Learning from Label and Feature Heterogeneity

Pei Yang  
The City College of New York  
South China University of Technology  
cs.pyang@gmail.com

Jingrui He  
Arizona State University  
jingrui.he@gmail.com

Hongxia Yang  
IBM Thomas J. Watson Research  
yangho@us.ibm.com

Haoda Fu  
Eli Lilly and Company  
fu_haoda@lilly.com

Abstract—Multiple types of heterogeneity, such as label heterogeneity and feature heterogeneity, often co-exist in many real-world data mining applications, such as news article categorization, gene functionality prediction. To effectively leverage such heterogeneity, in this paper, we propose a novel graph-based framework for Learning with both Label and Feature heterogeneities, namely $L^2F$. It models the label correlation by requiring that any two label-specific classifiers behave similarly on the same views if the associated labels are similar, and imposes the view consistency by requiring that view-based classifiers generate similar predictions on the same examples. To solve the resulting optimization problem, we propose an iterative algorithm, which is guaranteed to converge to the global optimum. Furthermore, we analyze its generalization performance based on Rademacher complexity, which sheds light on the benefits of jointly modeling the label and feature heterogeneity. Experimental results on various data sets show the effectiveness of the proposed approach.

Keywords—multi-label learning; multi-view learning; heterogeneity; Rademacher complexity.

I. INTRODUCTION

Many real-world applications exhibit both label and feature heterogeneities, such as text categorization, medical diagnosis, image/video annotation, gene functionality prediction, tag recommendation. One one hand, label heterogeneity means that each example is associated with a set of different class labels. For example, the FIFA World Cup news about the goal-line technology belong to both sports and technology categories; genes may have multiple functionalities which cause them to be associated with multiple diseases. On the other hand, feature heterogeneity means that the data is described by features from multiple views, or information sources. For example, news articles can be characterized by both the text content in the web pages, and the anchor text in the hyperlinks; proteins in given species have features that contain diverse information such as gene expression, protein-protein interactions, and sequence similarity, where some features are species-specific, and the others are cross-species.

The major challenge for addressing such problems is how to jointly model the multiple types of heterogeneity in mutually beneficial way. To address this problem, we propose a novel graph-based framework named $L^2F$ to leverage both label and feature heterogeneities. In particular, $L^2F$ accommodates multiple relationships, such as instance-instance, label-label, and view-view correlations. In this way, it is able to: (1) model the label correlation by requiring that any two label-specific classifiers behave similarly on the same views if the associated labels are similar, and (2) impose the view consistency by requiring that view-based classifiers generate similar predictions on the same examples. To solve the resulting optimization problem, we propose an iterative algorithm based on block coordinate descent. It is guaranteed to converge to the global optimum.

Furthermore, we aim to answer the fundamental question of whether the generalization performance can be improved by jointly modeling both label and feature heterogeneities. Our theoretical analysis based on Rademacher complexity shows that the error bound of the proposed framework could be improved by utilizing the label correlation and imposing the view consistency. We also empirically demonstrate the effectiveness of $L^2F$ on various data sets compared with state-of-the-art techniques. The main contributions of this paper can be summarized as follows.

- A graph-based framework named $L^2F$ for jointly modeling the label and feature heterogeneity;
- Theoretical analysis of $L^2F$ showing the benefits of simultaneously leveraging both types of heterogeneity;
- Experimental results on a variety of data sets showing the effectiveness of $L^2F$.

The rest of the paper is organized as follows. After a brief review of the related work in Section 2, we present the proposed $L^2F$ model in Section 3, and analyze its generalization performance in Section 4. Section 5 shows the experimental results on various datasets. Finally, we conclude in Section 6.

II. RELATED WORK

In this section, we survey the related work on heterogeneous learning from label or feature heterogeneity, such as multi-label learning and multi-view learning.

