# Heterogeneous Multi-task Semantic Feature Learning for Classification

Xin Jin[1,2], Fuzhen Zhuang[1], Sinno Jialin Pan[3], Changying Du[1], Ping Luo[1] and Qing He[1]

[1]Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, jinxin11@huawei.com, {zhuangfz, ducy, heq}@ics.ict.ac.cn, luop@ict.ac.cn
[2]Central Software Institute, Huawei Technologies Co. Ltd., Beijing 100085, China
[3]School of Computer Engineering, Nanyang Technological University, Singapore 639798, sinnopan@ntu.edu.sg

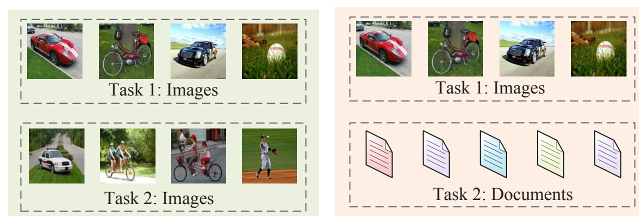
## ABSTRACT

Multi-task Learning (MTL) aims to learn multiple related tasks simultaneously instead of separately to improve generalization performance of each task. Most existing MTL methods assumed that the multiple tasks to be learned have the same feature representation. However, this assumption may not hold for many real-world applications. In this paper, we study the problem of MTL with heterogeneous features for each task. To address this problem, we first construct an integrated graph of a set of bipartite graphs to build a connection among different tasks. We then propose a multi-task nonnegative matrix factorization (MTNMF) method to learn a common semantic feature space underlying different heterogeneous feature spaces of each task. Finally, based on the common semantic features and original heterogeneous features, we model the heterogenous MTL problem as a multi-task multi-view learning (MTMVL) problem. In this way, a number of existing MTMVL methods can be applied to solve the problem effectively. Extensive experiments on three real-world problems demonstrate the effectiveness of our proposed method.

## 1. INTRODUCTION

In many real-world scenarios, one needs to collectively solve a number of related tasks, where little side information (e.g., labels) is available for each task. To solve this kind of problems, multi-task learning (MTL) has been proposed [5].

Most existing MTL methods assumed that all the tasks have the same feature representation as shown in Figure 1(a). Though this assumption holds for some applications, it may not hold for many other applications. For example, given an emergent event in social media, suppose that one task is to predict whether a social post in text is related to the event, and another task is to predict whether a social image is related to the event. On one hand, these two tasks are related as they are to make predictions on the same event. On the other hand, for each of these two tasks, label information is limited as the event is emergent. Therefore, MTL methods are desirable to solve these two tasks. However, most existing MTL methods are not applicable here because the feature representations of these two tasks are totally different (i.e., image pixels v.s. textual words) as shown in Figure 1(b). In this paper, we proposed a new method to solve the problem of MTL with heterogeneous feature spaces.




(a) *Homogeneous Features* (b) *Heterogeneous Features*

**Figure 1: Two kinds of MTL problems**

Our motivation is that though the instances of the different and related tasks are represented in different feature spaces, they may share a same semantic feature space. As in the example on social media event prediction described above, though an event-related image and an event-related post are represented differently, they should share the same semantic meanings because both of them describe the event. Once such a common semantic feature space is discovered, it can be used to share knowledge among multiple heterogeneous tasks, and thus improve their learning performance.

Specifically, in this paper, we assume multiple tasks share the same output space (i.e., the class labels of different tasks are the same or at least overlapping). Firstly, for each task, we build a bipartite graph to model the relationship between the labeled instances and the class labels. Secondly, based on the bipartite graphs of each task, we integrate them (i.e. the corresponding multiple tasks) through the layer of the class labels. After that we further build a correlation matrix between the class labels and the input features for each task based on the integrated graph. Finally, we propose a Multi-Task Nonnegative Matrix Factorization (MTNMF) method on the constructed correlation matrices as well as the original instance-feature matrices, which consists of both labeled and unlabeled data of each task, to learn a common semantic feature space underlying the multiple tasks. Once the semantic feature space is learned, together with the original heterogeneous features for each task, we can apply Multi-Task Multi-View Learning (MTMVL) methods to solve the target heterogenous MTL problem.

