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 space 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

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-
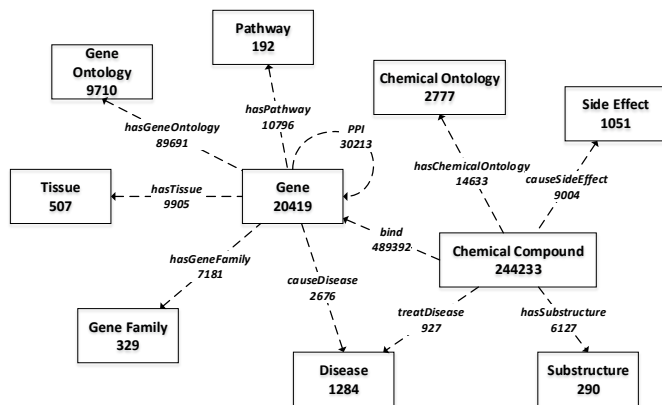
Figure 1: An example of heterogeneous information network (the data schema of SLAP network). Each box represents one type of nodes in the network, and each dashed line represents one type of links. Each number under the node/link represents the total number of nodes/links of the same type.

Figure 2: Multi-label classification by mining heterogeneous information networks.

tion network [1] as Figure 1 that contains abundant knowledge about the relationships among different types of entities including *chemical compounds* and *gene targets*, we can make use of the domain knowledge within this network to facilitate multi-label classification.

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 instances and labels 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.

---

[1]SLAP dataset [4]: this dataset is an information network that integrates many datasets into a single framework using Semantic Web technologies for drug discovery. It includes public datasets related to systems chemical biology: such as PubChem, DrugBank, PPI, SIDER, CTD diseases, KEGG Pathways, *etc.*
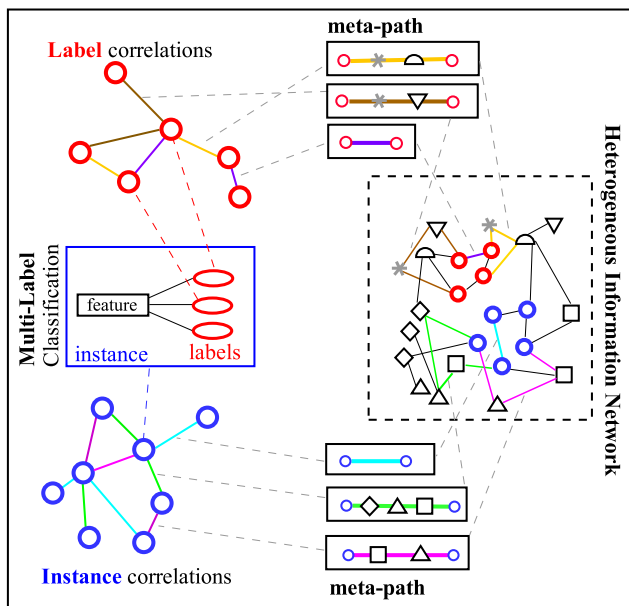
2.1 Multi-label Classification

Suppose we have a multi-label classification task with n instances and q labels. The set of all instances is denoted as $\mathcal{V}_I = \{I_1, \cdots, I_n\}$, where each instance I_i has a feature vector $\mathbf{x}_i$ in the input space. The set of all candidate labels is denoted as $\mathcal{V}_\ell = \{\ell_1, \cdots, \ell_q\}$. Without loss of generality, we assume the first n_ℓ instances in $\mathcal{V}_I$, denoted as $\mathcal{V}_I^\ell = \{I_1, \cdots, I_{n_\ell}\} \subset \mathcal{V}_I$, are labeled, where each instance I_i is associated with a label set, denoted as $\mathbf{Y}_i = \left(Y_i^1, \cdots, Y_i^q\right)^\top \in \{0, 1\}^q$. 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 $\mathcal{V}_I$ are unlabeled, i.e., $\mathcal{V}_I^u = \mathcal{V}_I - \mathcal{V}_I^\ell$. The task of multi-label classification is to predict the label sets for all the unlabeled instances in $\mathcal{V}_I^u$. In other words, for any possible label set $\mathbf{Y} \in \{0, 1\}^q$ and any instance I_i, we need to build a model to estimate the probability $P(\mathbf{Y}|\mathbf{x}_i)$ 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 $P(\mathbf{Y}|\mathbf{x}_i)$ for instance I_i approximately by assuming that all labels are independent:

$$\forall i, \quad P(\mathbf{Y}|\mathbf{x}_i) \approx \prod_{k=1}^{q} P(Y^k|\mathbf{x}_i)$$

