer**

   which situations, transferring skills should be done. Likewise, we are interested in knowing in which situations, knowledge should not be transferred. **p4**

Taxonomy of Transfer Learning Situations

1. **Inductive Transfer Learning** setting

   the target task is different from the source task, no matter when the source and target domains are the same or not. So, some labelled data in the target domain are required to induce an objective predictive model for use in the target domain. Indeed, we can further categorise this domain into two different fundamental setting.

   - A lot of labelled data in the source domain are available
   - No labelled data in the source domain are available

2. **Transductive Transfer Learning** setting

   the source and target tasks are the same, while the source and target domains are different. no labelled data in the target domain are available while a lot of labelled data in the source domain are available. Indeed, we can further categorise this domain into two different fundamental setting.

   - The feature spaces between the source and target domains are different
   - The feature spaces between domains are the same

3. **Unsupervised Transfer Learning** setting

the target task is different from but related to the source task. However, the unsupervised transfer learning focus on solving unsupervised learning tasks in the target domain, such as clustering, dimensionality reduction, and density estimation

Approaches to transfer learning in the above three different settings can be summarised into four cases based on “What to transfer.” **p5**

1. **Instance transfer**: to re-weight some labelled data in the source domain for use in the target domain **p5**

   [24] B. Zadrozny, “Learning and Evaluating Classifiers under Sample Selection Bias,” Proc. 21st Int’l Conf. Machine Learning, July 2004.

   [28] W. Dai, G. Xue, Q. Yang, and Y. Yu, “Transferring Naive Bayes Classifiers for Text Classification,” Proc. 22nd Assoc. for the Advancement of Artificial Intelligence (AAAI) Conf. Artificial Intelligence, pp. 540-545, July 2007. [29] J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N.D. Lawrence, Dataset Shift in Machine Learning. MIT Press, 2009.

   [30] J. Jiang and C. Zhai, “Instance Weighting for Domain Adaptation in NLP,” Proc. 45th Ann. Meeting of the Assoc. Computational Linguistics, pp. 264-271, June 2007.

   [31] X. Liao, Y. Xue, and L. Carin, “Logistic Regression with an Auxiliary Data Source,” Proc. 21st Int’l Conf. Machine Learning, pp. 505-512, Aug. 2005.

   [32] J. Huang, A. Smola, A. Gretton, K.M. Borgwardt, and B. Scho¨lkopf, “Correcting Sample Selection Bias by Unlabeled Data,” Proc. 19th Ann. Conf. Neural Information Processing Systems, 2007.

   [33] S. Bickel, M. Bru¨ ckner, and T. Scheffer, “Discriminative Learning for Differing Training and Test Distributions,” Proc. 24th Int’l Conf. Machine Learning, pp. 81-88, 2007.

   [34] M. Sugiyama, S. Nakajima, H. Kashima, P.V. Buenau, and M. Kawanabe, “Direct Importance Estimation with Model Selection and its Application to Covariate Shift Adaptation,” Proc. 20th Ann. Conf. Neural Information Processing Systems, Dec. 2008.

   [35] W. Fan, I. Davidson, B. Zadrozny, and P.S. Yu, “An Improved Categorization of Classifier’s Sensitivity on Sample Selection Bias,” Proc. Fifth IEEE Int’l Conf. Data Mining, 2005.

2. **Feature representation transfer**: find a "god" feature representation that reduces difference between the source and the target domains and the error of classification and regression models **p5**

   [8] J. Blitzer, M. Dredze, and F. Pereira, “Biographies, Bollywood, Boom-Boxes and Blenders: Domain Adaptation for Sentiment Classification,” Proc. 45th Ann. Meeting of the Assoc. Computational Linguistics, pp. 432-439, 2007.