In summary, our proposed MTNMF method has several advantages. 1) It solves the problem of heterogeneous multi-task learning without requiring any correspondences between tasks. 2) It fully makes use of both labeled and unlabeled instances to learn the common semantic feature space for multiple tasks. 3) The learned semantic features and the original features can be concatenated to form a MTMVL problem, where a number of existing MTMVL methods can be applied.


*CIKM'15,* October 19-23, 2015, Melbourne, VIC, Australia.
Copyright 2015 ACM. ISBN 978-1-4503-3794-6/15/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/2806416.2806644.

## 2. RELATED WORK

In the past decade, MTL has attracted a lot of attention. Previous work on MTL was focused on learning multiple tasks with homogeneous features. Most methods aim to learn common features among different tasks [5, 2, 15], or common predictive structure underlying different tasks [1, 6], or common prior of model parameters among different tasks [11, 21, 22].

Recently, several heterogeneous MTL methods have been proposed. Zhang and Yeung [24] proposed the Multi-task Discriminant Analysis (MTDA) algorithm, which aims to learn transforms for instances of heterogeneous features such that the transformed instances of the same class from different tasks are closer to each other. However, MTDA is a supervised learning method, which fails to exploit unlabeled instances to learn to the transformation. He *et al.* [14] also proposed the MUSH algorithm for heterogenous MTL. However, in MUSH, some correspondence among inputs of different tasks is assumed to be given in advance.

In transfer learning, there have been some methods proposed for cross- domain/task learning with heterogenous features [25, 10, 20, 8]. However, different from transfer learning, the objective of MTL is not to transfer knowledge from a domain/task to another domain/task, but learn a prediction model for each task simultaneously by exploiting relatedness among the tasks.

Multi-Task Multi-View Learning (MTMVL) is a special setting of MTL, where each task has multiple views rather than a single view. State-of-the-art approaches to MTMVL include IteM$^2$ [13], which is a transductive learning method, regMVMT [23], CSL-MTMV [16] and MAMUDA [17], which are inductive learning methods.

## 3. PROBLEM FORMULATION

In this paper, we denote by $X_{(i,j)}$ the element in the $i$-th row and $j$-th column of a matrix $X$, $\|\cdot\|$ the Frobenius norm, and $I_l$ the $l\times l$ identity matrix. In addition, we denote by $[N:M]$ $(N<M)$ a set of integers in the range between $N$ and $M$ inclusively.

Suppose we are given $T$ related classification tasks. For each task $t\in[1:T]$, there are $n_t$ labeled and $m_t$ unlabeled instances. The dimension of an instance of the $t$-th task is $d_t$. A nonnegative matrix $X_t\in\mathbb{R}^{n_t\times d_t}\ \geq 0$ is used to denote the labeled instances of task $t$, each row of which represents an instance. Accordingly, $P_t\in\mathbb{R}^{m_t\times d_t}\ \geq 0$ is used to denote the corresponding unlabeled instances of task $t$. For simplicity, we suppose that the classification tasks are binary, and different tasks have the same set of class labels. Let $Y_t\in[-1,1]^{n_t\times 1}$ be the label vector of the labeled instances of task $t$.

## 4. MULTI-TASK SEMANTIC FEATURE LEARNING

### 4.1 Connection among Multiple Tasks

Recall that, for each task $t$, we have a set of labeled instances $X_t$ and their corresponding labels $Y_t$. Based on $X_t$, we first build a bipartite graph to capture the relationship between the labeled instances and the class labels. To be specific, we use a matrix $U_t\in\mathbb{R}^{n_t\times C}$ to represent the bipartite graph for task $t$, where $C$ is the number of classes[1], $U_{t(i,j)}=1$ if the $i$-th instance belongs to the $j$-th class, otherwise, $U_{t(i,j)}=0$. As the class labels of multiple tasks are assumed to be the same or at least overlapping, the multiple bipartite graphs can be integrated into a unified graph through the layer of the class labels. An example of the integrated graph of a pair of tasks (i.e., a pair of bipartite graphs) is shown in Figure 2.