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 $P(\mathbf{Y}|\mathbf{x}_i)$ as follows:

$$P(\mathbf{Y}|\mathbf{x}_i) \approx P(Y^1|\mathbf{x}_i) \ P(Y^2|Y_i^1, \mathbf{x}_i) \ P(Y_i^3|Y^1, Y^2, \mathbf{x}_i) \cdots$$

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 $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. $\mathcal{V}$ denotes the set of nodes, which involves t types of objects: $\mathcal{V}_1 = \{v_{11}, \cdots, v_{1n_1}\}, \cdots, \mathcal{V}_t = \{v_{t1}, \cdots, v_{tn_t}\}$, where v_{ij} represents the j-th object of type i. $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ denotes the links between the objects in $\mathcal{V}$, which includes multiple types of links.

First we define the type of nodes in the heterogeneous information network $\mathcal{G}$ 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 binding 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 $R(v_{ip}, v_{jq})$ holds if object v_{ip} and v_{jq} 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 $R(v_{ip}, v_{jq})$ holds if the *gene* node v_{ip} has a link of type "hasPathway" to the pathway node v_{jq} in the network. We can write the link type as "gene $\xrightarrow{hasPathway}$ pathway" or "$\mathcal{V}_i \xrightarrow{R} \mathcal{V}_j$".

A *meta-path* $\mathcal{P}$ 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 $\xrightarrow{hasPathway}$ pathway $\xrightarrow{hasPathway^{-1}}$ 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 $\xrightarrow{PPI}$ gene" denotes the two *gene* nodes are connected through "PPI" links; while meta-path "gene $\xrightarrow{hasGO}$ GO $\xrightarrow{hasGO^{-1}}$ gene" corresponds to the semantic meaning that the two *gene* nodes share common *gene ontology* terms.

Input:
    $\mathcal{G}$: a heterogeneous information network, p_max,: maximum meta-path length (default=4) $\mathcal{X}$: attribute vectors for all instances.
    $\mathcal{Y}_\mathcal{L}$: label sets for the training instances,   A: a base learner for local model, $T_{\max}$: maximum # of iteration (default=10)
Meta-path Construction:
  - Construct the meta-path set for label correlations $\mathcal{S}_\ell = \{\mathcal{P}_1, \cdots, \mathcal{P}_{c_\ell}\}$:
    Breadth first search on schema graph of $\mathcal{G}$ starting from label type by adding short meta-paths that ends with label type into $\mathcal{S}_\ell$ first:
        1. If the length of meta-path is greater than p_max, exit the BFS;
        2. If current meta-path $\mathcal{P}_j$ cannot be reconstructed by the paths in $\mathcal{S}_\ell$, add $\mathcal{P}_j$ into $\mathcal{S}_\ell$; Otherwise, prune the current path from BFS.
  - Construct the meta-path set for instance correlations $\mathcal{S}_I = \{\mathcal{P}'_1, \cdots, \mathcal{P}'_{c_I}\}$:
    similar BFS, but starting and ending with instance type.
Training Initialization:
  - Learn the local model f:
    1. Construct q extended training sets $\forall 1 \leq k \leq q, \mathcal{D}_k = \left\{ (\mathbf{x}_i^k, y_i^k) \right\}$ by converting each instance $\mathbf{x}_i$ to $\mathbf{x}_i^k$ as follows:
        $\mathbf{x}_i^k = (\mathbf{x}_i, \text{LabelPathFeature}(\ell_k, \mathbf{Y}_i), \text{InstancePathFeature}(i, \mathcal{Y}_\mathcal{L}))$
    2. Let $f_k = A(\mathcal{D}_k)$ be the local model trained on $\mathcal{D}_k$.
Bootstrap:
  - Estimate the label sets, for $i \in \mathcal{U}$: produce an estimated values $\hat{\mathbf{Y}}_i$ for $\mathbf{Y}_i$ as follows: $\hat{Y}_i = f((\mathbf{x}_i, \mathbf{0}))$ using attributes only.
Iterative Inference:
  - Repeat until convergence or #iteration> $T_{\max}$
    1. Construct the extended testing instance by converting each instance $\mathbf{x}_i$ to $\mathbf{x}_i^k$'s ($i \in \mathcal{U}$) as follows:
        $\mathbf{x}_i^k = \left(\mathbf{x}_i, \text{LabelPathFeature}(\ell_k, \hat{\mathbf{Y}}_i), \text{InstancePathFeature}(i, \mathcal{Y}_\mathcal{L} \cup \{\hat{\mathbf{Y}}_i | i \in \mathcal{U}\}, \mathcal{S}) \right)$
    2. Update the estimated value $\hat{\mathbf{Y}}_i$ for $\mathbf{Y}_i$ on each testing instance ($i \in \mathcal{U}$) as follows: $\forall 1 \leq k \leq q, \hat{Y}_i^k = f_k(\mathbf{x}_i^k)$.
Output:
  $\hat{\mathcal{Y}}_\mathcal{U} = \left(\hat{\mathbf{Y}}_1, \cdots, \hat{\mathbf{Y}}_{n_u} \right)$:   the label sets of testing instances ($i \in \mathcal{U}$).