   [22] R. Raina, A. Battle, H. Lee, B. Packer, and A.Y. Ng, “Self-TaughtLearning: Transfer Learning from Unlabeled Data,” Proc. 24th Int’l Conf. Machine Learning, pp. 759-766, June 2007

   [36] W. Dai, G. Xue, Q. Yang, and Y. Yu, “Co-Clustering Based Classification for Out-of-Domain Documents,” Proc. 13th ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining, Aug. 2007.

   [37] R.K. Ando and T. Zhang, “A High-Performance Semi-Supervised Learning Method for Text Chunking,” Proc. 43rd Ann. Meeting on Assoc. for Computational Linguistics, pp. 1-9, 2005.

   [38] J. Blitzer, R. McDonald, and F. Pereira, “Domain Adaptation with Structural Correspondence Learning,” Proc. Conf. Empirical Methods in Natural Language, pp. 120 128, July 2006.

   [39] H. Daume´ III, “Frustratingly Easy Domain Adaptation,” Proc. 45th Ann. Meeting of the Assoc. Computational Linguistics, pp. 256-263, June 2007.

[40] A. Argyriou, T. Evgeniou, and M. Pontil, “Multi-Task Feature Learning,” Proc. 19th Ann. Conf. Neural Information Processing Systems, pp. 41-48, Dec. 2007.

[41] A. Argyriou, C.A. Micchelli, M. Pontil, and Y. Ying, “A Spectral Regularization Framework for Multi-Task Structure Learning,” Proc. 20th Ann. Conf. Neural Information Processing Systems, pp. 25- 32, 2008.

[42] S.I. Lee, V. Chatalbashev, D. Vickrey, and D. Koller, “Learning a Meta-Level Prior for Feature Relevance from Multiple Related Tasks,” Proc. 24th Int’l Conf. Machine Learning, pp. 489-496, July 2007.

[43] T. Jebara, “Multi-Task Feature and Kernel Selection for SVMs,” Proc. 21st Int’l Conf. Machine Learning, July 2004.

[44] C. Wang and S. Mahadevan, “Manifold Alignment Using Procrustes Analysis,” Proc. 25th Int’l Conf. Machine Learning ,pp. 1120-1127, July 2008.

3. **Parameter transfer**: discover shared parameters or priors between the source and the target domain models, which can benefit for transfer learning **p5**

   [45] N.D. Lawrence and J.C. Platt, “Learning to Learn with the Informative Vector Machine,” Proc. 21st Int’l Conf. Machine Learning, July 2004.

   [46] E. Bonilla, K.M. Chai, and C. Williams, “Multi-Task Gaussian Process Prediction,” Proc. 20th Ann. Conf. Neural Information Processing Systems, pp. 153-160, 2008.

   [47] A. Schwaighofer, V. Tresp, and K. Yu, “Learning Gaussian Process Kernels via Hierarchical Bayes,” Proc. 17th Ann. Conf. Neural Information Processing Systems, pp. 1209-1216, 2005.

   [48] T. Evgeniou and M. Pontil, “Regularized Multi-Task Learning,” Proc. 10th ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining, pp. 109-117, Aug. 2004.

   [49] J. Gao, W. Fan, J. Jiang, and J. Han, “Knowledge Transfer via Multiple Model Local Structure Mapping,” Proc. 14th ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining, pp. 283-291, Aug. 2008.

4. **Relational knowledge transfer**: build mapping of relational knowledge between the source and the target domains. Both domains are relational domains and i.i.d assumption is relaxed in each domain **p5**

   [50] L. Mihalkova, T. Huynh, and R.J. Mooney, “Mapping and Revising Markov Logic Networks for Transfer Learning,” Proc. 22nd Assoc. for the Advancement of Artificial Intelligence (AAAI) Conf. Artificial Intelligence, pp. 608-614, July 2007.

   [51] L. Mihalkova and R.J. Mooney, “Transfer Learning by Mapping with Minimal Target Data,” Proc. Assoc. for the Advancement of Artificial Intelligence (AAAI ’08) Workshop Transfer Learning for Complex Tasks, July 2008.