Multi-label learning studies the problem where each example is associated with a set of labels [17]. One key issue is to exploit correlations or dependencies among multiple labels. According to [20], existing strategies for label correlation exploitation can be grouped into three categories: first-order, second-order, and high-order approaches. First-order methods assume that labels are independent, and multi-label learning problem can be transformed into a number of independent binary classification problems, e.g., ML-kNN [21]. Second-order approaches consider the pairwise relations between labels. Then the multi-label learning problem is transformed into the label ranking problem which aims at properly ranking every relevant-irrelevant label pair for each training instance, e.g., Rank-SVM [4]. Various methods have been proposed for high-order label correlation learning. For example, LEAD [20]

This work was done when the first author conducted the postdoctoral research in The City College of New York.
employed Bayesian network to encode the conditional dependencies of the labels as well as the feature set, with the feature set as the common parent of all labels. The LS-ML algorithm was proposed for multi-label learning to extract common subspace shared among multiple labels [7]. A hypergraph spectral learning formulation was proposed for multi-label classification to exploit the correlation information among different labels using hypergraph [15]. TRAM [8] studied the problem of transductive multi-label learning by utilizing the information from both labeled and unlabeled data. LIFT [19] constructed features specific to each label by conducting clustering analysis on its positive and negative instances, and then performed training and testing by querying the clustering results. MAHR [6] aimed to discover the label relationship via a boosting approach with a hypothesis reuse mechanism. A generic empirical risk minimization (ERM) framework was proposed for large-scale multi-label learning [18].

Multi-view learning has been studied extensively in the literature. Co-training [1] is one of the earliest multi-view learning algorithm. SVM-2K [5] combined KCCA with SVM in an optimization framework. CoMR [13] was proposed for multi-view learning, which was based on a Reproducing Kernel Hilbert Space (RKHS) with a data-dependent co-regularization norm. MMH [2] was a large-margin learning framework for discovering a predictive latent subspace representation shared by multiple views. An information-theoretic framework [14] was proposed for multi-view learning, which showed that minimizing the incompatibility over unlabeled data helped reduce expected loss on the test data. The PAC generalization bound [3] was provided for co-training, which upper-bounded the error of classifiers learned from two views.

### III. THE PROPOSED $L^2F$ MODEL

In this section, we will introduce the proposed $L^2F$ model. The basic idea of $L^2F$ is to encode the label correlation and view consistency in a graph-based framework.

Let $n, m$ denote the number of examples and labels, respectively. Let $\mathcal{X}$ be an example space, and $\mathcal{L} = \{L_1,L_2,\cdots,L_m\}$ be a finite set of class labels. An example $x \in \mathcal{X}$ is described from $V$ views, i.e., $x = \{x^{(j)}|1 \leq j \leq V\}$ where $x^{(j)}$ is the instance in $j$th view, which is a feature vector. For the $j^{th} (1 \leq j \leq V)$ view, the feature dimension is denoted by $d_j$. Each example $x$ is associated with a subset of labels $L(x) \in 2^\mathcal{L}$, which is the set of relevant labels of $x$. In practice, the relevant labels $L(x)$ can be denoted by a binary label vector $Y(x) = [Y_1(x),Y_2(x),\cdots,Y_m(x)]$ where $Y_i(x) = 1$ if $x \in L_i$, $Y_i(x) = -1$ otherwise for $1 \leq i \leq m$. Let $\mathcal{Y} = \{-1,1\}^m$ be the set of all such possible labelings.

Given a data set $D = \{(x,Y(x))|x \in \mathcal{X}, Y(x) \in \mathcal{Y}\}$, consisting of $n_l$ labeled examples and $n_u$ unlabeled examples which are i.i.d drawn from some unknown distribution $P$, our goal is to build a multi-label classifier $h : \mathcal{X} \to \mathcal{Y}$ as accurately as possible. Without loss of generality, assume that the labels of the first $n_l$ examples are known. For the compactness of representation, we denote the $i^{th} (1 \leq i \leq m)$ label vector of all the examples by $y_i = [Y_1(x_1),Y_2(x_1),\cdots,Y_l(x)]^T \in \mathbb{R}^{n_x \times 1}$. Let $f_{ij} \in \mathbb{R}^{n_x \times 1}$ be the prediction vector of all the examples for the $i^{th} (1 \leq i \leq m)$ label and the $j^{th} (1 \leq j \leq V)$ view. Denote $f = [f_{11}^T,\cdots,f_{1V}^T,\cdots,f_{m1}^T,\cdots,f_{mV}^T]^T \in \mathbb{R}^{nm \times 1}$. Let $\|A\|_F$ be the Frobenius norm for the matrix $A$.

