Multi-Label Classification by Mining Label and Instance Correlations from Heterogeneous Information Networks

Xiangnan Kong
University of Illinois at Chicago
Chicago, IL, USA
xkong4@uic.edu

Bokai Cao
Renmin University of China
Beijing, China
caobokai@ruc.edu.cn

Philip S. Yu
University of Illinois at Chicago
Chicago, IL, USA
King Abdulaziz University
Jeddah, Saudi Arabia
psyu@cs.uic.edu

ABSTRACT

Multi-label classification is prevalent in many real-world applications, where each example can be associated with a set of multiple labels simultaneously. The key challenge of multi-label classification comes from the large size of all possible label sets, which is exponential to the number of candidate labels. Most previous work focuses on exploiting correlations among different labels to facilitate the learning process. It is usually assumed that the label correlations are given beforehand or can be derived directly from data samples by counting their label co-occurrences. However, in many real-world multi-label classification tasks, the label correlations are not given and can be hard to learn directly from data samples within a moderate-sized training set. Heterogeneous information networks can provide abundant knowledge about relationships among different types of entities including data samples and class labels. In this paper, we propose to use heterogeneous information networks to facilitate the multi-label classification process. By mining the linkage structure of heterogeneous information networks, multiple types of relationships among different class labels and data samples can be extracted. Then we can use these relationships to effectively infer the correlations among different class labels in general, as well as the dependencies among the label sets of data examples inter-connected in the network. Empirical studies on real-world tasks demonstrate that the performance of multi-label classification can be effectively boosted using heterogeneous information networks.

Categories and Subject Descriptors
H.2.8 [Database Management]: Database Applications-Data Mining

General Terms
Algorithm, Experimentation, Performance

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

KDD'13, August 11–14, 2013, Chicago, Illinois, USA.
Copyright 2013 ACM 978-1-4503-2174-7/13...$15.00.

Keywords
Multi-label Classification, Heterogeneous Information Network, Label Correlation, Data Mining

1. INTRODUCTION

Multi-label classification has become increasingly important in recent years, where each example can be associated with multiple labels simultaneously. It has a wide range of real-world applications. For example, in drug discovery, one molecular drug can bind with multiple protein targets, and researchers would like to predict which protein targets that one chemical compound can bind with in order to discover new drugs for a certain disease; in gene-disease association prediction, one gene sequence can involve in multiple diseases, and researchers are interested in predicting which diseases that each gene is related to.

The key challenge of multi-label classification comes from the large space of all possible label sets, which is exponential to the number of candidate labels. To tackle this problem, conventional multi-label classification approaches [33, 11, 15] focus on exploiting the correlations among different class labels to facilitate the learning process. It is usually assumed that the label correlations are given beforehand or can be derived directly from data samples by counting their label co-occurrences.

However, in many real-world multi-label classification tasks, the label correlations are not given and can be hard to learn directly from data samples within a moderate-sized training set. Heterogeneous information networks [31] can usually provide abundant knowledge about the relationships among different types of entities including data samples and class labels. In this paper, we propose to use heterogeneous information networks to facilitate the multi-label classification process. We first introduce an example of real-world multi-label classification task and show how a heterogeneous information network can facilitate the learning process of multi-label classification.

Example (Drug-Target Binding Prediction): In drug discovery, one molecular drug can bind with a set of gene targets. The multi-label classification task here is to predict which set of gene targets that one chemical compound can bind with in order to discover new drugs for a certain disease. The number of candidate labels are usually very large, involving hundreds or thousands of gene targets. The correlations among these gene targets are hard to learn directly from data samples within a moderate-sized training set. However, if we already have a heterogeneous informa-
First, the heterogeneous information network can provide abundant knowledge about the relationships among different gene targets. In the network, gene targets are interconnected with many other types of nodes, such as diseases and pathways. The gene targets, that are linked with similar diseases or pathways, are more likely to appear together in the same label set than those without such connections. Second, the heterogeneous information network can also provide abundant knowledge about the relationships among different chemical compounds. In the network, chemical compounds are also connected with other types of objects, such as side effects and chemical ontologies. The chemical compounds, that are linked with similar side effects or chemical ontologies, are more likely to have similar label sets than the chemicals without such connections.

By mining the linkage structure of heterogeneous information networks, multiple types of relationships among different class labels and data samples can be extracted. Such relationships can then be used to infer the correlations among different class labels in general, and the dependencies among the label sets of different data samples. In this paper, we focus on studying the problem of multi-label classification by mining label and instance correlations from heterogeneous information networks. The major research challenges are as follows:

Mining heterogeneous label correlations: In multi-label classification, the multiple label concepts can be correlated with each other through multiple types of relationships. For example, different gene targets can be correlated for various reasons: 1) they belong to the same gene family; 2) they share similar pathways; 3) they are inter-connected through PPI links, etc. Heterogeneous information networks can provide complex relationships among the label concepts, involving multiple types of label correlations [28]. How to exploit the linkage semantics is a very challenging problem [28], which has not yet been explored in this context.