   [52] J. Davis and P. Domingos, “Deep Transfer via Second-Order Markov Logic,” Proc. Assoc. for the Advancement of Artificial Intelligence (AAAI ’08) Workshop Transfer Learning for Complex Tasks, July 2008.

We will look into all three problems settings mentioned above from now on.

3. Inductive Transfer Learning

Definition 2 (Inductive Transfer Learning).

Given a source domain D_S and a learning task T_S, a target domain D_T and a learning task T_T, inductive transfer learning aims to help improve the learning of the target predictive function $f(cdot)$ in D_T using the knowledge in D_S and T_S, where $T_S \neq T_T$. **p6**

So, put it differently, a few labelled data in the target domain are required as the training data to induce the target predictive function. **p6**

Transferring Knowledge of Instances(instance-transfer approach)

- Dai et al. [6] proposed a boosting algorithm, **TrAdaBoost**, which is an extension of the *AdaBoost* algorithm, to address the inductive transfer learning problems. **TrAdaBoost** assumes that the source and target-domain data use exactly the same set of features and labels, but the distributions of the data in the two domains are different. **p6**

  [6] W. Dai, Q. Yang, G. Xue, and Y. Yu, "Boosting for Transfer Learning," Proc. 24th Int'l Conf. Machine Learning, pp. 193-200, June 2007.

- Jiang and Zhai [30] proposed a heuristic method to remove "misleading" training examples from the source domain based on the difference between conditional probabilities $P(y_T|x_T)$and $P(y_S|x_S)$. **p6**

  [30] J. Jiang and C. Zhai, "Instance Weighting for Domain Adaptation in NLP," Proc. 45th Ann. Meeting of the Assoc. Computational Linguistics, pp. 264-271, June 2007.

- Liao et al. [31] proposed a new active learning method to select the unlabelled data in a target domain to be labelled with the help of the source domain data. **p6**

  [31] X. Liao, Y. Xue, and L. Carin, "Logistic Regression with an Auxiliary Data Source," Proc. 21st Int'l Conf. Machine Learning, pp. 505-512, Aug. 2005.

- Wu and Dietterich [53] integrated the source domain (auxiliary) data an Support Vector Machine (SVM) framework for improving the classification performance. **p6**

  [53] P. Wu and T.G. Dietterich, "Improving SVM Accuracy by Training on Auxiliary Data Sources," Proc. 21st Int'l Conf. Machine Learning, July 2004.

Transferring Knowledge of Feature Representations(Feature representation transfer)

The feature-representation-transfer approach to the inductive transfer learning problem aims at finding "good" feature representations to minimise domain divergence and classification or regression model error. Strategies to find "good" feature representations are different for different types of the source domain data. **p6**

Supervised Feature Construction

The basic idea is to learn a low-dimensional representation that is shared across related tasks. In addition, the learned new representation can reduce the classification or regression model error of each task as well. **p6**

- Argyriou et al. [40] proposed a sparse feature learning method for multitask learning. In the inductive transfer learning setting, the common features can be learned by solving an optimisation problem. **p6**

  [40] A. Argyriou, T. Evgeniou, and M. Pontil, "Multi-Task Feature Learning," Proc. 19th Ann. Conf. Neural Information Processing Systems, pp. 41-48, Dec. 2007.

- In a follow-up work, Argyriou et al. [41] proposed a spectral regularisation framework on matrices for multitask structure learning. **p6**

[41] A. Argyriou, C.A. Micchelli, M. Pontil, and Y. Ying, “A Spectral Regularisation Framework for Multi-Task Structure Learning,” Proc. 20th Ann. Conf. Neural Information Processing Systems, pp. 25- 32, 2008.