In $L^2F$, we model the multiple types of relationships including instance-instance, label-label, and view-view correlations in a graph-based framework. The goal is to maximize the smoothness consistency of the instances together with label correlation and view consistency, and simultaneously minimize the empirical loss on the training data. Thus, the objective is to minimize,

$$J(f) = J_C(f) + \alpha J_L(f) + \beta J_V(f) + \gamma J_{emp}(f)$$

(1)

where $J_C, J_L, J_V$, and $J_{emp}$ correspond to instance consistency, label correlation, view consistency, and empirical loss, respectively. The non-negative parameters $\alpha, \beta$, and $\gamma$ balance the importance of the corresponding terms. Next we will give a detailed setup of each loss function.

#### Instance Consistency on the Graph:

Let $G^{(C)}_j = \{V_j,E_j\}$ be the graph for the instances in the $j$th view, where $V_j$ is the set of instances, and $E_j$ is the set of edges. For an edge $e \in E_j$ connecting the instance pair $(x^{(j)}_i, x^{(j)}_k)$, its weight is determined by the similarity between the two instances denoted by $k(x^{(j)}_i, x^{(j)}_k)(1 \leq i, k \leq m)$, which can be estimated using the instance-feature correlation in various ways (e.g., we use Gaussian RBF function). Let $W_j \in \mathbb{R}^{n \times n}$ be the affinity matrix for the instance-instance graph $G^{(C)}_j$ whose $(i,k)$ element is $k(x^{(j)}_i, x^{(j)}_k)$. Define the Laplacian matrix $L_j = D^{-\frac{1}{2}}(D-W_j)D^{-\frac{1}{2}}$ where $D$ is a diagonal matrix with element $D_{ii} = \sum_{k=1}^n W_{ij}(i,k)$. Intuitively, similar instances should have similar predictions. Following the random walk model [23], we model the instance consistency as follows,

$$J_C(f) = \sum_{i=1}^m \sum_{j=1}^V f_{ij}^T L_j f_{ij} = f^T Q C f$$

(2)

where $Q_C$ is a block diagonal matrix with its entry $[Q_C]_{ij,ij}$ is $L_j$ for $1 \leq i \leq m, 1 \leq j \leq V$. Since the Laplacian matrix $L_j$ is positive semi-definite, $Q_C$ is also positive semi-definite.

#### Label Correlation:

Let $G^{(L)} = \{V,E\}$ be the graph for the labels, where $V = \mathcal{L}$ is the set of labels, and $E$ is the set of edges. For an edge $e \in E$ connecting the label pair $(L_i,L_k)$, its weight is determined by the similarity between the two labels denoted by $k(L_i,L_k)(1 \leq i, k \leq m)$, which can be estimated using the example-label correlation in various ways. Let $S \in \mathbb{R}^{m \times m}$ be the affinity matrix for the label-label graph $G^{(L)}$ whose $(i,k)$ element is $k(L_i,L_k)$. The degree of a label $L_i (1 \leq i \leq m)$ is defined as $d_i = \sum_{j=1}^m S_{ij}$.

Based on the graph $G^{(L)}$, we model the label correlations by requiring that any two label-specific classifiers behave similarly on the same views if the associated labels are similar. In specific, if two labels $L_i$ and $L_k$ are similar, the label-specific classifiers $f_{ij}$ and $f_{kj}$ should keep close to each other on the same $j$th view. Therefore, we model the correlation among multiple labels as follows,

$$J_L(f) = \sum_{j=1}^V \sum_{i,k=1}^m S_{ik} \left\| \frac{f_{ij}}{\sqrt{d_i}} - \frac{f_{kj}}{\sqrt{d_k}} \right\|_F^2 = f^T Q L f$$