Figure 3: The PIPL algorithm

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$\xrightarrow{hasGO}$GO$\xrightarrow{hasGO^{-1}}$gene". It indicates another type of label correlations: *gene* labels that share similar *gene ontology* terms can be more likely (or 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, $\mathcal{S}_\ell = \{\mathcal{P}_1, \cdots, \mathcal{P}_{c_\ell}\}$, the meta-path-based label correlations can be used as follows:

$$\forall i, \ P(\mathbf{Y}_i | \mathbf{x}_i) \approx \prod_{k=1}^{q} P(\mathbf{Y}_i^k | \mathbf{x}_i, \mathbf{Y}_i^{\mathcal{P}_1(k)}, \cdots, \mathbf{Y}_i^{\mathcal{P}_{c_\ell}(k)})$$

where $\mathcal{P}_j(k)$ denotes the index set of labels that are linked to the k-th label through meta-path $\mathcal{P}_j \in \mathcal{S}_\ell$.

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:

$$P(\mathcal{Y}|\mathcal{X}) \approx \prod_i P(\mathbf{Y}_i | \mathbf{x}_i)$$

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$\xrightarrow{causeSideEffect}$ Side Effect $\xrightarrow{causeSideEffect^{-1}}$ 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, $\mathcal{S}_I = \{\mathcal{P}'_1, \cdots, \mathcal{P}'_{c_I}\}$, the meta-path-based instance correlations can be used as follows:

$$P(\mathcal{Y}|\mathcal{X}) \approx \prod_i P(\mathbf{Y}_i | \mathbf{x}_i, \mathbf{Y}_{\mathcal{P}'_1(i)}, \cdots, \mathbf{Y}_{\mathcal{P}'_{c_I}(i)})$$

where $\mathcal{P}'_j(i)$ denotes the index set of instances that are linked to the i-th instance through meta-path $\mathcal{P}'_j \in \mathcal{S}_I$.