- Lee et al. [42] proposed a convex optimisation algorithm for simultaneously learning meta-priors and feature weights from an ensemble of related prediction tasks. The meta-priors can be transferred among different tasks. **p6**

  [42] S.I. Lee, V. Chatalbashev, D. Vickrey, and D. Koller, “Learning a Meta-Level Prior for Feature Relevance from Multiple Related Tasks,” Proc. 24th Int’l Conf. Machine Learning, pp. 489-496, July 2007.

- Jebara [43] proposed to select features for multitask learning with SVMs. **p6**

  [43] T. Jebara, “Multi-Task Feature and Kernel Selection for SVMs,” Proc. 21st Int’l Conf. Machine Learning, July 2004.

- Ruckert and Kramer [54] designed a kernel-based approach to inductive transfer, which aims at finding a suitable kernel for the target data. **p6**

  [54] U. Ruckert and S. Kramer, “Kernel-Based Inductive Transfer,” Proc. European Conf. Machine Learning and Knowledge Discovery in Databases (ECML/PKDD ’08), pp. 220-233, Sept. 2008.

Unsupervised Feature Construction

- In [22], Raina et al. proposed to apply sparse coding [55], which is an unsupervised feature construction method, for learning higher level features for transfer learning. **p6**

  [22] R. Raina, A. Battle, H. Lee, B. Packer, and A.Y. Ng, “Self-TaughtLearning: Transfer Learning from Unlabeled Data,” Proc. 24th Int’l Conf. Machine Learning, pp. 759-766, June 2007

  [55] H. Lee, A. Battle, R. Raina, and A.Y. Ng, “Efficient Sparse Coding Algorithms,” Proc. 19th Ann. Conf. Neural Information Processing Systems, pp. 801-808, 2007.

- Recently, manifold learning methods have been adapted for transfer learning. In [44], Wang and Mahadevan proposed a Procrustes analysis-based approach to manifold alignment without correspondences, which can be used to transfer the knowledge across domains via the aligned manifolds. **p6**

  [44] C. Wang and S. Mahadevan, “Manifold Alignment Using Procrustes Analysis,” Proc. 25th Int’l Conf. Machine Learning ,pp. 1120-1127, July 2008.

Transferring Knowledge of Parameters(Parameters transfer)

Most parameter-transfer approaches to the inductive transfer learning setting assume that individual models for related tasks should share some parameters or prior distributions of hyper-parameters. **p7**

- Lawrence and Platt [45] proposed an efficient algorithm known as MT-IVM, which is based on Gaussian Processes (GP), to handle the multitask learning case. MT-IVM tries to learn parameters of a Gaussian Process over multiple tasks by sharing the same GP prior. **p7**

  [45] N.D. Lawrence and J.C. Platt, “Learning to Learn with the Informative Vector Machine,” Proc. 21st Int’l Conf. Machine Learning, July 2004.

- Bonilla et al. [46] also investigated multitask learning in the context of GP. The authors proposed to use a free-form covariance matrix over tasks to model intertask dependencies, where a GP prior is used to induce correlations between tasks. **p7**

  [46] E. Bonilla, K.M. Chai, and C. Williams, “Multi-Task Gaussian Process Prediction,” Proc. 20th Ann. Conf. Neural Information Processing Systems, pp. 153-160, 2008.

- Schwaighofer et al. [47] proposed to use a hierarchical Bayesian framework (HB) together with GP for multitask learning. **p7**

  [47] A. Schwaighofer, V. Tresp, and K. Yu, "Learning Gaussian Process Kernels via Hierarchical Bayes," Proc. 17th Ann. Conf. Neural Information Processing Systems, pp. 1209-1216, 2005.

- Evgeniou and Pontil [48] borrowed the idea of HB to SVMs for multitask learning. And the base idea is that transferring the priors of the GP models, some researchers also proposed to transfer parameters of SVMs under a regularisation framework **p7**

  [48] T. Evgeniou and M. Pontil, "Regularized Multi-Task Learning," Proc. 10th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 109-117, Aug. 2004.