Mining heterogeneous instance correlations: In multi-label classification, the label sets of different instances can also be correlated with each other through multiple types of relationships. For example, different chemical compounds can be correlated for various reasons: 1) they have similar side effects; 2) they have similar chemical ontologies; 3) they have similar substructures, etc. Heterogeneous information networks can provide complex relationships among different instances, involving multiple types of correlations.

In this paper, we study how we can facilitate the multi-label classification process by mining the correlations among labels and instances from heterogeneous information networks. We propose a novel solution, called PIPL, to assign a set of candidate labels to a group of related instances in heterogeneous information networks. Different from previous work, the proposed PIPL, as shown in Figure 2, can exploit various types of dependencies among both of instances and labels based upon different meta-paths in heterogeneous information networks. By explicitly exploiting these meta-path based dependencies, our PIPL method can effectively capture the diverse and complex relationships among instances and labels. Empirical studies on real-world tasks show that the proposed method can significantly boost the performance of multi-label classification by incorporating information from heterogeneous information networks.

2. PROBLEM DEFINITION

In this section, we first introduce some related concepts and notations, then define the problem.
2.1 Multi-label Classification

Suppose we have a multi-label classification task with \( n \) instances and \( q \) labels. The set of all instances is denoted as \( V_I = \{ i_1, \ldots, i_n \} \), where each instance \( i_I \) has a feature vector \( x_i \) in the input space. The set of all candidate labels is denoted as \( V_Y = \{ l_1, \ldots, l_q \} \). Without loss of generality, we assume the first \( n_I \) instances in \( V_I \), denoted as \( V_I' = \{ i_1, \ldots, i_{n_I} \} \subseteq V_I \), are labeled, where each instance \( i_I \) is associated with a label set, denoted as \( Y_i = \{ l_1^i, \ldots, l_q^i \} \in \{ 0, 1 \}^q \). If instance \( i_I \) has the \( j \)-th label \( l_j \) in its label set, \( Y_{i,j} = 1 \), otherwise \( Y_{i,j} = 0 \). The remaining instances in \( V_I \) are unlabeled, i.e., \( V_{I,U} = V_I - V_I' \). The task of multi-label classification is to predict the label sets for all the unlabeled instances in \( V_{I,U} \). In other words, for any possible label set \( Y_i \in \{ 0, 1 \}^q \) and any instance \( i_I \), we need to build a model to estimate the probability \( P(Y_i|X_i) \) based upon the training dataset.

The key challenge of multi-label classification comes from the fact that the output space of all possible label sets \( \{ 0, 1 \}^q \) is very large, which is exponential to the number of candidate labels \( q \). In order to tackle this problem, one simplest solution is to decompose the multi-label classification problem into a set of binary classification problems (one for each label), e.g., Bsvm [2]. This method can estimate \( P(Y_i|X_i) \) for instance \( i_I \) approximately by assuming that all labels are independent:

\[
\forall i_I, \quad P(Y_i|X_i) \approx \prod_{k=1}^q P(Y_{i,k}|X_i)
\]

However, the correlations among different labels are totally ignored in these methods.

Most previous multi-label classification approaches focus on exploiting the label correlations to facilitate the learning process. For example, the ECC method [27] captures label correlations directly from data samples by approximate the probability \( P(Y_i|X_i) \) as follows:

\[
P(Y_i|X_i) \approx P(Y_i^1|X_i)P(Y_i^2|Y_i^1, X_i)P(Y_i^3|Y_i^1, Y_i^2, X_i) \ldots
\]

It is usually assumed implicitly that the label correlations can be derived directly from the observed data samples by counting their label co-occurrences.

However, in many real-world multi-label classification tasks, the label correlations are hard to learn directly from data samples with a moderate-sized training set. Heterogeneous information networks [31] can usually provide abundant knowledge about the relationships among different types of entities including data samples and class labels. In this paper, we propose to use heterogeneous information networks to facilitate the multi-label classification process.

2.2 Heterogeneous Information Network

A heterogeneous information network is a special kind of information network with multiple types of nodes and multiple types of links [31, 28]. It can be represented as a directed graph \( G = (V, E) \), where \( V \) denotes the set of nodes, which involves \( t \) types of objects: \( V_I = \{ v_{i_1}, \ldots, v_{i_n} \} \), \( V_Y = \{ v_{l_1}, \ldots, v_{l_q} \} \), where \( v_{i_j} \) represents the \( j \)-th object of type \( t \). \( E \subseteq V \times V \) denotes the links between the objects in \( V \), which includes multiple types of links.

First we define the type of nodes in the heterogeneous information network \( G \) corresponds to the instances of the multi-label classification task as the instance type. While we define type of nodes corresponds to the label concepts for the multi-label classification task as the label type. For example, in the task of drug-target prediction, the chemical compound nodes are the instances and the gene nodes are the labels; in the task of gene-disease association prediction, the gene nodes are the instances and the disease nodes are the labels.