[1]Note that in general, $C$ can be larger than 2. For simplicity, here we assume the multiple classification tasks be binary.

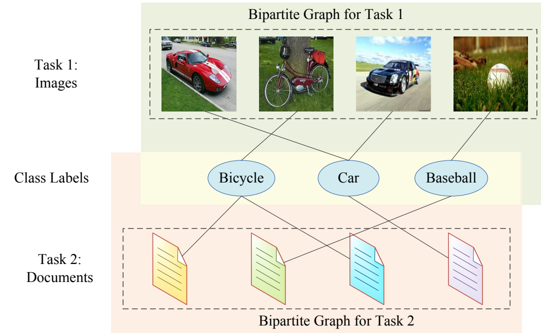



**Figure 2: Bipartite Graphs for Two Tasks**

Based on the matrix $U_t$, which is used to represent the relationship between instances and classes, we can further generate a correlation matrix $G_t$ between input features and class labels for task $t$ by setting $G_t={X_t}^\top U_t\in\mathbb{R}^{d_t\times C}$, where $G_{t(i,j)}=\sum_{k=1}^{n_t}X_{t(k,i)}\times U_{t(k,j)}$, and $X_{t(k,i)}\geq 0$ is the value of the $i$-th feature of the $k$-th labeled instance of task $t$. It can be shown that $G_{t(i,j)}$ is large if there are many instances whose the values of the $i$-th feature are large and class labels are the $j$-th class. This implies that $G_{t(i,j)}$ is large if the $i$-th feature and the $j$-th class have strong correlation.

### 4.2 Semantic Feature Learning

For each task, to extract latent semantic features, one can apply NMF to decompose $G_t$ to latent factor matrices,

$$G_t=W_tH_t,\ \text{s.t.}\ W_t\geq 0, H_t\ \geq 0,\ {W_t}^\top W_t=I. \qquad (1)$$

where $W_t\in\mathbb{R}^{d_t\times k}$, $H_t\in\mathbb{R}^{k\times C}$, and $k$ is the dimension of the latent semantic space. Each column in $W_t$ can be referred to as a base vector of the latent space, which is represented by a linear combination of the original task-specific features. The $i$-th column of $H_t$ can be referred to as the latent semantic representation for $i$-th class. The constraint ${W_t}^\top W_t=I$ is to ensure the solution to be unique and reduce redundancy [9].

For MTL, because the multiple tasks to be learned are assumed to be related, it is more desirable to learn the semantic features for each task jointly by exploiting their relatedness. In addition, as from the integrated graph, the layer of label classes is shared by all the tasks, intuitively, we can collectively learn $W_t$ for each task by enforcing the class representations $\{H_t\}$'s to be the same. Therefore, we propose to decompose $\{G_t\}$'s jointly as follows:

$$\min_{W_t\geq 0, H\geq 0, {W_t}^\top W_t=I}\sum_{t=1}^{T}\|G_t-W_tH\|^2. \qquad (2)$$

Note that the optimization problem (2) can facilitate knowledge transfer among multiple tasks through sharing $H$ in semantic feature learning for different tasks. However, (2) only makes use of labeled data because $G_t$ is constructed based on labeled instance only. However, in MTL, because labeled data for each task are scarce, the semantic features learned from (2) may not be robust and reliable. Therefore, how to exploit unlabeled data in semantic feature learning for different tasks is crucial. Intuitively, besides performing collective NMF on $\{G_t\}$'s to learn $W_t$ jointly, one can also apply NMF on the original data, which include both labeled and unlabeled instances, to learn $W_t$ for each task. To be specific, by defining $X_t'=[X_t;P_t]\in\mathbb{R}^{(n_t+m_t)\times d_t}$, we can learn $W_t$ by solving the following NMF problem:

$$X_t'=V_t{W_t}^\top,\ \text{s.t.}\ W_t\geq 0,\ V_t\geq 0,\ {W_t}^\top W_t=I, \qquad (3)$$

where $V_t\in\mathbb{R}^{(n_t+m_t)\times k}$, $W_t\in\mathbb{R}^{d_t\times k}$, and $V_t$ is the new representation of $X_t'$ under the new bases $\{W_t\}$'s. By combining (2) and (3), our proposed Multi-Task NMF method for semantic feature learning can be written as follows:

$$\min_{\substack{W_t{}^\top W_t=I\\ W_t,V_t,H\geq 0}} \sum_{t=1}^{T}\left(\|G_t-W_tH\|^2+\alpha_t\|X'_t-V_tW_t{}^\top\|^2\right), \quad (4)$$

where $\alpha_t > 0$ is a tradeoff parameter to balance the importance between the labeled and unlabeled data. For simplicity, $\alpha_t$ is set to 1 in this paper, which means that we assume the two terms in (4) be equally important. Note that the optimization problem (4) is not convex for $V_t$, $W_t$ and $H$ jointly. To solve (4), we use an alternative optimization approach to alteratively optimize one variable while fixing the other variables. The update rules of the alterative optimization approach are summarized as follows:

$$W_{t(i,j)} \leftarrow W_{t(i,j)}\sqrt{\frac{(G_tH^\top+\alpha_tX_t'^\top V_t)_{(i,j)}}{\left(W_tW_t^\top(G_tH^\top+\alpha_tX_t'^\top V_t)\right)_{(i,j)}}}, \quad (5)$$

$$H_{(i,j)} \leftarrow H_{(i,j)}\frac{\left(\sum_{t=1}^{T}(W_t{}^\top G_t)\right)_{(i,j)}}{\left(\sum_{t=1}^{T}(W_t{}^\top W_tH)\right)_{(i,j)}}, \quad (6)$$

$$V_{t(i,j)} \leftarrow V_{t(i,j)}\frac{(X'_tW_t)_{(i,j)}}{(V_tW_t{}^\top W_t)_{(i,j)}}, \quad (7)$$

where $W_{t(i,j)}$, $V_{t(i,j)}$ and $H_{(i,j)}$ denote the $(i,j)$-th element of the matrices $W_t$, $V_t$, and $H$, respectively. By applying the update rules in (5), (6) and (7), the solution converges to a local optima. Existing approaches to proving the convergence of the NMF algorithm can be adapted to prove the convergence of our proposed algorithm.[2]

## 4.3 Learning Classifiers

For each task $t$, one can use the learned matrix $W_t$ to map the original data to the common semantic space underlying all the $T$ tasks via $X_t^{(c)} = X_tW_t$. Combining with the original features, in total, there are $T+1$ views for a heterogeneous MTL problem with $T$ tasks. MTMVL techniques can be applied. In this paper, we adopt the regMVMT algorithm [23][3].

# 5. EXPERIMENTS

## 5.1 Datasets and Preprocessing

In Table 1, $N_p$ and $N_n$ are the numbers of positive and negative instances, respectively. On the second and third datasets, different tasks have some overlapping features. Besides comparing with heterogeneous MTL methods on the three datasets, we also compare MTNMF with homogeneous multi-task learning baselines on the second and third datasets.

**Table 1: Statistics of Data Sets with Heterogeneous Features**

| Problem | Task # | $N_p$ | $N_n$ | Feature # |
|---|---|---|---|---|
| 20News & Image | 6 | 986 ∼ 1,555 | 993 ∼ 1,427 | 900 ∼ 3000 |
| Email Spam | 15 | 200 | 200 | 996 ∼ 2583 |
| Sentiment | 4 | 1000 | 1000 | 1611 ∼ 2793 |

**20Newsgroups & ImageNet Classification:** The documents are from the 20 Newsgroups dataset[4], while the images are from the ImageNet dataset[5].