- Gao et al. [49] proposed a locally weighted ensemble learning framework to combine multiple models for transfer learning, where the weights are dynamically assigned according to a model's predictive power on each test example in the target domain. **p7**

  [49] J. Gao, W. Fan, J. Jiang, and J. Han, "Knowledge Transfer via Multiple Model Local Structure Mapping," Proc. 14th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 283-291, Aug. 2008.

Transferring Relational Knowledge(relational knowledge transfer)

Different from other three contexts, the relational-knowledge-transfer approach deals with transfer learning problems in relational domains, where the data are **non-i.i.d**. and can be represented by multiple relations, such as networked data and social network data. **p7**

- Mihalkova et al. [50] proposed an algorithm TAMAR that transfers relational knowledge with Markov Logic Networks (MLNs) across relational domains **p7**

  [50] L. Mihalkova, T. Huynh, and R.J. Mooney, "Mapping and Revising Markov Logic Networks for Transfer Learning," Proc. 22nd Assoc. for the Advancement of Artificial Intelligence (AAAI) Conf. Artificial Intelligence, pp. 608-614, July 2007.

- MLNs [56] is a powerful formalism, which combines the compact expressiveness of first-order logic with flexibility of probability, for statistical relational learning. In MLNs, entities in a relational domain are represented by predicates and their relationships are represented in first-order logic. **p7**

  [56] M. Richardson and P. Domingos, "Markov Logic Networks," Machine Learning J., vol. 62, nos. 1/2, pp. 107-136, 2006.

- In the AAAI-2008 workshop on transfer learning for complex tasks,4 Mihalkova and Mooney [51] extended TAMAR to the single-entity-centred setting of transfer learning, where only one entity in a target domain is available **p7**

  [51] L. Mihalkova and R.J. Mooney, "Transfer Learning by Mapping with Minimal Target Data," Proc. Assoc. for the Advancement of Artificial Intelligence (AAAI '08) Workshop Transfer Learning for Complex Tasks, July 2008.

- Davis and Domingos [52] proposed an approach to transferring relational knowledge based on a form of second-order Markov logic. The basic idea of the algorithm is to discover structural regularities in the source domain in the form of Markov logic formulas with predicate variables, by instantiating these formulas with predicates from the target domain. **p8**

  [52] J. Davis and P. Domingos, "Deep Transfer via Second-Order Markov Logic," Proc. Assoc. for the Advancement of Artificial Intelligence (AAAI '08) Workshop Transfer Learning for Complex Tasks, July 2008.

4. Transductive Transfer Learning

Although in the conventional *transductive transfer learning* firstly proposed by Arnold et al. [58], where they required that the source and target tasks be the same. In this paper, the author relaxed this condition to that , we only require that part of the unlabelled target data be seen at training time in order to obtain the marginal probability for the target data. **p8**

[58] A. Arnold, R. Nallapati, and W.W. Cohen, "A Comparative Study of Methods for Transductive Transfer Learning," Proc. Seventh IEEE Int'l Conf. Data Mining Workshops, pp. 77-82, 2007.

Note that we use the term transductive to emphasise the concept that in this type of transfer learning, the tasks must be the same and there must be some unlabelled data available in the target domain. **p8**

Definition 3 (Transductive Transfer Learning)

Given a source domain D_S and a corresponding learning task T_S, a target domain D_T and a corresponding learning task T_T , transductive transfer learning aims to improve the learning of the target predictive function $f(\cdot)$ in D_T using the knowledge in D_S and T_S, where $D_S \neq D_T$ and $T_S = T_T$. In addition, some unlabelled target-domain data must be available at training time. **p8**

Transferring the Knowledge of Instances(instance-transfer)

Most instance-transfer approaches to the transductive transfer learning setting are motivated by **importance sampling**. **p8**

- Zadrozny [24] proposed to estimate the terms $P(x_{z_i})$ and $P(x_{T_i})$ independently by constructing simple classification problems **p8**

  [24] B. Zadrozny, "Learning and Evaluating Classifiers under Sample Selection Bias," Proc. 21st Int'l Conf. Machine Learning, July 2004..