3. PROPOSED METHOD

Heterogeneous information networks can provide abundant knowledge about the relationships among different types of entities including data samples and class labels. By mining the linkage structure of heterogeneous information networks, multiple types of relationships among different class labels and the data samples can be extracted. Such relationships can be used to infer the correlations among different class labels in general, and the dependencies among the label sets of different data samples.

In the following subsections, we first discuss how to extract heterogeneous relationships among different labels from heterogeneous information networks in section 3.2. We then discuss how to extract heterogeneous relationships among the instance from heterogeneous information networks in section 3.3. We then integrate these two parts into a unified model in section 3.4.

3.1 Meta-path

Before discussing in detail about the proposed method, we first introduce the concept of meta-path in heterogeneous information networks. Following the works [28, 20, 18, 30], we first briefly review the concept of meta-path. Then we use meta-paths to describe the possible relationships among instances and labels derived from a heterogeneous information network.

Generally, meta-path is defined as a sequence of relations in the network schema. The objects in heterogeneous information networks (including the instances and labels) are inter-connected through multiple types of links. Each type of links from node type \( i \) to node type \( j \) corresponds to a binary relation \( R \), where \( R(v_{i,p}, v_{j,q}) \) holds if object \( v_{i,p} \) and \( v_{j,q} \) are linked by a link of type \( R \). For example, in Figure 1, the link type “hasPathway” is a relation between gene nodes and pathways nodes, where \( R(v_{i,p}, v_{j,q}) \) holds if the gene node \( v_{i,p} \) has a link of type “hasPathway” to the pathway node \( v_{j,q} \) in the network. We can write the link type as “gene \( \xrightarrow{\text{hasPathway}} \) pathway” or “\( \forall i_p \xrightarrow{R} v_{j,q} \)”.

A meta-path \( P \) corresponds a type of path within the network graph, containing a certain sequence of link types. For example, in Figure 1, a meta-path “gene \( \xrightarrow{\text{hasPathway}} \) pathway \( \xrightarrow{\text{hasGO}} \) gene” denotes the composite relationship between gene nodes where the semantic meaning of this meta-path is that the two gene nodes share a common pathway node. The link type “hasPathway\(^{-1}\)” represents the inverted relation of “hasPathway”.

Different meta-paths usually represent different relationships among the linked objects with totally different semantic meanings. For example, the meta-path “gene \( \xrightarrow{\text{PPI}} \) gene” denotes the two gene nodes are connected through “PPI” links; while meta-path “gene \( \xrightarrow{\text{hasGO}} \) gene” corresponds to the semantic meaning that the two gene nodes share common gene ontology terms.
Input:  
\( G \): a heterogeneous information network, \( p_{\text{max}} \): maximum meta-path length (default=4)  
\( Y_C \): label sets for the training instances, \( A \): a base learner for local model,  
\( T_{\text{max}} \): maximum # of iteration (default=10)  

**Meta-path Construction:**  
- Construct the meta-path set for label correlations \( S_l = \{P_1, \ldots, P_{c_l}\} \):  
  - Breadth first search on schema graph of \( G \) starting from label type by adding short meta-paths that ends with label type into \( S_l \) first:  
    1. If the length of meta-path is greater than \( p_{\text{max}} \), exit the BFS;  
    2. If current meta-path \( P_j \) cannot be reconstructed by the paths in \( S_l \), add \( P_j \) into \( S_l \); Otherwise, prune the current path from BFS.  
- Construct the meta-path set for instance correlations \( S_i = \{P'_1, \ldots, P'_{c_i}\} \):  
  - similar BFS, but starting and ending with instance type.  

**Training Initialization:**  
- Learn the local model \( f \):  
  1. Construct \( q \) extended training sets \( \forall 1 \leq k \leq q, D_k = (\{x^k_i, y^k_i\}) \) by converting each instance \( x_i \) to \( x^k_i \) as follows:  
    \[ x^k_i = (x_i, \text{LabelPathFeature}(i, Y_i), \text{InstancePathFeature}(i, Y_i)) \]  
  2. Let \( I_k = f(D_k) \) be the local model trained on \( D_k \).  

**Bootstrap:**  
- Estimate the label sets, for \( i \in U \): produce an estimated values \( \hat{Y}_i \) for \( Y_i \) as follows:  
  \[ \hat{Y}_i = f((x_i, 0)) \]  

**Iterative Inference:**  
- Repeat until convergence or \#iteration > \( T_{\text{max}} \):  
  1. Construct the extended testing instance by converting each instance \( x_i \) to \( x^k_i \)'s (\( i \in U \)) as follows:  
    \[ x^k_i = (x_i, \text{LabelPathFeature}(i, Y_i), \text{InstancePathFeature}(i, Y_i)) \]  
  2. Update the estimated value \( \hat{Y}_i \) for \( Y_i \) on each testing instance (\( i \in U \)) as follows: \( \forall 1 \leq k \leq q, \hat{Y}^k_i = f_k(x^k_i) \).  