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-

- $\mathbf{x}_\ell = \text{LabelPathFeature}(\ell_k, \mathbf{Y}_i)$
  For each meta-path $\mathcal{P}_j \in \mathcal{S}_\ell$:
  1. Get related labels for node ℓ_k through meta-path $\mathcal{P}_j$
     i.e., the related index set $\mathcal{C} = \mathcal{P}_j(i)$
  2. $\mathbf{x}_j = Aggregation\left(\{Y_i^k | k \in \mathcal{C}\}\right)$
     Return $(\cdots, \mathbf{x}_j^\top, \cdots)^\top$

- $\mathbf{x}_I = \text{InstancePathFeature}(i, \mathcal{Y})$
  For each meta-path $\mathcal{P}_j' \in \mathcal{S}_I$:
  1. Get related instances for node I_i through meta-path $\mathcal{P}_j'$,
     i.e. the related index set $\mathcal{C} = \mathcal{P}_j'(i)$
  2. $\mathbf{x}_j = Aggregation\left(\{\mathbf{Y}_i | i \in \mathcal{C}\}\right)$
     Return $(\cdots, \mathbf{x}_j^\top, \cdots)^\top$

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

bel correlations and meta-path-based intance correlations simultaneously.

$$P(\mathcal{Y}|\mathcal{X}) \approx \prod_i \prod_{k=1}^q P(\mathbf{Y}_i^k | \mathbf{x}_i, \mathbf{Y}_i^{\mathcal{P}_j(k)}, \cdots, \mathbf{Y}_{\mathcal{P}_{j'}'(i)}^k, \cdots)$$

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 $\mathcal{P}_j$ in $\mathcal{S}_\ell$ (or $\mathcal{P}_j'$ in $\mathcal{S}_I$) is non-redundant if $\mathcal{P}_j$ (or $\mathcal{P}_j'$) cannot be reconstructed by combining any subset of the meta-paths in $\mathcal{S}_\ell$ (or $\mathcal{S}_I$). We only extract short meta-paths with a maximum path length p_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 $\forall 1 \leq k \leq q, \mathcal{D}_k = \left\{(\mathbf{x}_i^k, y_i^k)\right\}$ by converting each instance $\mathbf{x}_i$ to $\mathbf{x}_i^k$ using the functions in Figure 4. 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 over $290K$ 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 gene 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 $\mathcal{D}_U$ involving n instances, *i.e.*, $(\boldsymbol{x}_i, \boldsymbol{Y}_i)$. Here $\boldsymbol{Y}_i \in \{0,1\}^q$ $(i = 1, \cdots, n)$. Denote $h(\boldsymbol{x}_i)$ as the predicted label set for $\boldsymbol{x}_i$ 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.

$$\text{micro-F1}(h, \mathcal{D}_U) = \frac{2 \times \sum_{i=1}^n \|h(\boldsymbol{x}_i) \cap \boldsymbol{Y}_i\|_1}{\sum_{i=1}^n \|h(\boldsymbol{x}_i)\|_1 + \sum_{i=1}^n \|\boldsymbol{Y}_i\|_1}$$

The larger the value, the better the performance.

• *Hamming loss* [8, 37]: evaluates the symetric difference between true labels and predicted labels.

$$\text{HammingLoss}(h, \mathcal{D}_U) = \frac{1}{n} \sum_{i=1}^n \frac{1}{q} \|h(\boldsymbol{x}_i) \oplus \boldsymbol{Y}_i\|_1$$

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.

• *Subset 0/1 Loss* [8, 10]: evaluates 0/1 loss over the label set prediction.

$$\text{SubsetLoss}(h, \mathcal{D}_U) = \frac{1}{n} \sum_{i=1}^n I\left(h(\boldsymbol{x}_i) \neq \boldsymbol{Y}_i\right)$$

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

Table 1: Summary of compared methods.

Method	Type of Classification	Type of Correlations Exploited	Publication
Bsvm	Binary Classification	all independent	[2]
Ecc	Multi-Label Classification	① label correlcation from data samples	[27]
PIsl	Collective Classification	② instance correlations from heterogeneous network	[18]
Icml	Multi-Label Collective Classification	① label correlation from data samples ③ instance correlation from homogeneous network	[17]
PIml	Multi-Label Collective Classification	② instance correlations from heterogeneous network	This paper
PIPL	Multi-Label Collective Classification	② meta-path-based instance correlation ④ label correlations from heterogeneous network	This paper

Table 2: Classification performances "average score $\pm$ 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.