- Fan et al. [35] further analyzed the problems by using various classifiers to estimate the probability ratio. **p9**

  [35] W. Fan, I. Davidson, B. Zadrozny, and P.S. Yu, "An Improved Categorization of Classifier's Sensitivity on Sample Selection Bias," Proc. Fifth IEEE Int'l Conf. Data Mining, 2005.

- Huang et al. [32] proposed a kernel-mean matching (KMM) algorithm to learn $\frac{P(x_{s_i})}{P(x_{T^i})}$ directly by matching the means between the source domain data and the target domain data in a reproducing-kernel Hilbert space (RKHS). **p9**

  [32] J. Huang, A. Smola, A. Gretton, K.M. Borgwardt, and B. Scho¨lkopf, "Correcting Sample Selection Bias by Unlabeled Data," Proc. 19th Ann. Conf. Neural Information Processing Systems, 2007.

- Bickel et al. [33] combined the two steps in a unified framework by deriving a kernel-logistic regression classifier. **p9**

  [33] S. Bickel, M. Bru¨ ckner, and T. Scheffer, "Discriminative Learning for Differing Training and Test Distributions," Proc. 24th Int'l Conf. Machine Learning, pp. 81-88, 2007.

- Dai et al. [28] extended a traditional Naive Bayesian classifier for the transductive transfer learning problems. For more information on importance sampling and reweighting methods for covariate shift or sample selection bias, readers can refer to a recently published book [29] by Quionero-Candela et al. One can also consult a tutorial on Sample Selection Bias by Fan and Sugiyama in ICDM-08 **p9**

[28] W. Dai, G. Xue, Q. Yang, and Y. Yu, “Transferring Naive Bayes Classifiers for Text Classification,” Proc. 22nd Assoc. for the Advancement of Artificial Intelligence (AAAI) Conf. Artificial Intelligence, pp. 540-545, July 2007.

[29] J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N.D. Lawrence, Dataset Shift in Machine Learning. MIT Press, 2009.

Transferring Knowledge of Feature Representations(Feature representation transfer)

Most feature-representation-transfer approaches to the transductive transfer learning setting are under unsupervised learning frameworks. **p9**

- Blitzer et al. [38] proposed a structural correspondence learning (SCL) algorithm, which extends [37], to make use of the unlabeled data from the target domain to extract some relevant features that may reduce the difference between the domains. **p9**

  [36] W. Dai, G. Xue, Q. Yang, and Y. Yu, “Co-Clustering Based Classification for Out-of-Domain Documents,” Proc. 13th ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining, Aug. 2007.

  [37] R.K. Ando and T. Zhang, “A High-Performance Semi-Supervised Learning Method for Text Chunking,” Proc. 43rd Ann. Meeting on Assoc. for Computational Linguistics, pp. 1-9, 2005.

- In [63], Xing et al. proposed a novel algorithm known as bridged refinement to correct the labels predicted by a shift-unaware classifier toward a target distribution and take the mixture distribution of the training and test data as a bridge to better transfer from the training data to the test data. **p9**

  [63] D. Xing, W. Dai, G.-R. Xue, and Y. Yu, “Bridged Refinement for Transfer Learning,” Proc. 11th European Conf. Principles and Practice of Knowledge Discovery in Databases, pp. 324-335, Sept. 2007.

- In [64], Ling et al. proposed a spectral classification framework for cross-domain transfer learning problem, where the objective function is introduced to seek consistency between the in-domain supervision and the out-of-domain intrinsic structure. **p9**

  [64] X. Ling, W. Dai, G.-R. Xue, Q. Yang, and Y. Yu, “Spectral Domain Transfer Learning,” Proc. 14th ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining, pp. 488-496, Aug. 2008. **p9**

- In [65], Xue et al. proposed a cross-domain text classification algorithm that extended the traditional probabilistic latent semantic analysis (PLSA) algorithm to integrate labelled and unlabelled data from different but related domains, into a unified probabilistic model. The new model is called Topic-bridged PLSA, or TPLSA.