**Output:**  
\( \hat{Y}_U = (\hat{Y}_1, \cdots, \hat{Y}_n) \): the label sets of testing instances (\( i \in U \)).

![Figure 3: The PIPL algorithm](image)

### 3.2 Meta-path-based Label Correlations

Heterogeneous information networks can involve various existing knowledge about the candidate label concepts, where the complex correlations among label concepts are embedded within the network structures. For example, in Figure 1, the label nodes (i.e., the gene nodes) are linked with each other directly through “PPI” links.

It indicates one type of label correlation within the network structure: gene labels that are linked by “PPI” links can be more likely (or less likely) to appear together in the label set of a chemical compound than those without “PPI” links.

To generalize, the label concepts can also be linked by certain meta-paths, thus being correlated. For example, the gene labels are also linked with each other through the meta-path “gene hasGO\rightarrow GO hasGO^{-1}\rightarrow gene”.

It indicates another type of label correlation: gene labels that share similar gene ontology terms can be more likely (less likely) to appear together in the label set of a chemical compound than those without such meta-path links. Thus meta-paths among the label nodes, i.e., the meta-path starting and ending with the label node type, can effectively capture different types of label correlations embedded in heterogeneous information networks.

We propose to exploit meta-path-based label correlations for multi-label classification. Given a set of meta-paths among the label nodes, \( S_l = \{P_1, \cdots, P_{c_l}\} \), the meta-path-based label correlations can be used as follows:

\[ P(Y_i|x_i) \approx \prod_{k=1}^{c_l} P(Y^k_i|x_i, Y^k_{P_1(i)}, \cdots, Y^k_{P_{c_l}(i)}) \]

where \( P_j(k) \) denotes the index set of labels that are linked to the \( k \)-th label through meta-path \( P_j \in S_l \).

### 3.3 Meta-path-based Instance Correlations

Existing approaches for multi-label classification usually have i.i.d. assumptions, where the label set predictions on different instances are assumed to be independent:

\[ P(Y|X) \approx \prod_i P(Y_i|x_i) \]

However, in heterogeneous information networks, there are complex correlations not only among different labels but also among different instances. Heterogeneous information networks can involve various existing knowledge about the instances, where the complex correlations among the label sets of different instances are embedded within the network structures.

Similar to the previous subsection, the instance nodes can also be linked by certain meta-paths, thus the label sets of the linked instances can be correlated. For example, the chemical compound nodes are linked with each other through the meta-path “chemical compound\rightarrow causeSideEffects\rightarrow SideEffect\rightarrow causeSideEffects^{-1}\rightarrow chemical compounds”.

It indicates the type of instance correlations: chemical drugs that share similar side effects can be more likely (or less likely) to have similar label sets than those without such meta-path links. Thus meta-paths among the instance nodes, i.e., the meta-path starting and ending with the instance node type, can effectively capture different types of instance correlations embedded in heterogeneous information networks.

We propose to exploit meta-path-based instance correlations for multi-label classification. Given a set of meta-paths among the instance nodes, \( S_i = \{P'_1, \cdots, P'_{c_i}\} \), the meta-path-based instance correlations can be used as follows:

\[ P(Y|x) \approx \prod_i P(Y_i|x_i, Y'_{P'_1(i)}, \cdots, Y'_{P'_{c_i}(i)}) \]

where \( P'_j(i) \) denotes the index set of instances that are linked to the \( i \)-th instance through meta-path \( P'_j \in S_i \).

### 3.4 The Unified Model

In order to perform multi-label collective classification more effectively in heterogeneous information networks, in this paper, we explicitly consider both meta-path-based la-
We construct networks capturing the linkage structure of heterogeneous information shown in [28] that long meta-paths are not quite useful in simultaneously. The label set of each gene is defined in the dataset, each gene can cause or be related to multiple diseases as the instances, and diseases as the labels. In SLAP iterative inference, we first extract all non-redundant meta-paths with a maximum path length \( S \) set of the meta-paths in \( S \), then averaged with equal importance. The macro average means that the score is first computed on each label separately and then averaged with equal importance.

\[
\text{macro-F1}(h, D_U) = \frac{2 \times \sum_{i=1}^{n} \| h(x_i) \cap Y_i \|_1}{\sum_{i=1}^{n} \| h(x_i) \|_1 + \sum_{i=1}^{n} \| Y_i \|_1}
\]

The larger the value, the better the performance.

\bullet \text{Hamming loss} [8, 37]: evaluates the symmetric difference between true labels and predicted labels.

\[
\text{HammingLoss}(h, D_U) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{q} \| h(x_i) \oplus Y_i \|_1
\]

where \( \oplus \) stands for the symmetric difference of two sets. \( \| \cdot \|_1 \) denotes the \( l_1 \)-norm. The smaller the value, the better the performance.