criteria	#label	methods					
		Bsvm	Ecc	Icml	PIsl	PIml	PIPL
Micro-F1 ↑	10	0.360 ± 0.082 (6)	0.387 ± 0.073 (4)	0.366 ± 0.079 (5)	0.390 ± 0.115 (3)	0.399 ± 0.107 (2)	0.400 ± 0.106 (1)
	20	0.385 ± 0.046 (6)	0.406 ± 0.043 (4)	0.389 ± 0.066 (5)	0.417 ± 0.055 (3)	0.426 ± 0.055 (2)	0.433 ± 0.066 (1)
	30	0.317 ± 0.035 (6)	0.359 ± 0.027 (2)	0.343 ± 0.039 (5)	0.342 ± 0.037 (4)	0.355 ± 0.013 (3)	0.360 ± 0.007 (1)
	40	0.342 ± 0.045 (5)	0.386 ± 0.032 (3)	0.339 ± 0.042 (6)	0.382 ± 0.034 (4)	0.387 ± 0.030 (2)	0.391 ± 0.030 (1)
	50	0.303 ± 0.055 (6)	0.346 ± 0.059 (4)	0.321 ± 0.063 (5)	0.348 ± 0.064 (3)	0.360 ± 0.075 (2)	0.366 ± 0.078 (1)
Hamming Loss ↓	10	0.011 ± 0.002 (1)	0.013 ± 0.002 (6)	0.011 ± 0.002 (1)	0.011 ± 0.003 (1)	0.011 ± 0.003 (1)	0.011 ± 0.002 (1)
	20	0.008 ± 0.001 (1)	0.010 ± 0.000 (6)	0.008 ± 0.000 (1)	0.008 ± 0.001 (1)	0.008 ± 0.001 (1)	0.008 ± 0.001 (1)
	30	0.008 ± 0.000 (5)	0.009 ± 0.000 (6)	0.007 ± 0.000 (1)	0.007 ± 0.001 (1)	0.007 ± 0.001 (1)	0.007 ± 0.000 (1)
	40	0.007 ± 0.000 (4)	0.007 ± 0.000 (4)	0.007 ± 0.000 (4)	0.006 ± 0.000 (1)	0.006 ± 0.000 (1)	0.006 ± 0.000 (1)
	50	0.006 ± 0.001 (1)	0.007 ± 0.001 (6)	0.006 ± 0.001 (1)	0.006 ± 0.001 (1)	0.006 ± 0.001 (1)	0.006 ± 0.001 (1)
Subset 0/1 Loss ↓	10	0.108 ± 0.023 (5)	0.125 ± 0.020 (6)	0.107 ± 0.020 (4)	0.103 ± 0.024 (1)	0.103 ± 0.024 (1)	0.103 ± 0.023 (1)
	20	0.153 ± 0.008 (4)	0.180 ± 0.011 (6)	0.154 ± 0.006 (5)	0.148 ± 0.005 (2)	0.148 ± 0.009 (2)	0.147 ± 0.007 (1)
	30	0.197 ± 0.009 (4)	0.212 ± 0.006 (6)	0.191 ± 0.010 (3)	0.195 ± 0.011 (5)	0.192 ± 0.010 (2)	0.191 ± 0.009 (1)
	40	0.227 ± 0.018 (4)	0.238 ± 0.016 (5)	0.228 ± 0.017 (1)	0.219 ± 0.018 (3)	0.218 ± 0.014 (2)	0.216 ± 0.015 (1)
	50	0.255 ± 0.026 (5)	0.275 ± 0.029 (6)	0.250 ± 0.030 (4)	0.244 ± 0.028 (3)	0.243 ± 0.028 (2)	0.241 ± 0.028 (1)

otherwise $I(\pi) = 0$. The smaller the value, the better the performance.

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): We compare with another baseline using binary decomposition. For each binary classification task, we use the meta-path based collective

classification method in heterogenous information networks [18] to perform the binary collective classification. This method can exploit the meta-path based instance correlation within a heterogenous information network.