(3)

where $Q_L$ is a block matrix with its entry,

$$[Q_L]_{ij,kj} = \begin{cases} 2(1-S_{ik}/d_i) I_{i=x} & i = k \\ -2S_{ik} I_{i=x}/\sqrt{d_i} & i \neq k \end{cases}$$

for $1 \leq i, k \leq m, 1 \leq j \leq V$. Since $f^T Q_L f \geq 0$, $Q_L$ is positive semi-definite.
View Consistency: In order to maximize the view consistency, we require that for any view pairs, the difference of predictions resulting from their view-based classifiers should keep small as much as possible. Hence, we model the consistency among multiple views as follows,

\[ J_V (f) = \sum_{m=1}^{m} \sum_{j,k=1}^{V} \| f_{ij} - f_{ik} \|^2_F = f^T Q f \]  

(4)

where \( Q \) is a block matrix with its entry,

\[ [Q]_{ij,ik} = \begin{cases} 2(1-V)^{-1} I_{n \times n}, & j = k \\ -2I_{n \times n}, & j \neq k \end{cases} \]

for \( 1 \leq i \leq m, 1 \leq j, k \leq V \). Since \( f^T Q f \geq 0 \), \( Q \) is positive semi-definite.

Empirical Loss: Various empirical loss functions, such as hinge loss, least square loss, logistic loss, etc., can be used to measure the consistency with known label information.

Overall Objective: In summary, the overall goal is to minimize the following objective function:

\[ J (f) = J_C (f) + \alpha J_L (f) + \beta J_V (f) + \gamma J_{emp} (f) \]

\[ = f^T Q f + \gamma J_{emp} (f) \]  

(5)

where \( Q = \alpha Q_C + \beta Q_V \).

A nice property of the proposed method is that its objective function is joint convex as shown in the following theorem. Due to space limitation, we omit the proofs for all the theorems, which will be given in a long version of this paper.

Theorem 3.1 (Optimality): When using convex empirical loss, the objective in Eq. 5 is convex with respect to \( f \).

When using least square loss as empirical loss function, the objective function in Eq. 5 can be solved analytically. For the least square loss, we have

\[ J_{emp} (f) = \sum_{i=1}^{m} \sum_{j=1}^{V} \| f_{ij} - y_i \|^2_F = f^T Q f + 2 f^T p + q \]

where \( Q_{emp} \) is block diagonal matrix with its entry \([Q_{emp}]_{ij,ij} = I_{n \times n}, p \) is a block vector with its entry \( p_{ij} = y_i \), and \( q \) is a constant block vector with its entry \([q_{ij}] = y_i y_i^T 1_{n \times 1} \) for \( 1 \leq i \leq m, 1 \leq j \leq V \). Obviously, \( Q_{emp} \) is positive semi-definite. Then, the objective function in Eq. 5 can be rewritten into

\[ J (f) = J_C (f) + \alpha J_L (f) + \beta J_V (f) + \gamma J_{emp} (f) \]

\[ = f^T Q f + \gamma J_{emp} (f) \]  

(6)

where \( Q_A = \alpha Q_C + \beta Q_V + \gamma Q_{emp} \). Obviously, \( Q_A \) is positive semi-definite. By taking derivative of Eq. 6 with respect to \( f \), we have \( \nabla f J (f) = 0 \Rightarrow f^* = Q_A^{-1} p \).

Optimization using block coordinate descent: Since \( Q_A \in \mathbb{R}^{nm \times nm} \), the space complexity of the above method is \( O(n^2 m^2 V^2) \). To reduce the space complexity, we resort to block coordinate descent (BCD) method [10], [16] to iteratively solve the optimization problem. We first rewrite the objective in Eq. 6 as follows

\[ J (f) = J_C (f) + \alpha J_L (f) + \beta J_V (f) + \gamma J_{emp} (f) \]

\[ = \sum_{m=1}^{m} \sum_{j=1}^{V} \sum_{i=1}^{m} \sum_{k=1}^{V} S_{ik} \left( \frac{f_{ij}^T f_{ik}}{\| f_{ij} \|^2} - \frac{1}{\sqrt{d_i d_k}} \right) + \beta \sum_{m=1}^{m} \sum_{j,k=1}^{V} (f_{ij}^T f_{ik} + f_{ik}^T f_{ij}) \]

\[ + \gamma \sum_{m=1}^{m} \sum_{j,k=1}^{V} (f_{ij}^T f_{ik} - 2 f_{ij}^T y_i + y_i^T y_i) \]