\bullet \text{Subset 0/1 Loss} [8, 10]: evaluates 0/1 loss over the label set prediction.

\[
\text{SubsetLoss}(h, D_U) = \frac{1}{n} \sum_{i=1}^{n} I(h(x_i) \neq Y_i)
\]

\( I(\cdot) \) denotes the indicator function, i.e., \( I(\pi) = 1 \) iff \( \pi \) holds.

Figure 4: The functions of constructing relational features for meta-path-based label correlations and meta-path-based instance correlations simultaneously.

\[
P(Y|X) = \prod_{k=1}^{q} P(Y_i^k | X_i, Y_i^{P_i^k}, \ldots, Y_i^{P_i^p}(i)) \cdot \ldots
\]

In Figure 3, we summarize the proposed multi-label collective classification algorithm, called PIPL. The algorithm includes the following steps:

Meta-path Construction: Given a heterogeneous information network, we first extract all non-redundant meta-paths for label correlations and instance correlations separately. A meta-path \( P \) in \( S \) (or \( P' \) in \( S \)) is non-redundant if \( P \) (or \( P' \)) cannot be reconstructed by combining any subset of the meta-paths in \( S \) (or \( S \)). We only extract short meta-paths with a maximum path length \( p_{\text{max}} \). It has been shown in [28] that long meta-paths are not quite useful in capturing the linkage structure of heterogeneous information networks.

Training Initialization: We construct \( q \) extended training sets \( \forall 1 \leq k \leq q, \mathcal{D}_k = \{ (x^k_i, y^k_i) \} \) for each instance \( x_i \), to \( x^k_i \) using the extended training sets. We train one classifier on each label, using the extended training sets.

Iterative Inference: Overall, it is an iterative classification algorithm [22] for the inference step. During the inference, we iteratively update the label set predictions of the testing instances, and the relational features corresponding to the label and instance correlations.

4. EXPERIMENTS

4.1 Data Collections

In order to evaluate the performances of multi-label collective classification in heterogeneous information networks, we had our algorithm tested on a bioinformatic dataset SLAP [4], which is a heterogeneous network composed by 290K nodes and 720K edges. As shown in Figure 1, the SLAP dataset contains integrated data related to chemical compounds, genes, diseases, side effects, pathways etc. Specifically, there are two different prediction tasks studied in this section:

- Gene-Disease Association Prediction: The first task we studied is gene-disease association prediction, where we treat genes as the instances, and diseases as the labels. In SLAP dataset, each gene can cause or be related to multiple diseases simultaneously. The label set of each gene is defined as the set of diseases that the gene can cause. The task of gene-disease association prediction is that, we are given a set of training gene instances, and for each unlabeled gene instance, we want to predict which set of diseases the gene can cause. In this way, we could discover what kinds of diseases might be caused by a specific gene. In details, we extracted 3000 gene ontology terms (GO terms) and used them as the features of each gene instance. The top-50 diseases that are associated with the largest number of genes were used as the candidate labels. The labels of this task is highly incomplete. Thus, we kept all the genes which had at least one label, and randomly sampled additional disease instances without any direct link connection to any disease, according to 1 : 10 ratio. The final dataset contains 1943 instances, 3000 features and 50 labels in total. All the remaining nodes are kept in the network, and can be used in meta-paths.

- Drug-Target Binding Prediction: The second task we studied is drug-target binding prediction, where we treat chemical compounds as the instances and genes as the labels. We first represent the structure of each chemical compound as a graph object. Then we used the multi-label subgraph mining algorithm [19] to extract 1500 subgraphs from the dataset and used them as features of the instances. Similar to the previous task, the top-50 genes that are binded with the largest number of chemical compounds were used as the candidate labels. We removed the instances with no feature. The dataset contains 5651 instances, 1500 features and 50 labels in total. All the remaining nodes are kept in the network, and can be used in meta-paths.

4.2 Evaluation Metrics

Multi-label classification problems require more complicated criteria for performance evaluation than conventional single-label classification problems. We use some evaluation criteria in [10, 14, 21, 37, 8] to verify the multi-label classification performance in heterogeneous information networks.

Given a multi-label dataset \( D_U \) involving \( n \) instances, i.e., \( (x_i, Y_i) \), where \( Y_i \in \{0,1\}^n \) (\( i = 1, \ldots, n \)). Denote \( h(x_i) \) as the predicted label set for \( x_i \), by a multi-label classifier \( h \).

We have the following evaluation criteria:

- Micro \( F1 \) [10, 14, 21]: is the harmonic mean of micro average of Precision and Recall. The micro average means that the score is first computed on each label separately and then averaged with equal importance.

- Hamming Loss [8, 37]: evaluates the symmetric difference between true labels and predicted labels.