- Icml (simple label correlation + instance correlation in homogeneous network): This method was proposed in [17] which can exploit relational features for *inter-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 *inter-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 correla-

Table 3: Classification performances "average score ± std (rank)" on drug-target binding prediction task. "↓" indicates the smaller the value the better the performance; "↑" indicates the larger the value the better the performance.

criteria	#label	methods				
		Bsvm	Ecc	PIsl	PIml	PIPL
Micro-F1 ↑	10	0.532±0.046 (5)	0.576±0.053 (4)	0.608±0.046 (3)	0.611±0.040 (2)	0.625±0.042 (1)
	20	0.553±0.019 (5)	0.588±0.018 (4)	0.696±0.016 (3)	0.714±0.011 (2)	0.724±0.011 (1)
	30	0.536±0.052 (5)	0.585±0.054 (4)	0.674±0.032 (3)	0.695±0.025 (2)	0.706±0.026 (1)
	40	0.523±0.018 (5)	0.568±0.022 (4)	0.599±0.022 (3)	0.618±0.022 (2)	0.642±0.022 (1)
	50	0.521±0.028 (5)	0.571±0.036 (4)	0.603±0.031 (3)	0.635±0.028 (2)	0.653±0.026 (1)
Hamming Loss ↓	10	0.024±0.003 (5)	0.021±0.003 (4)	0.020±0.003 (2)	0.020±0.002 (3)	0.018±0.002 (1)
	20	0.019±0.001 (5)	0.017±0.000 (4)	0.012±0.001 (3)	0.012±0.001 (2)	0.011±0.001 (1)
	30	0.018±0.002 (5)	0.016±0.002 (4)	0.012±0.001 (3)	0.011±0.000 (2)	0.010±0.000 (1)
	40	0.017±0.001 (5)	0.015±0.001 (4)	0.014±0.001 (3)	0.013±0.001 (2)	0.012±0.001 (1)
	50	0.016±0.001 (5)	0.014±0.001 (4)	0.013±0.001 (3)	0.012±0.001 (2)	0.011±0.001 (1)
Subset 0/1 Loss ↓	10	0.147±0.012 (5)	0.128±0.017 (4)	0.123±0.011 (2)	0.124±0.010 (3)	0.113±0.010 (1)
	20	0.222±0.009 (5)	0.193±0.006 (4)	0.165±0.011 (3)	0.163±0.010 (2)	0.148±0.004 (1)
	30	0.265±0.019 (5)	0.223±0.029 (4)	0.214±0.007 (3)	0.207±0.004 (2)	0.182±0.003 (1)
	40	0.305±0.008 (5)	0.250±0.004 (2)	0.268±0.010 (4)	0.257±0.010 (3)	0.223±0.010 (1)
	50	0.351±0.009 (5)	0.288±0.018 (2)	0.306±0.013 (4)	0.288±0.020 (3)	0.261±0.017 (1)

tion): The proposed method for multi-label collective classification in heterogenous information networks. The only difference between PIPL and PIml is that PIml does not consider the meta-path based label correlation.

For a fair comparison, we use LibSVM [3] with linear kernel and default parameter as the base classifier for all the compared methods. The maximum number of iterations in the methods PIPL, PIml and PIsl are all set as 10.

4.4 Performances of Multi-Label Classification

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 three 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 heterogenous information networks, the label sets of related instances through instance meta-pahts 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 predic-

tion task, meta-path based methods designed for heterogeous 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