By setting the first-order derivative of the above equation with respect to \( f_{ij} (1 \leq i \leq m, 1 \leq j \leq V) \) to zero, we obtain the analytical solution as follows

\[ f_{ij}^* = H^{-1} p_{ij} \]  

(7)

where \( H_{ij} = 2L_j + \left[ 4\alpha + 4\beta (V-1) + 2\gamma \right] I_{n \times n} \) and \( p_{ij} = \sum_{k=1,k \neq j}^{V} \frac{S_{ik}}{\sqrt{d_i d_k}} f_{ik} + 4\beta \sum_{k=1,k \neq j}^{V} f_{ik} + 2\gamma y_i \).

Prediction: For the test example, the final prediction is the expectation of predictions resulting from view-based classifiers. For the example \( x \), the prediction for its \( h \)th label is as follows

\[ h_i(x) = \operatorname{sgn} \left( \frac{1}{V} \sum_{j=1}^{V} f_{ij}^* (x) \right) \]  

(8)

The following theorems show the convergence property and algorithm complexity of the proposed method.

Theorem 3.2 (Convergence): The \( L^2 \)F method converges to the global optimum.

Theorem 3.3 (Time Complexity): The time complexity of the \( L^2 \)F method is \( O(n + m V n m^2) \) where \( n \) is the iteration of the algorithm.

Theorem 3.4 (Space Complexity): The space complexity of the \( L^2 \)F method is \( O(V n^2 + m^2) \).

IV. THEORETIC ANALYSIS

In this section, we analyze the generalization performance of the proposed framework, which shows the benefits of simultaneously modeling label and feature heterogeneity. To be specific, we will demonstrate that the upper bound of empirical Rademacher complexity together with the error bound of the proposed \( L^2 \)F model can be reduced by incorporating the label correlation and enhancing the view consistency.

Let \( \mathcal{H} \) be the space of functions with the norm defined as \( \| f \|_{\mathcal{H}}^2 = f^T Q f \). Based on \( \mathcal{H} \), we define \( \tilde{\mathcal{H}} \) to be the space of functions with the norm \( \| f \|_{\tilde{\mathcal{H}}}^2 = \| f \|_{\mathcal{H}}^2 = f^T Q f + \alpha J_C + \beta J_V f = f^T Q f \). Suppose that \( Q_C, Q_L, Q_V \) are invertible 1. The following theorem will show that both \( \mathcal{H} \) and \( \tilde{\mathcal{H}} \) are RKHS.

Theorem 4.1 (RKHS): Both \( \mathcal{H} \) and \( \tilde{\mathcal{H}} \) are RKHS with kernel matrix \( K = Q_C^{-1} \), and \( \tilde{K} = [Q_C + \alpha Q_L + \beta Q_V]^{-1} \), respectively.

Hence, based on Theorem 4.1, the overall objective in Eq. 5 can be reduced to standard supervised learning problem:

\[ f^* = \arg \min_{f \in \tilde{\mathcal{H}}} \| f \|_{\tilde{\mathcal{H}}}^2 + \gamma J_{emp} (f) \]  

(9)

Let \( \mathcal{F} := \{ f \in \tilde{\mathcal{H}} : \| f \| \leq r \} \) denote the ball of radius \( r \) in \( \mathcal{H} \). According to Theorem 4.12 in [12], we can easily obtain the following theorem regarding the Rademacher complexity of the proposed method.

Theorem 4.2 (Rademacher complexity): The empirical Rademacher complexity of the proposed \( L^2 \)F method is upper bounded by:

\[ \hat{R} (F) \leq \frac{2r}{nm V} \sqrt{\text{tr} \left( [Q_C + \alpha Q_L + \beta Q_V]^{-1} \right)} \]  

1 Otherwise, a practical approach is to add a small regularization term \( \lambda I (\lambda \geq 0) \) to it.
Theorem 4.9 in [12] together with Theorem 4.2 show that: 

\[
\text{the overall objective as defined in Eq. 5. An application of}
\]

incorporating the label correlation and view consistency into 

\[
\text{er complexity of the proposed}
\]

\[
L
\]

views. From Theorem 4.3, we can see that the Rademach-

\[
\text{labels, while}
\]

\[
Q_L
\]

counts, where the number of latent topics is set to 100. 