- Subset 0/1 Loss [8, 10]: evaluates 0/1 loss over the label set prediction.
Type of Classification

- Multi-Label Collective Classification
- Collective Classification
- Binary Classification

Type of Correlations Exploited

- label correlation from data samples
- instance correlation from heterogeneous network
- instance correlation from homogeneous network
- label correlation from heterogeneous network
- meta-path-based instance correlation
- label correlations from heterogeneous network

Publication

- This paper
- [27]
- [18]
- [17]

4.3 Compared Methods

In order to demonstrate the effectiveness of our multi-label classification approach, we compared the following methods (summarized in Table 1):

- **BSVM** (binary SVM): The first baseline method uses binary decomposition method to solve multi-label classification problems, which is similar to [2]. The multi-label dataset is first divided into multiple single-label datasets by one-vs-all binary decomposition. For each binary classification task, we use the SVM as the base classifier. Then the predictions of SVMs for all labels are combined to make the final prediction. BSVM assumes all the labels and all instances are independent.

- **ECC** (multi-label classification + ensemble): This baseline method is an ensemble of classifier chains (CC) [27]. The ensemble is created by training different classifier chains using randomly sampled subset of instances with random label orders.

- **PISL** (binary decomposition + meta-path based instance correlation): For each binary classification method, we use the meta-path based collective classification method in heterogeneous information networks [18] to perform the binary collective classification. This method can exploit the meta-path based instance correlation within a heterogeneous information network.

- **ICML** (simple label correlation + instance correlation in homogeneous network): This method was proposed in [17] which can exploit relational features for instance dependencies based upon homogeneous network for multi-label collective classification. Since this method requires a homogeneous network among the instances, we can only run this method on the Gene-Disease association prediction task, where the PPI network is used as the homogeneous network among the gene instances.

- **PIML** (simple label correlation + meta-path based instance correlation): This method is extended from PISL [18] by adding relational features according to instance-cross-label dependencies for multi-label collective classification [17]. This method can only exploit the instance correlations from heterogeneous networks. However, the label correlations used in this method are directly derived from data samples instead of using heterogeneous networks.

- **PIPL** (meta-path based instance and label correlations in heterogeneous networks): This method combines meta-path based instance correlations from heterogeneous networks with label correlations from heterogeneous networks.

Table 1: Summary of compared methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>Type of Classification</th>
<th>Type of Correlations Exploited</th>
<th>Publication</th>
</tr>
</thead>
<tbody>
<tr>
<td>BSVM</td>
<td>Binary Classification</td>
<td>all independent</td>
<td>[2]</td>
</tr>
<tr>
<td>ECC</td>
<td>Multi-Label Classification</td>
<td>label correlation from data samples</td>
<td>[27]</td>
</tr>
<tr>
<td>PISL</td>
<td>Collective Classification</td>
<td>instance correlation from heterogeneous network</td>
<td>[18]</td>
</tr>
<tr>
<td>ICML</td>
<td>Multi-Label Collective Classification</td>
<td>instance correlation from data samples</td>
<td>[17]</td>
</tr>
<tr>
<td>PIPL</td>
<td>Multi-Label Collective Classification</td>
<td>meta-path-based instance correlation</td>
<td>This paper</td>
</tr>
<tr>
<td>PIML</td>
<td>Multi-Label Collective Classification</td>
<td>label correlations from heterogeneous network</td>
<td>This paper</td>
</tr>
</tbody>
</table>

Table 2: Classification performances “average score ± std (rank)” on gene-disease association prediction task. “↓” indicates the smaller the value the better the performance; “↑” indicates the larger the value the better the performance.