Task	Meta-path	Correlation
Gene-Disease Prediction	Disease $\xrightarrow{treated}$ Chemical compound $\xrightarrow{treat}$ Disease	label correlation
	Disease $\xrightarrow{treated}$ Chemical compound $\xrightarrow{has}$ Substructure $\xrightarrow{in}$ Chemical compound $\xrightarrow{treat}$ Disease	label correlation
	Gene $\xrightarrow{cause}$ Disease $\xrightarrow{treated}$ Chemical compound $\xrightarrow{bind}$ Gene	instance correlation
	Gene $\xrightarrow{binded}$ Chemical compound $\xrightarrow{cause}$ Side effect $\xrightarrow{caused}$ Chemical compound $\xrightarrow{bind}$ Gene	instance correlation
Drug-Target Prediction	Gene $\xrightarrow{PPI}$ Gene	label correlation
	Gene $\xrightarrow{has}$ pathway $\xrightarrow{has}$ Gene	label correlation
	Chemical compound $\xrightarrow{bind}$ Gene $\xrightarrow{PPI}$ Gene $\xrightarrow{binded}$ Chemical compound	instance correlation
	Chemical compound $\xrightarrow{bind}$ Gene $\xrightarrow{has}$ Tissue $\xrightarrow{has}$ Gene $\xrightarrow{binded}$ Chemical compound	instance correlation
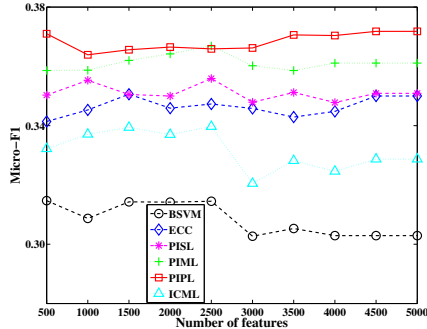
(a) Gene-Disease Association Prediction
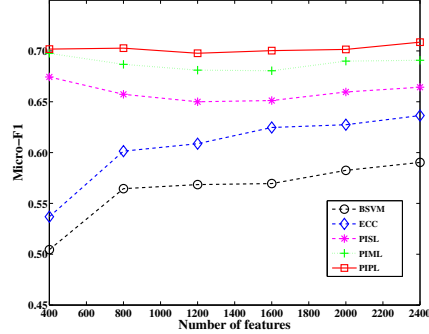
(b) Drug-Target Binding Prediction

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

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

dataset	publication	network	type of classification	#types (node,link)	#Nodes	Node Types
Gene	[34]	homogeneous	single-label	(1,1)	1,243	gene
Citeseer	[22]	homogeneous	single-label	(1,1)	3,312	paper
WebKB	[7]	homogeneous	single-label	(1,1)	3,877	webpage
DBLP	[17]	homogeneous	multi-label	(1,1)	4638	author
ACM	[18]	heterogeneous	single-label	(5,5)	12,499	paper,author,conference,proceeding,school
Cora	[23]	heterogeneous	single-label	(5,5)	4,330	paper, author, journal, publisher, editor
IMDB	[1]	heterogeneous	single-label	(5,5)	1,382	movie, actor, director, studio, producer
NASD	[24]	heterogeneous	single-label	(5,6)	22,000	broker, branck, firm, disclosure, regulator
SLAP	this paper	heterogeneous	multi-label	(10,11)	280,792	drug, disease, gene, side effect, pathway tissue, GO, gene family, substructure,CO

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

8. REFERENCES

[1] http://www.imdb.com.

[2] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown. Learning multi-label scene classification. *Pattern Recognition*, 37(9):1757–1771, 2004.

[3] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[4] B. Chen, Y. Ding, and D. Wild. Assessing drug target association using semantic linked data. *PLoS Comput Biology*, 8(7), 2012.

[5] G. Chen, Y. Song, F. Wang, and C. Zhang. Semi-supervised multi-label learning by solving a sylvester equation. In *SDM*, 2008.

[6] W. Cheng and E. Hüllermeier. Combining instance-based learning and logistic regression for multilabel classification. *Machine Learning*, 76(2-3):211–225, 2009.

[7] M. Craven, D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam, and S. Slattery. Learning to extract symbolic knowledge from the world wide web. In *AAAI*, 1998.

[8] K. Dembczyński, W. Cheng, and E. Hüllermeier. Bayes optimal multilabel classification via probabilistic classifier chains. In *ICML*, 2010.

[9] A. Elisseeff and J. Weston. A kernel method for multi-labelled classification. In *NIPS*. 2002.

[10] N. Ghamrawi and A. McCallum. Collective multi-label classification. In *CIKM*, 2005.

[11] S. Godbole and S. Sarawagi. Discriminative methods for multi-labeled classification. In *PAKDD*, 2004.

[12] W. Indyk, T. Kajdanowicz, and P. Kazienko. Relational large scale multi-label classification method for video categorization. *Multimedia Tools and Applications*, 2012.