**Email Spam Detection** [3]: each task has a set of specific features which only include the words appear in the corresponding person's emails. When conducting comparison experiments with homogeneous MTL methods, we use another feature representation for emails that is based on a unified vocabulary for all the tasks.

**Sentiment Classification:** we use the multi-domain sentiment classification dataset [4]. The features are similarly constructed as for Email Spam dataset.

## 5.2 Experimental Setting

For each configuration, we perform 10 random trials and report the average classification accuracy.

### *5.2.1 The First Group:*

**TSVM:** Transductive SVM (TSVM) [18] is a semi-supervised learning method, we use the SVM-light[6] implementation for the TSVM classifier.

**NMF:** we first apply NMF on both the labeled and unlabeled data to learn semantic features for each task separately as shown in Eq.(3), and then train a TSVM classifier for each task with the learned semantic features separately.

**MTDA:** Multi-task Discriminant Analysis (MTDA) [24] is a multi-task learning algorithm that can deal with heterogeneous features across different tasks.

### *5.2.2 The Second Group:*

Four multi-task learning algorithms aim at problems with homogeneous features are tested, they are GMTL [19], rMTFL [12], DirtyMTL [15] and RMTL [7].

## 5.3 Experimental Results

### *5.3.1 MTL Problems with Heterogeneous Features*

Comparison results of MTNMF with the first group of baseline methods on the three datasets are shown in Tables 2, 3, 4 respectively. As shown on the tables, MTNMF performs better than NMF though both of these two methods are based on nonnegative matrix factorization. This is because the semantic features for different tasks extracted by MTNMF are not only based on the factorization on the original data matrix, but also based on the factorization on the integrated bipartite graphs which capture the correlation among different tasks. Moreover, in general, multi-task learning methods, MTNMF and MTDA, outperform the methods that learn different tasks individually, such as TSVM and NMF. This is because for each task, labeled information is too sparse to learn a precise prediction model. Last but not least, the superiority of MTNMF over MTDA suggests that the semantic features learned by MTNMF is more effective for solving multi-task learning problems with heterogeneous features.

**Table 3: Experimental Results for 20Newsgroups&Imagenet**

| Task | 1 | 2 | 3 | 4 | 5 | 6 | mean |
|---|---|---|---|---|---|---|---|
| TSVM | 70.2 | 81.9 | 76.5 | 72.6 | 78.5 | 81.3 | 76.8 |
| NMF | 52.1 | 70.1 | 63.3 | 62.7 | 64.5 | 65.5 | 63.0 |
| MTDA | 77.3 | 83.6 | 77.7 | 77.2 | 80.7 | 83.2 | 79.9 |
| MTNMF | **78.9** | **84.5** | **78.5** | **78.6** | **82.8** | **86.6** | **81.6** |

**Table 4: Experimental Results for Sentiment Problem**

| Task | 1 | 2 | 3 | 4 | mean |
|---|---|---|---|---|---|
| TSVM | 62.5 | 61.8 | 67.0 | 74.3 | 66.4 |
| NMF | 58.8 | 59.4 | 59.8 | 57.7 | 58.9 |
| MTDA | 67.4 | 68.0 | 71.8 | 74.0 | 70.3 |
| MTNMF | **68.6** | **69.2** | **73.8** | **75.6** | **71.8** |

In the second series of experiments, we compare performance between different methods under varying numbers of labeled and unlabeled training data. Note that the sizes of labeled and unlabeled data are set to be the same for this series of experiments. The results are shown in Figure 3, where both average results and standard deviation of 10 random runs are reported. As can be seen from the

[2]Due to the limited space, the detailed proof is omitted.

[3]In general, any MTMVL method can be adopted.