<table>
<thead>
<tr>
<th>criteria</th>
<th># label</th>
<th>BSVM</th>
<th>ECC</th>
<th>ICML</th>
<th>PISL</th>
<th>PIPL</th>
<th>PIML</th>
</tr>
</thead>
<tbody>
<tr>
<td>Micro-F1</td>
<td>10</td>
<td>0.360±0.082</td>
<td>0.387±0.073</td>
<td>0.366±0.079</td>
<td>0.390±0.115</td>
<td>0.399±0.107</td>
<td>0.400±0.106</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0.385±0.046</td>
<td>0.406±0.043</td>
<td>0.389±0.066</td>
<td>0.417±0.055</td>
<td>0.426±0.055</td>
<td>0.433±0.066</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>0.317±0.035</td>
<td>0.359±0.027</td>
<td>0.343±0.039</td>
<td>0.354±0.047</td>
<td>0.355±0.013</td>
<td>0.360±0.007</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>0.342±0.045</td>
<td>0.386±0.032</td>
<td>0.339±0.042</td>
<td>0.382±0.034</td>
<td>0.387±0.030</td>
<td>0.391±0.030</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>0.303±0.055</td>
<td>0.346±0.059</td>
<td>0.321±0.063</td>
<td>0.348±0.064</td>
<td>0.360±0.075</td>
<td>0.366±0.078</td>
</tr>
<tr>
<td>Hamming Loss</td>
<td>10</td>
<td>0.011±0.002</td>
<td>0.013±0.002</td>
<td>0.011±0.002</td>
<td>0.011±0.003</td>
<td>0.011±0.003</td>
<td>0.011±0.002</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0.008±0.001</td>
<td>0.010±0.000</td>
<td>0.008±0.001</td>
<td>0.008±0.001</td>
<td>0.008±0.001</td>
<td>0.008±0.001</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>0.008±0.000</td>
<td>0.009±0.000</td>
<td>0.007±0.000</td>
<td>0.007±0.001</td>
<td>0.007±0.001</td>
<td>0.007±0.000</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>0.007±0.000</td>
<td>0.007±0.000</td>
<td>0.007±0.000</td>
<td>0.006±0.000</td>
<td>0.006±0.000</td>
<td>0.006±0.000</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>0.006±0.001</td>
<td>0.007±0.001</td>
<td>0.006±0.001</td>
<td>0.006±0.001</td>
<td>0.006±0.001</td>
<td>0.006±0.001</td>
</tr>
<tr>
<td>Subset 0/1 Loss</td>
<td>10</td>
<td>0.108±0.023</td>
<td>0.125±0.020</td>
<td>0.107±0.020</td>
<td>0.103±0.024</td>
<td>0.103±0.024</td>
<td>0.103±0.023</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0.153±0.008</td>
<td>0.180±0.011</td>
<td>0.154±0.006</td>
<td>0.148±0.005</td>
<td>0.148±0.005</td>
<td>0.147±0.007</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>0.197±0.009</td>
<td>0.212±0.006</td>
<td>0.191±0.010</td>
<td>0.195±0.011</td>
<td>0.192±0.010</td>
<td>0.191±0.009</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>0.227±0.018</td>
<td>0.238±0.016</td>
<td>0.228±0.017</td>
<td>0.219±0.018</td>
<td>0.218±0.014</td>
<td>0.216±0.015</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>0.255±0.026</td>
<td>0.275±0.029</td>
<td>0.250±0.030</td>
<td>0.244±0.028</td>
<td>0.243±0.028</td>
<td>0.241±0.028</td>
</tr>
</tbody>
</table>

otherwise $I(π) = 0$. The smaller the value, the better the performance.
We first study the effectiveness of the proposed PIPL method on multi-label collective classification in heterogeneous information networks. In our experiments, 5-fold cross validations are performed on each task to evaluate the performances of multi-label collective classification. We show the detailed results in Table 2 and Table 3, with the number of candidate labels varying from 10 to 50. The average performances of the five methods with standard deviations are reported with respect to ten different evaluation metrics.

In both tasks, we can observe that the Bsvm method, which assumes all instances and labels are independent, is outperformed by other methods, which explicitly consider correlations among predictions from various aspects. These results demonstrate the importance of exploiting different types of correlations for multi-label collective classification in heterogeneous information networks. In details, PIsl can improve performances over Bsvm because it exploits the meta-path based instance correlation. These results support our first assumption that in multi-label heterogeneous information networks, the label sets of related instances through instance meta-paths are not independent, and should be classified collectively. Similar conclusions can also be drawn from the results that PIml outperforms Ecc. Though both methods can exploit simple label correlation, PIml can further exploit the meta-path based instance correlations, while Ecc assumes that the label sets of different instances are independent, and are predicted independently. Thus it is necessary and important to exploit the meta-path based correlations among the related instances.

We also observe that in gene-disease association prediction task, meta-path based methods designed for heterogeneous networks (i.e., PIsl, PIml, PIPL) can achieve better performances than the ICml that only exploits the homogeneous network among instances.

We further observe that the proposed PIPL performs the best among all compared methods. Especially, PIPL outperforms PIml and PIsl, by taking meta-path based label correlation into consideration. In heterogeneous information networks, both of instances and candidate labels can be correlated with each other via diverse semantic meanings. In Table 4, we show four examples of the meta paths used by PIPL method in both tasks, which correspond to label correlations and instance correlations separately. Such complex correlations can be effectively exploited through instance and label meta-paths and thus boost the classification performances of multi-label collective classification in heterogeneous information networks.

### 4.5 Sensitivity to Number Features

In previous subsection, we extracted 3000 GO terms and used them as features for Gene-Disease association prediction task. For Drug-Target binding prediction task, we extracted 1500 substructures of the chemical compounds and used them as features. To test the stability of the performances of PIPL method, we show the micro-F1 scores of each method under different numbers of features (ranging from 500 to 5000 in Gene-Disease association prediction task, and from 400 to 2400 in Drug-Target binding prediction task) with 50 candidate labels. Figure 5(a) and Figure 5(b) illustrate that PIPL performs quite well and stably, and it is not sensitive to the number of features.