are described from two views: one corresponds to the TF-IDF 

\[
\text{the multi-view data as follows. For each dataset, the instances}
\]

\[
\text{labels. The above two datasets are available online}
\]

\[
3
\]

\[
\text{data instances on average and with a total number of 101 class}
\]

\[
\text{label cardinality by the the number of labels. Label diversity}
\]

\[
\text{different datasets. Label cardinality is the average number}
\]

\[
2
\]

\[
\text{optimization problem of estimating label concept composition-}
\]

They formulate the transductive multilabel classification as an 

\[
\text{special features instead of the original ones. LS-ML [7] learns}
\]

\[
\text{ML-kNN by building the classifier on each label with label-}
\]

\[
\text{kNN is a first-order approach which ignores the correlation}
\]

\[
\text{among multiple labels. In contrast, all the other algorithms}
\]

\[
\text{performs better than ML-kNN by leveraging the label}
\]

\[
\text{of all, a common trend observed from these figures is that}
\]

\[
\text{show the results on different subset of Reuters Corpus V olume}
\]

\[
\text{RCV) datasets, respectively. The results on Diabetes dataset}
\]

\[
\text{figure 7. In each figure, x-axis represents the}
\]

\[
\text{ratio which is used to randomly sample a subset of instances}
\]

\[
\text{from the training data, and y-axis denotes the}
\]

\[
\text{indicating better performance.}
\]

\[
\text{F_1-score (harmonic mean of precision and}
\]

\[
\text{recall) [22] on the test data as the evaluation metric, which}
\]

\[
\text{as follows:}
\]

\[
F_1 = \frac{1}{n_u} \sum_{k=1}^{n_u} \frac{2|L(x_k) \cap Z(x_k)|}{|L(x_k)| + |Z(x_k)|}
\]

\[
\text{where}
\]

\[
Z(x) = \{L_i | b_i(x) = 1, 1 \leq i \leq m\}
\]

\[
\text{is the predicted label set for example x. Note that larger value of F}_1\text{-score are}
\]

\[
\text{indicating better performance.}
\]

\[
\text{Figure 1 shows the results on Medical dataset. Figures 2-6}
\]

\[
\text{show the results on different subset of Reuters Corpus Volume}
\]

\[
\text{RCV datasets, respectively. The results on Diabetes dataset}
\]

\[
\text{are shown in Figure 7. In each figure, x-axis represents the}
\]

\[
\text{ratio which is used to randomly sample a subset of instances}
\]

\[
\text{from the training data, and y-axis denotes the}
\]

\[
\text{indicating better performance.}
\]

\[
\text{The results show that}
\]

\[
L^2F
\]

\[
\text{performs better than the other}
\]

\[
\text{algorithms in most cases. ML-kNN [21] performs worst among}
\]

\[
\text{all the algorithms. This is likely due to the fact that ML-}
\]

\[
kNN is a first-order approach which ignores the correlation
\]

\[
\text{among multiple labels. In constrast, all the other algorithms}
\]

\[
\text{usually perform better than ML-kNN by leveraging the label}
\]

\[
\text{correlations. In principle, the classifiers induction process of}
\]

\[
\text{LIFT [19] is similar to ML-kNN. But LIFT improves upon}
\]

\[
\text{ML-kNN by building the classifier on each label with label-}
\]

\[
\text{specific features instead of the original ones. LS-ML [7] learns}
\]

\[
\text{a common subspace shared among multiple labels, which}
\]

\[
\text{helps improve the learning performance for the multi-label}
\]

\[
\text{data. However, since its objective function is non-convex,}
\]

\[
\text{the performance of LS-ML would be limited by the local}
\]

\[
\text{optimum problem. Different from other approaches, TRAM [8]}
\]

\[
\text{is a transductive multi-label learning method which tries to}
\]

\[
\text{exploit the information from both labeled and unlabeled data.}
\]

\[
\text{They formulate the transductive multilabel classification as an}
\]

\[
\text{optimization problem of estimating label concept composition-}
\]