[13] M. Ji, J. Han, and M. Danilevsky. Ranking-based classification of heterogeneous information networks. In *KDD*, 2011.

[14] F. Kang, R. Jin, and R. Sukthankar. Correlated label propagation with application to multi-label learning. In *CVPR*, 2006.

[15] H. Kazawa, T. Izumitani, H. Taira, and E. Maeda. Maximal margin labeling for multi-topic text categorization. In *NIPS*. 2005.

[16] X. Kong, B. Cao, P. Yu, Y. Ding, and D. Wild. Meta path-based collective classification in heterogeneous information networks. *CoRR*, 2013.

[17] X. Kong, X. Shi, and P. Yu. Multi-label collective classification. In *SDM*, 2011.

[18] X. Kong, P. Yu, Y. Ding, and D. Wild. Meta path-based collective classification in heterogeneous information networks. In *CIKM*, 2012.

[19] X. Kong and P. S. Yu. Multi-label feature selection for graph classification. In *ICDM*, 2010.

[20] N. Lao and W. Cohen. Relational retrieval using a combination of path-constrained random walks. *Machine Learning*, 81(2):53–67, 2010.

[21] Y. Liu, R. Jin, and L. Yang. Semi-supervised multi-label learning by constrained non-negative matrix factorization. In *AAAI*, 2006.

[22] Q. Lu and L. Getoor. Link-based classification. In *ICML*, 2003.

[23] A. McCallum, K. Nigam, J. Rennie, and K. Seymore. Building domain-specific search engines with machine learning techniques. In *AAAI Spring Symposium on Intelligent Agents in Cyberspace 1999*, 1999.

[24] J. Neville, O. Simsek, D. Jensen, J. Komoroske, K. Palmer, and H. Goldberg. Using relational knowledge discovery to prevent securities fraud. In *KDD*, 2005.

[25] S. Peters, Y. Jacob, L. Denoyer, and P. Gallinari. Iterative multi-label multi-relational classification algorithm for complex social networks. *Social Network Analysis and Mining*, 2:17–29, 2012.

[26] J. Read, B. Pfahringer, and G. Holmes. Multi-label classification using ensembles of pruned sets. In *ICDM*, 2008.

[27] J. Read, B. Pfahringer, G. Holmes, and E. Frank. Classifier chains for multi-label classification. In *ECML-PKDD*, 2009.

[28] Y. Sun, J. Han, X. Yan, P. Yu, and T. Wu. PathSim: Meta path-based top-k similarity search in heterogeneous information networks. In *VLDB*, 2011.

[29] Y. Sun, J. Han, P. Zhao, Z. Yin, H. Cheng, and T. Wu. Rankclus: Integrating clustering with ranking for heterogeneous information network analysis. In *EDBT*, 2009.

[30] Y. Sun, B. Norick, J. Han, X. Yan, P. Yu, and X. Yu. Integrating meta-path selection with user-guided object clustering in heterogeneous information networks. In *KDD*, 2012.

[31] Y. Sun, Y. S. Yu, and J. Han. Ranking-based clustering of heterogeneous information networks with star network schema. In *KDD*, pages 797–806, 2009.

[32] G. Tsoumakas and I. Vlahavas. Random k-labelsets: An ensemble method for multilabel classification. In *ECML-PKDD*, 2007.

[33] N. Ueda and K. Saito. Parametric mixture models for multi-labeled text. In *NIPS*. 2003.

[34] X. Xiang and J. Neville. Pseudolikelihood em for within-network relational learning. In *ICDM*, 2008.

[35] M.-L. Zhang and K. Zhang. Multi-label learning by exploiting label dependency. In *KDD*, 2010.

[36] M.-L. Zhang and Z.-H. Zhou. Ml-knn: A lazy learning approach to multi-label learning. *Pattern Recognition*, 40(7):2038–2048, 2007.

[37] X. Zhang, Q. Yuan, S. Zhao, W. Fan, W. Zheng, and Z. Wang. Multi-label classification without the multi-label cost. In *SDM*, 2010.