### 5. RELATED WORK

Multi-label classification has been extensively studied in recent years [33, 11, 15, 32, 5, 25, 12]. Conventional methods can be roughly categorized as follows: (1) binary approaches, which converts the multi-label problem into multiple independent binary classification problems (one for each label) [2, 36] (2) pairwise approaches which exploit the pairwise relation between different labels [10, 9]. (3) High-order approaches: The third type of approaches considers
Table 4: Examples of meta-paths used in PIPL method

<table>
<thead>
<tr>
<th>Task</th>
<th>Meta-path</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene-Disease Prediction</td>
<td>Disease $\rightarrow$ Disease, Chemical compound $\rightarrow$ Chemical compound $\rightarrow$ Disease</td>
<td>label correlation</td>
</tr>
<tr>
<td></td>
<td>Gene $\rightarrow$ Disease, Chemical compound $\rightarrow$ Cancer $\rightarrow$ Disease</td>
<td>instance correlation</td>
</tr>
<tr>
<td>Drug-Target Prediction</td>
<td>Gene $\rightarrow$ Disease, Chemical compound $\rightarrow$ Chemical compound $\rightarrow$ Gene</td>
<td>label correlation</td>
</tr>
<tr>
<td></td>
<td>Chemical compound $\rightarrow$ Gene, Chemical compound $\rightarrow$ Gene</td>
<td>instance correlation</td>
</tr>
</tbody>
</table>

Figure 5: Micro-F1 scores with different number of features.

(a) Gene-Disease Association Prediction (b) Drug-Target Binding Prediction

Table 5: Summary of related datasets used in network classification.

<table>
<thead>
<tr>
<th>dataset</th>
<th>publication</th>
<th>network</th>
<th>type of classification</th>
<th>#types</th>
<th>#Nodes</th>
<th>Node Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene</td>
<td>[34]</td>
<td>homogeneous</td>
<td>single-label</td>
<td>(1,1)</td>
<td>1,243</td>
<td>gene</td>
</tr>
<tr>
<td>Citeseer</td>
<td>[22]</td>
<td>homogeneous</td>
<td>single-label</td>
<td>(1,1)</td>
<td>3,312</td>
<td>paper</td>
</tr>
<tr>
<td>WebKB</td>
<td>[7]</td>
<td>homogeneous</td>
<td>single-label</td>
<td>(1,1)</td>
<td>3,877</td>
<td>webpage</td>
</tr>
<tr>
<td>DBLP</td>
<td>[17]</td>
<td>homogeneous</td>
<td>multi-label</td>
<td>(1,1)</td>
<td>4638</td>
<td>author</td>
</tr>
<tr>
<td>ACM</td>
<td>[18]</td>
<td>heterogeneous</td>
<td>single-label</td>
<td>(5,5)</td>
<td>12,499</td>
<td>paper, author, conference, proceeding, school</td>
</tr>
<tr>
<td>Cora</td>
<td>[23]</td>
<td>heterogeneous</td>
<td>single-label</td>
<td>(5,5)</td>
<td>4,330</td>
<td>paper, author, journal, publisher, editor</td>
</tr>
<tr>
<td>IMDB</td>
<td>[1]</td>
<td>heterogeneous</td>
<td>single-label</td>
<td>(5,5)</td>
<td>1,382</td>
<td>movie, actor, director, studio, producer</td>
</tr>
<tr>
<td>NASD</td>
<td>[24]</td>
<td>heterogeneous</td>
<td>single-label</td>
<td>(5,5)</td>
<td>22,000</td>
<td>broker, branch, firm, disclosure, regulator</td>
</tr>
<tr>
<td>SLAP</td>
<td>this paper</td>
<td>heterogeneous</td>
<td>multi-label</td>
<td>(10,11)</td>
<td>280,792</td>
<td>drug, disease, gene, side effect, pathway, tissue, GO, gene family, substructure, CO</td>
</tr>
</tbody>
</table>

the high-order correlations among different labels. Such approaches includes random subset ensemble approaches [26, 27], Bayesian network based approach [35] and full-order approaches [6, 8]. [17] studied the multi-label collective classification using homogeneous networks.

Heterogeneous information networks have attracted much attention in recent years [31, 29, 13]. Sun et al. [31, 28] studied the clustering problem and top-k similarity problem in heterogeneous information networks. Ji et al. studied a specialized classification problem on heterogeneous networks, where different types of nodes share a same set of label concepts [13]. [18, 16] studied collective classification problem in heterogeneous information networks.

In Table 5 we reviewed the datasets used in previous works on network classification. It shows that among all datasets, SLAP dataset is one of the best choices for studying multi-label classification in heterogeneous information networks.

6. CONCLUSION

In this paper, we studied how to exploit heterogeneous information networks for multi-label collective classification. Conventional multi-label learning approaches usually assume that the label correlations are either given or can be derived directly from data samples by counting their co-occurrences. We proposed to use heterogeneous information networks to facilitate the learning process of multi-label classification by mining label correlations and instance correlations from the network. We propose a novel solution to multi-label classification, called PIPL by exploiting complex linkage information in heterogeneous information networks. Empirical
studies on real-world tasks demonstrate that the proposed multi-label classification approach can effectively boost classification performances.

7. ACKNOWLEDGEMENTS

This work is supported in part by NSF through grants HS-0905215, CNS-1115234, IIS-0914934, DBI-0960443, and OSE-1129076, US Department of Army through grant W911NF-12-1-0066, and Huawei grant.

8. REFERENCES