[4]http://people.csail.mit.edu/jrennie/20Newsgroups/

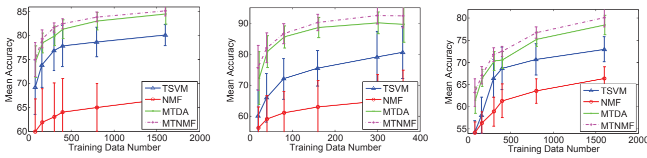
[5]http://www.image-net.org/download-features

[6]http://svmlight.joachims.org/

**Table 2: Experimental Results for Email Spam Problem (300 instances, 150 labeled and 150 unlabeled)**

| Task | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| TSVM | 84.9 | 83.1 | 78.5 | 78.4 | 86.9 | 65.1 | 76.3 | 85.7 | 86.2 | 85.4 | 73.0 | 76.8 | 87.2 | 64.0 | 73.8 | 79.0 |
| NMF | 83.8 | 67.7 | 62.3 | 65.0 | 60.7 | 55.2 | 65.0 | 73.1 | 61.6 | 80.0 | 60.0 | 53.3 | 63.3 | 64.3 | 56.4 | 64.8 |
| MTDA | 91.0 | 93.0 | 93.0 | 93.7 | 88.2 | 87.5 | 85.8 | 93.4 | 93.1 | **92.8** | 87.3 | 82.6 | 93.6 | **86.8** | 89.4 | 90.1 |
| MTNMF | **94.2** | **95.7** | **96.0** | **95.7** | **91.4** | **89.0** | **91.7** | **95.7** | **95.2** | **92.8** | **90.7** | **85.3** | **94.9** | 86.7 | **91.5** | **92.4** |

figure, MTNMF performs best under different number of training data for these 3 problems, which shows the advantage of learning a shared latent semantic space from multiple tasks.

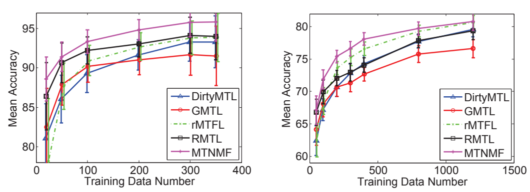



(a) *20Newsgroups & ImageNet* (b) *Email Spam* (c) *Sentiment*

**Figure 3: Experimental Results for 3 Problems (Heterogeneous Features)**

### 5.3.2 *MTL Problems with Homogeneous Features*

To conduct comparison experiments with homogeneous MTL methods, i.e., the second group of baseline methods, we use the **Email Spam Detection** and **Sentiment Classification** dataset with a unified feature representation for different tasks as described in Section 5.1. As the baseline methods are supervised learning approaches, in this series of experiments, we only use labeled training data for all the comparison methods including MTNMF. The comparison results in terms of classification accuracy are shown in Figure 4 under varying sizes of training instances. As can be seen from the figures, by extracting semantic features for each task collectively, MTNMF can also boost the performance of homogeneous MTL.




(a) *Email Spam* (b) *Sentiment*

**Figure 4: Experimental Results (Homogeneous Features)**

## 6. CONCLUSIONS

In this paper, we propose the Multi-Task Nonnegative Matrix Factorization (MTNMF) method to solve multi-task learning problems with heterogeneous feature spaces. In MTNMF, a set of integrated bipartite graphs are built based on the labeled data to model the relationship between original features and class labels among multiple tasks. A collective NMF method is then proposed to extract common semantic features from the integrated bipartite graphs as well as the unlabeled data for different tasks. Experiments on 3 real-world problems demonstrate the effectiveness of the proposed method.

## 7. ACKNOWLEDGMENTS


This work is supported by the National Natural Science Foundation of China (No. 61473273, 61473274, 61175052, 61203297), National High-tech R&D Program of China (863 Program) (No. 2014AA015105, 2013AA01A606).


## 8. REFERENCES


[1] R. K. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *JMLR*, 6:01, 2005.

[2] A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. In *NIPS*, pages 41 – 48, Vancouver, BC, Canada, 2007.

[3] S. Bickel. Ecml-pkdd discovery challenge 2006 overview. In *ECML-PKDD Discovery Challenge Workshop*, pages 1–9, 2006.