  [65] G.-R. Xue, W. Dai, Q. Yang, and Y. Yu, “Topic-Bridged PLSA for Cross-Domain Text Classification,” Proc. 31st Ann. Int’l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 627-634, July 2008.

- Transfer learning via dimensionality reduction was recently proposed by Pan et al. [66]. In this work, Pan et al. exploited the Maximum Mean Discrepancy Embedding (MMDE) method, originally designed for dimensionality reduction, to learn a low-dimensional space to reduce the difference of distributions between different domains for transductive transfer learning. However, MMDE may suffer from its computational burden. Thus, in [67], Pan et al. further proposed an efficient feature extraction algorithm, known as Transfer Component Analysis (TCA) to overcome the drawback of MMDE. **p10**

  [66] S.J. Pan, J.T. Kwok, and Q. Yang, “Transfer Learning via Dimensionality Reduction,” Proc. 23rd Assoc. for the Advancement of Artificial Intelligence (AAAI) Conf. Artificial Intelligence, pp. 677- 682, July 2008.

[67] S.J. Pan, I.W. Tsang, J.T. Kwok, and Q. Yang, "Domain Adaptation via Transfer Component Analysis," Proc. 21st Int'l Joint Conf. Artificial Intelligence, 2009.

5. UNSUPERVISED TRANSFER LEARNING

No labelled data are observed in the source and target domains in training.

Definition 4 (Unsupervised Transfer Learning).

Given a source domain D_S with a learning task D_S, a target domain D_T and a corresponding learning task T_T , unsupervised transfer learning aims to help improve the learning of the target predictive function $f_T(\cdot)$ ** in D_T using the knowledge in D_S and T_S, where $T_S \neq T_T$ and Y_S and Y_T are not observable.

** In unsupervised transfer learning, the predicted labels are latent variables, such as clusters or reduced dimensions. **p10**

Transferring Knowledge of Feature Representations(Feature representation transfer)

Recently, Self-taught clustering (STC) [26] and transferred discriminative analysis (TDA) [27] algorithms are proposed to transfer clustering and transfer dimensionality reduction problems, respectively. **p10**

[26] W. Dai, Q. Yang, G. Xue, and Y. Yu, "Self-Taught Clustering," Proc. 25th Int'l Conf. Machine Learning, pp. 200-207, July 2008.

[27] Z. Wang, Y. Song, and C. Zhang, "Transferred Dimensionality Reduction," Proc. European Conf. Machine Learning and Knowledge Discovery in Databases (ECML/PKDD '08), pp. 550-565, Sept. 2008.

6. TRANSFER BOUNDS AND NEGATIVE TRANSFER

- **Transfer Bounds**: the limit of the power of transfer learning
- **Negative transfer** happens when the source domain data and task contribute to the reduced performance of learning in the target domain. Despite the fact that how to avoid negative transfer is a very important issue, little research work has been published on this topic. **p10**

7. Open Problems and Conclusions

In the future, several important research issues need to be addressed.

- how to avoid negative transfer is an open problem. As mentioned in Section 6, many proposed transfer learning algorithms assume that the source and target domains are related to each other in some sense. However, if the assumption does not hold, negative transfer may happen, which may cause the learner to perform worse than no transferring at all. **p13**
- how to measure the similarity/dissimilarity between domains. **p13**
- **Heterogeneous Transfer Learning**: most existing transfer learning algorithms so far assumed that the feature spaces between the source and target domains are the same. However, in many applications, we may wish to transfer knowledge across domains or tasks that have different feature spaces, and transfer from multiple such source domains. We refer to this type of transfer learning as **heterogeneous transfer learning**.