\[
\text{The results show that unlabeled data can provide helpful}
\]

\[
\text{information to build the multi-label classifier.}
\]

\[
\text{In comparison with the other methods, the key advantage of}
\]

\[
L^2F
\]

\[
\text{is that it models both label and feature heterogeneities
in a principled framework. First, by leveraging the consistency among multiple views, the view-based classifiers can mutually improve each other. On the contrary, all the other comparison methods do not consider the view consistency, simply concatenating features from different views cannot gain much additional improvement. Second, by considering the correlation among multiple labels, the performance of label-specific classifiers in $L^2F$ can benefit from each other. In the next subsection, we will also show how the performances of $L^2F$ vary with the trade-off parameters, $\alpha$ and $\beta$, which control the weight of label correlation and view consistency, respectively. Another competency of $L^2F$ is that it is capable of finding the global optimum due to the joint convexity of the objective function.

In addition, we have a few different observations on the Diabetes dataset. The performance of three algorithms, i.e., ML-kNN, LS-ML, and LIFT, are poor, indicating that this is a more challenging task. In contrast, both TRAM and $L^2F$ perform better, and their results are comparable on this dataset. This might due to the fact both TRAM and $L^2F$ take advantage of the unlabeled data information. TRAM utilizes the unlabeled data in a transductive way, while $L^2F$ leverages the smoothness consistency among the nearest instances.
F1-score

0.45

0.55

0.4

0.2

0.5

0.3

0.3

0.4

0.4

0.4

0.6

0.5

0.6

0.6

0.7

0.7

0

0

0.1

0.2

0.3

0.4

0.5

0.6

0.7

α

β

γ

Fig. 8: $F_1$-score varies with $\alpha$ and $\beta$ (log2 scale).

Fig. 9: $F_1$-score varies with $\gamma$ (log2 scale) and iteration.

C. Parameter Sensitivity and Convergence

We study the parameter sensitivity on the Medical dataset. $\alpha$ and $\beta$ are used to balance the importance of label correlation and view consistency, respectively. We tune $\alpha$ and $\beta$ on the grid $2^{-4:1:4}$. Figure 8 shows the results. Comparing with $\alpha = 0$, the algorithm performs better when $\alpha$ increases, and the best case occurs when $\alpha = 1$, which indicates that modeling label correlation could significantly improve the multi-label learning performance. However, if $\alpha$ is very large such as $\alpha = 16$, the label correlation part will dominate the entire objective function, making the model hard to keep certain level of accuracy. Nevertheless, the performance is robust over a wide range of values for $\alpha$. In Figure 8, we observe a similar trend for $\beta$, which suggests that the learner could benefit from enhancing the consistency among multiple views. $\gamma$ is used to control the weight of empirical loss. We tune $\gamma$ on the grid $2^{-4:1:4}$. The result is shown in the left panel of Figure 9. As expected, the performance is poor when $\gamma = 0$, and the $F_1$-score first increases and then decreases when $\gamma$ is increased. As a result, we tune the parameters on each dataset using standard cross-validation.

We empirically study the convergence property of $L^2F$ algorithm on the Medical dataset. The result is shown in the right panel of Figure 9. From this figure, we can see that $L^2F$ converges fast and its performance becomes stable after 10 iterations.

VI. CONCLUSION AND FUTURE WORK

In this paper, we propose a graph-based approach $L^2F$ for learning from both label and feature heterogeneities. An iterative algorithm is presented to solve the convex problem, which is guaranteed to converge to the global optimum. We analyze its performance in terms of its generalization error rate. Both the theoretic analysis and the comparison experiments with state-of-the-art methods demonstrate the effectiveness of the proposed method. One of our on-going work is to extend the proposed framework to the transductive setting. Due to its label and feature heterogeneity, transductive learning is particularly challenging in this situation.

ACKNOWLEDGMENT

This material is partially supported by the National Science Foundation under Grant No. IIS1017415, by the Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053, by DARPA under Contract No. W911NF-11-C-0200 and W911NF-12-C-0028, by Region II University Transportation Center under the project No. 49997-33 25, and by NSFC under Grant No. 61473123.

REFERENCES