[4] J. Blitzer, M. Dredze, and F. Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *ACL*, pages 440 – 447, Prague, Czech republic, 2007.

[5] R. Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

[6] J. Chen, L. Tang, J. Liu, and J. Ye. A convex formulation for learning shared structures from multiple tasks. In *ICML*, pages 137 – 144, Montreal, QC, Canada, 2009.

[7] J. Chen, J. Zhou, and J. Ye. Integrating low-rank and group-sparse structures for robust multi-task learning. In *ACM SIGKDD*, pages 42 – 50, San Diego, CA, United states, 2011.

[8] W. Dai, Y. Chen, G.-R. Xue, Q. Yang, and Y. Yu. Translated learning: Transfer learning across different feature spaces. In *NIPS*, pages 353–360, 2008.

[9] C. Ding, T. Li, W. Peng, and H. Park. Orthogonal nonnegative matrix t-factorizations for clustering. In *ACM SIGKDD*, pages 126–135. ACM, 2006.

[10] L. Duan, D. Xu, and I. W. Tsang. Learning with augmented features for heterogeneous domain adaptation. In *ICML*, 2012.

[11] T. Evgeniou and M. Pontil. Regularized multi-task learning. In *ACM SIGKDD*, pages 109 – 117, Seattle, WA, United states, 2004.

[12] P. Gong, J. Ye, and C. Zhang. Robust multi-task feature learning. In *ACM SIGKDD*, pages 895–903. ACM, 2012.

[13] J. He and R. Lawrence. A graph-based framework for multi-task multi-view learning. In *ICML*, pages 25 – 32, Bellevue, WA, United states, 2011.

[14] J. He, Y. Liu, and Q. Yang. Linking heterogeneous input spaces with pivots for multi-task learning. In *SDM*, 2014.

[15] A. Jalali, P. Ravikumar, S. Sanghavi, and C. Ruan. A dirty model for multi-task learning. In *NIPS*, Vancouver, BC, Canada, 2010.

[16] X. Jin, F. Zhuang, S. Wang, Q. He, and Z. Shi. Shared structure learning for multiple tasks with multiple views. In *ECML PKDD*, volume 8189 LNAI, pages 353 – 368, Prague, Czech republic, 2013.

[17] X. Jin, F. Zhuang, H. Xiong, C. Du, P. Luo, and Q. He. Multi-task multi-view learning for heterogeneous tasks. In *CIKM*, pages 441–450, New York, NY, USA, 2014. ACM.

[18] T. Joachims. Transductive inference for text classification using support vector machines. In *ICML*, pages 200–209, San Francisco, CA, USA, 1999.

[19] Z. Kang, K. Grauman, and F. Sha. Learning with whom to share in multi-task feature learning. In *ICML*, pages 521–528, 2011.

[20] G.-J. Qi, C. Aggarwal, and T. Huang. Towards semantic knowledge propagation from text corpus to web images. In *WWW*, pages 297–306. ACM, 2011.

[21] G. Skolidis and G. Sanguinetti. Bayesian multitask classification with gaussian process priors. *IEEE Transactions on Neural Networks*, 22(12):2011–2021, 2011.

[22] S. Yu, V. Tresp, and K. Yu. Robust multi-task learning with t-processes. In *ICML*, volume 227, pages 1103 – 1110, Corvalis, OR, United states, 2007.

[23] J. Zhang and J. Huan. Inductive multi-task learning with multiple view data. In *ACM SIGKDD*, pages 543 – 551, Beijing, China, 2012.

[24] Y. Zhang and D.-Y. Yeung. Multi-task learning in heterogeneous feature spaces. In *AAAI*, volume 1, pages 574 – 579, San Francisco, CA, United states, 2011.

[25] Y. Zhu, Y. Chen, Z. Lu, S. J. Pan, G.-R. Xue, Y. Yu, and Q. Yang. Heterogeneous transfer learning for image classification. In *AAAI*, 2011.